WOULD ACCOUNTABILITY BASED ON TEACHER VALUE-ADDED BE SMART POLICY? AN EXAMINATION OF THE STATISTICAL PROPERTIES AND POLICY ALTERNATIVES

Douglas N. Harris
University of Wisconsin at Madison
June 19, 2008

Abstract: With annual standardized student testing, it now may be feasible to measure the contributions to student achievement made by individual teachers. But will these "teacher value-added" measures help to improve student achievement and learning? I address this question through a "policy validity" framework" that includes three factors: (a) *statistical validity*, i.e., how well teacher value-added measures actually measure true teacher contributions to achievement; (b) *purposes*, i.e., whether the measures are intended to signal effective teachers or provide a path to improvement; and (c) *costs*. Regarding statistical properties, I outline many of the key assumptions of value-added, most of which are rejected by empirical evidence. However, there is some evidence that the assumption violations may not be severe and that the measures contain useful information. I also compare the policy validity of teacher value-added accountability to three main policy alternatives: teacher credentials, school value-added accountability, and formative uses of test data. This analysis shows that teacher value-added is likely to increase student achievement more effectively and efficiently than a teacher credentials-only strategy, but it is unclear whether teacher value-added would raise achievement more than alternative uses of student test scores. Resolving this issue will require studies of how different uses of student test scores influence instruction and student learning in practice.

Acknowledgements: The author was chair of the National Conference on Value-Added Modeling, which included events at UW-Madison, April 22-24, 2008 and the Urban Institute in Washington, DC, May 23, 2008. He wishes to thank co-chairs Adam Gamoran and Stephen Raudenbush, program committee members Henry Braun, JR Lockwood, Robert Meyer, and Tim Sass, and Jane Hannaway who led the Urban Institute event. For their useful comments on this article, the author wishes to thank Bob Floden, Adam Gamoran, Drew Gitomer, David Monk, Tim Sass and participants in the National Conference on Value-Added, the 2007 Educational Testing Service Invitational Conference in San Francisco, CA and a seminar at the University of Colorado at Boulder. The author also gratefully acknowledges financial support from the Carnegie Corporation, Educational Testing Service, Joyce Foundation, and Spencer Foundation.

Introduction

Annual standardized student testing is a pervasive, and probably permanent, piece of the U.S. K–12 education system. But this has not resolved the age-old issue about whether and how policymakers should use the test results in accountability systems. During the 1970s, many states expanded student testing and adopted minimum competency exams that students had to pass to graduate from high school. In the 1980s, they added school report cards to the accountability mix, reporting point-in-time snapshots of average school achievement. This trend toward test-based school-level accountability accelerated in the 1990s with state policies such as school grades, reconstitution, takeovers, and other incentives (Harris & Herrington, 2007). *No Child Left Behind* (NCLB) appears to have cemented school-level test-based accountability as a key lever in the national education strategy, but this is a broad strategy and leaves the door open for a wide variety of policies regarding the use of standardized test scores. One new policy option that has intrigued players on all sides of the education debate is accountability based on how much "value-added" teachers and schools contribute to student to achievement.

A fundamental problem in holding schools accountable for student achievement is that, as economists put it, education is jointly produced by schools, families, and communities (Hanushek, 1979). Students' socio-economic status is arguably the strongest predictor of their educational outcomes, an observation dating at least as far back as Coleman (1966). There is also strong evidence that this correlation reflects a causal relationship. The achievement levels of Black kindergarteners are half a standard deviation below the levels of White kindergartners (Fryer & Levitt, 2004). Since such differences occur before students enter school, they must be due to

family, community and other factors outside of school control.[1] Rothstein (2004) describes the myriad ways that family and community factors influence student learning. It is therefore no surprise that a school serving White students from middle and high income families is 89 times more likely than a high-minority, high-poverty school to be among a state's top-third on achievement tests (Harris, 2007).

These facts pose difficulties for school accountability systems whose expressed goals are to measure and reward school performance. If school performance measures substantially reflect non-school contributors to student success, as is the case with NCLB and typical state school report cards, then genuine improvements in school performance will not show up in higher performance measures, leaving schools with weak incentives to improve and perverse incentives to "cream" the most socio-economically advantaged students (Harris, 2007) and push out low performers (Figlio, forthcoming). In this sense, school accountability based on such misleading performance measures is not only unfair to the schools but unfair to the students as well.

Value-added modeling has drawn wide interest in recent years as a way to solve this problem and isolate the contribution of schools. The basic logic is simple: if each student's achievement is measured every year, then, in trying to determine each school's performance, we can take into account where students started at the beginning of each year and therefore indirectly account for the family and community factors that contribute to achievement. This differs from typical school report cards, and to some degree NCLB, which do not account for where students start.[2]

---

[1] Lee and Burkham (2002) also provide extensive evidence on these "starting gate inequalities" using the same database as Fryer and Levitt (2004).

[2] Some state school report cards such as Florida and Kentucky do include achievement growth as part of the school performance measures, but these are exceptions. NCLB includes several pilots with "growth models" but this name is

Value-added can also be used to measure the performance of teachers and hold them accountable as individuals. The U.S. education system has a long history with one form of teacher accountability—teacher merit pay—dating to the early part of the last century (Murnane and Cohen, 1986). Showing renewed interest, many districts such as Denver and states such as Florida are again experimenting with merit pay, using student achievement as a key component of the performance measures on which merit and compensation bonuses are paid. The federal government has also taken up the idea; some of the district efforts are being funded by the federal government's Teacher Incentive Fund and new efforts in Congress are poised to create new federal supports.[3] Also, Gordon, Kane & Staiger (2006) proposed using teacher value-added as a primary basis for teacher tenure decisions, an idea that some school districts such as New York City have considered.[4] More than just adding accountability, these policies and proposals take aim at some of the oldest and firmly established teacher-related policy traditions—nearly guaranteed job security and compensation based on credentials (i.e., the single salary schedule).

Even many advocates of test-based accountability, however, acknowledge that measuring teacher contributions to student test scores is difficult. Teacher value-added is intended to address

---

misleading and schools are still primarily held accountable for achievement levels rather than growth even in growth model states.

[3] Full disclosure: The author is a member of the Technical Advisory Committee for the Teacher Incentive Fund. Congressman George Miller (D-CA), a leader on education issues in the U.S. House of Representatives, recently proposed a program in which school districts could apply for funds to provide additional compensation to teachers in low-performing schools if teachers demonstrated performance on measure such as student test scores. For more recent discussions of the background and evidence on teacher merit pay, see Figlio and Kenny (2007) and Podgursky and Springer (2007).

[4] The local and state teacher unions strongly opposed a proposal by NYC Chancellor Joel Klein to use student test scores in tenure decisions. A state law was quickly adopted placing a two-year moratorium on the idea. Gordon, Kane, and Staiger (2006) write specifically that "We propose federal support to help states measure the effectiveness of individual teachers—based on their impact on student achievement, subjective evaluations by principals and peers, and parental evaluations. States would be given considerable discretion to develop their own measures, as long as student achievement impacts (using so-called "value-added" measures) are a key component. The federal government would pay for bonuses to highly rated teachers willing to teach in high-poverty schools. In return for federal support, schools would not be able to offer tenure to new teachers who receive poor evaluations during their first two years on the job without obtaining district approval and informing parents in the schools. States would open further the door to teaching for those who lack traditional certification but can demonstrate success on the job" (p.2).

these difficulties but, as I describe in the next section, many assumptions have to hold in order to interpret value-added measures as true (causal) contributions to student achievement. Many of the assumptions seem unrealistic and have been rejected, most recently in studies commissioned as part of the National Conference on Value-Added Modeling, which are the main subjects of this volume and the review of evidence. On the other hand, it is still possible for the results to contain useful information, so long as the assumptions are not violated too severely. Later in the next section, I discuss evidence that provides a limited degree of confidence in the validity of value-added measures.

No performance measure or accountability policy is perfect, of course, and it may be that teacher value-added accountability is better than the alternatives. I compare rewards for teacher credentials, accountability based on school value-added, and formative uses of standardized tests as alternatives to teacher value-added accountability, using the "policy validity" framework outlined by Harris (forthcoming). The first element of the framework, *statistical validity*, refers to the relationship between any teacher quality measure and the construct it attempts to measure, which researchers and policymakers increasingly define as the contribution to student achievement. The second element of the framework is the *purpose*. There are two main purposes with regard to student achievement scores. First, accountability, by definition, involves creating signals or summative assessments of effectiveness—that is, determining who is performing well, as the basis for incentives.[5] But knowing who is performing well does not provide a path to improvement; or, borrowing the language of teacher educators, it is not "formative" in nature. The two types are interrelated in that any path to improvement may be of little use unless teachers have incentives to

[5] The term "signal" in economics typically refers to the indicators of effectiveness that precede employment and observations of actual performance. Here, I use the term somewhat more broadly so that it encompasses signals such as credentials, which fit the economics definition, and measures such as value-added that are based on observed performance. The use here is therefore more akin to the statistical meaning as in "signal-to-noise" ratio.

improve; likewise, providing incentives for teachers to improve without a path to improvement may do little to drive performance. The third and final element of the framework is cost. While several researchers emphasize the fact that the costs of education programs are just as important as their effects (Harris, 2008; Levin & McEwan, 2001), there remains little evidence on the cost-effectiveness of education programs (Levin, 1991; Rice, 2002). I compare teacher value-added to the various policy alternatives using this three-factor framework.

After discussing the policy validity of teacher value-added accountability, I provide brief discussions of the three policy alternatives. Many of the assumptions and statistical issues that arise with teacher value-added also arise with school value-added, though the task appears more feasible when the focus is on whole schools. For this and other reasons, it is unclear whether teacher accountability based on value-added is superior to the alternatives in raising student achievement.

Policy Validity of Teacher Value-Added

Value-Added and Its Assumptions

From a statistical standpoint, isolating the impact of schools from that of families and communities is a problem of non-random assignment of teachers to students. If all teachers had the same chance be assigned to any given student, and if we had enough observations on each teacher, we could draw conclusions about each teacher's effectiveness simply by looking at the end-of-year test results and complex statistical adjustments would be entirely unnecessary. But there is ample evidence of non-random assignment, e.g., that the most disadvantaged students are assigned to the least qualified teachers (Clotfelter et al., 2005) and "tracked" into classrooms within schools that have other disadvantaged students (Gamoran, 1986; Oakes, 1985; Ogbu, 2003). The potential

attraction of value-added is that it may allow us to indirectly account for family and community factors even when these types of non-random assignment arise.

While the popularity of the term value-added is relatively new, the ideas and their application to education date at least as far back to Hanushek (1979) and Boardman and Murnane (1979). For more recent discussions, which discuss theoretical and technical issues, see for example Harris and Sass (2005) and Todd and Wolpin (2003). Below, I briefly describe some of the key assumptions of the models and recent evidence about their validity, including evidence from many of the articles in this volume. While the focus below is on the economics-based value-added model, the discussion is applicable to most value-added models in common use.

It is important to emphasize that even most of the strongest advocates of test-based accountability express some concern about designing accountability that involves pitting teachers against one another within the same school (D.C. Harris, 2007). This means that calculating teacher value-added for accountability cannot be done in a way that compares teachers only within schools.[6] Thus, only comparisons of teachers across schools are relevant here, and in statistical terms this means omitting school fixed effects. As we will see below, this constraint on the estimation of value-added models has significant consequences for the statistical validity of the models.

Assumption #1: School administration and teamwork among teachers do not have a significant impact on student achievement. The first implication of the need to compare teachers across schools is that it becomes quite difficult to account for the impact of school administration.

[6] Harris (forthcoming) distinguished value-added for accountability (VAM-A) from value-added for program evaluation (VAM-P). A strong case can be made for comparing teachers within schools in VAM-P models so that the impact of school administration and the non-random assignment of teachers to schools can be accounted for; in contrast to VAM-A, the results from VAM-P studies are only used to set broad policies rather than evaluate teachers in ways that might lead teachers to compete with one another in unproductive ways.

If we were making within-school comparisons, then we might reasonably assume that the impact of school administration affects all teachers relatively equally so that there is no need to account for it. But when estimating value-added models for accountability, this approach fails because very few teachers are observed in multiple schools. This leaves one of two options: (a) measure the quality of school administration directly (e.g., through surveys of teachers and parents) and include these in the value-added model; or (b) assume that the impact of administration is small. Information from surveys is rarely, if ever, used in external accountability systems, which limits the practicality of the first option. This leaves the second option, which I label Assumption #1.

A similar problem arises with teacher teamwork. The purpose here is to measure how much each teacher contributes to student achievement, but it is possible, contrary to the assumptions of value-added, that teachers contribute to the achievement of students of other teachers, e.g., by mentoring. The only rigorous evidence I am aware of on this point is Harris and Sass (2007b) who find that the number of NBPTS teachers in a school has no impact on the value-added of other teachers within the same schools, but this is far from definitive.[7] It is possible that neither administration nor teamwork play a significant role; some researchers describe teaching as "loose coupled" meaning that teachers mainly work on their own in their classrooms, making it difficult for anyone else to have a significant impact on what they do or how well they do it. The evidence on Assumption #1 is quite thin, but there is really no way that any subsequent improvements in the methodology of value-added could ever allow the assumption to be relaxed.

Assumption #2: Controlling for previous achievement levels is sufficient to account for the impact of past school resources. Education is a cumulative process. The educational resources students receive early in life affect their academic success later in life. But as a practical matter, it

[7] There are other studies of mentoring, but these are formal mentoring programs where teachers have formal mentoring roles. Here, we are interested in the general effects of teachers.

is impossible to explicitly measure the whole range of resources students receive at any given time, let alone in past years. It may be possible to account for resources indirectly, however, because the effects of all resources should be reflected in subsequent achievement. This means that when trying to explain why students reached achievement level *A* at time *t* (A_t), we can account for past school resources by controlling for achievement in the previous time period (A_{t-1}). The effects of all school resources experienced up to time *t-1* should be reflected in A_{t-1}.

Notice that Assumption #1 involves only "school resources." Controlling for past achievement also accounts for family and community factors, but only under the assumption that students are assigned to teachers based solely on their previous achievement, and not based on unobserved student characteristics that may also be related to students' subsequent achievement. This is implausible. For example, Feng (2005) finds that students are assigned to teachers partly based on students' discipline problems, which are generally unobserved. Also, Harris and Sass (2005) and McCaffrey, Sass, and Lockwood (this volume), show that the findings regarding teacher value-added are quite different when relying solely on Assumption #1 to account for student differences. Because of the centrality of family and community factors, researchers have developed value-added models that do not require such restrictive assumptions (see Assumption #3 below).

Related to this is the need to make an assumption about the rate at which the impact of past school resources decays or fades out.[8] At the one extreme, past schooling resources (e.g., last year's teacher) may have very little impact on current achievement. At the other extreme, past schooling resources could be just as important as current ones in affecting current achievement. Kane and Staiger (2008) show that the impacts that individual teachers decay by 50 percent or more per year.

[8] Value-added models assume that the decay is geometric.

That is, the impact of having a good teacher does not seem to last. As Rothstein (this volume) points out, this could be because the variation in teacher value-added is driven by differences in instruction that have only ephemeral impacts, e.g., how much teachers teach to the test. Another possible explanation is that the content of achievement tests is somewhat independent across years. To take an extreme example, suppose students need not understand any of the academic content covered on the 3rd grade test in order to learn the academic content on the 4th grade test. In that case, we would not expect the 3rd grade teacher's contribution to the 3rd grade achievement score to have any impact at all on the 4th grade test. Whatever the explanation, the apparently high rate of decay is not an assumption of value-added models, but rather an outgrowth of the estimation of the models.[9] Harris and Sass (2005) find that the impact of school resources is relatively insensitive to any decay assumption that might be imposed.

Assumption #3: Students' contributions to their own achievement can be measured with student fixed effects that account for the non-random assignment of students to teachers ("static selection").[10] As noted above, it would be unrealistic to assume that students are assigned to teachers based on fixed, observable qualities only. To address this, economists typically include student "fixed effects" in their value-added models, which represent the conditional average rate of achievement growth over all the years they are in the database. They are "conditional" in the sense that these control for students' average school resources and other factors that might influence

[9] This statement refers to models in which lagged achievement is included as an independent variable. Value-added models that use the change in score as the dependent variable do assume zero decay. When lagged achievement is included on the right-hand side, the rate of decay can be estimated directly or restricted to a specific value. The empirically estimated rate of decay depends on other aspects of the model specification. For example, the rate of decay in a model with student fixed effects is likely to be lower because, in the absence of student fixed effects, lagged achievement reflects both average achievement and the year-specific deviation. With student fixed effects, lagged achievement reflects only the latter.

[10] Value-added modeling accounts for student growth both before the teacher taught the student and afterward. For example, if we were studying the effect of a fourth-grade teacher, the student's average rate of growth would be estimated to account for student learning in third and fifth grade as well as fourth.

student learning in any give year that are largely outside teachers' control, and that might influence student learning.[11] Rothstein (this volume) describes this as the "static selection" assumption. This does not preclude changes over time in students' propensities to make learning gains, but it does mean that any time-varying propensities are randomly distributed among teachers. Otherwise, there is what Rothstein (this volume) calls "dynamic selection" which may introduce bias into the teacher value-added measures even when student fixed effects are included.

The static selection assumption with student fixed effects is almost certainly more realistic than the alternative of no selection based on unobservable qualities that is required in the absence of student fixed effects (see Assumption #2 and Harris and Sass (2007a)), which explains why economists typically include student fixed effects in their models. But the matter is still not completely settled. Kane and Staiger (2008) report on an experiment involving 78 classrooms in the Los Angeles School District. The researchers solicited school principals willing to randomly assign teachers to classrooms within their schools. The researchers then compared the value-added measured before the experiment to those calculated on the basis for random assignment which, so long as the random assignment was carried out with fidelity, cannot be driven by systematic assignment of students to teachers. Specifically, they regressed mean end-of-year test scores on previous value-added. A coefficient of one on the value-added variable would seem to suggest that value-added is a perfect predictor of teacher contributions when random assignment is used. For some value-added specifications they indeed find coefficients close to one, but the student fixed effects model did not perform as well as some others.

[11] As discussed by Harris (forthcoming-a), "The fixed student contribution is often called 'innate ability' by economists and is akin to what psychologists consider general intelligence, or *g*. The more general term, 'fixed student contribution,' is used here because it is virtually impossible with education data sets to estimate anything like innate ability. No data sets include measures of student abilities at birth or, in their absence, sufficiently measure family and other environmental factors well enough to distinguish innate from environmental differences."

Also, Rothstein (this volume) tests the dynamic selection assumption by considering whether the teacher assignment in any given year predicts *past* achievement growth. While we would expect the *current* teacher to affect *current* achievement, a current teacher cannot change what has already happened—or "rewrite history"—and will only appear to do so when students are non-randomly assigned. Rothstein estimates value-added models with student fixed effects and indeed finds that current teacher assignment does predict past student achievement and therefore rejects the static selection assumption.

It is worth considering how violations of the static selection assumption might arise in practice. The most obvious explanation, and the example commonly given to explain the above findings, is that school principals "track" students and do not randomly assign teachers to tracks. Monk (1987) finds that most school principals randomly or evenly distribute students in elementary grades, apparently because principals want to even out the workload among teachers. But he also finds that some principals try to match students to teachers who have skills particularly well suited to students' needs, thus violating the static selection assumption.[12]

This assumption becomes even more problematic when we recall that value-added for accountability is only practical when comparing teachers across schools. In these cases, even if teachers are evenly or randomly assigned within schools, the non-random selection of teachers to schools creates the same type of problem. Principals cannot randomly select teachers from the

[12] Rothstein also discusses the issue of principal assignment decisions, writing that "it requires in effect that principals decide on classroom assignments for the remainder of a child's career on the day that child begins kindergarten" (p.10). This statement unintentionally makes the assumption seem less realistic than it is. As noted above, the assumption of value-added models is satisfied under the "even distribution" assumption, even if the decisions about even distribution are made "dynamically" such that principals take into account time-varying information about students. It would therefore be more accurate to say, in the context of within-school comparisons of teachers, that the models assume that some principals randomly assign students and the remaining principals make decisions about each year's track based solely on the previous year's track, without making use of any new information. This still seems implausible, but a little less so than Rothstein's formulation. Also note that Rothstein's evidence seems to reject even the weaker assumption.

entire population of potential teachers, or even from the entire pool within their respective school districts. Rather, they can only choose from among the teachers in their schools. There is ample evidence of non-random assignment of teachers to schools and that assignment is correlated with factors (teacher experience, etc.) that are sometimes related to teacher value-added (Clotfelter et al. 2005). One possible solution to this problem is to compare teachers across *similar* schools. This approach is used at the school level (that is, comparing schools with similar student demographics) in England (see Evans, this volume).

Assumption #4: A one-point increase in test scores represents the same amount of learning regardless of the students' initial level of achievement or the test year. Value-added models are, at a basic level, models of student achievement. Therefore, it is unsurprising that value-added requires strong assumptions about the measurement of student achievement. Specifically, it is assumed that a one-point change in the score is the same on every point on the test scale—that is, the test is interval-scaled. Even the psychometricians who are responsible for test scaling shy away from making this assumption in the strict sense.

Some adjustments can be made in the value-added analysis to account for the scale problems. For example, some researchers add grade-by-year fixed effects, which adjust each teacher's value-added based on the mean achievement of all students in the respective grade and year. However, this amounts to simply shifting teachers' value-added based on the mean gain in the years and grades in which they have taught. This approach is sufficient so long as the scaling problems influence only the mean gain and not, for example, the distribution around the mean. In that case, an arguably better approach is to "normalize" all the test scores to a mean of zero and standard deviation of one, based on the standard deviation of the respective grades and years. This

approach requires the assumption that the differences in the standard deviations (and means) are due to changes in the scale rather than any genuine changes in the learning distribution.[13]

Ballou (this volume) argues that the assumptions of traditional scaling techniques, based on Item Response Theory (IRT), are inherently difficult to test. Further, even the plausibility of the resulting test scales from these methods is questionable and other reasonable approaches yield quite different measures of achievement gain. Ballou describes an alternative non-IRT method of measuring student progress, requiring less restrictive assumptions, in which students are ranked based on their achievement gains and then teacher value-added is calculated based on these rankings rather than the gains themselves.[14] He finds that the rankings of teachers on their value-added often vary dramatically between the traditional IRT approach and the student rank-order approach, even though cases can be made for each. Briggs and Wiley (this volume) also examine sensitivity to test scaling and find less sensitivity than Ballou, but this likely due to: (a) the narrower range of assumptions that they consider (all fall within the IRT paradigm); and (b) the fact that they focus on school value-added rather than teacher value-added. Given that the variation in true student gains is larger across teachers than across schools, teacher value-added is almost necessarily more sensitive to the test scale than school value-added.

Assumption #5: Teachers are equally effective with all types of students. The fact that students and teachers are not randomly assigned has already been established. One potential problem that arises from this is that some teachers might be assigned to students who are less likely to make achievement gains. Even if the value-added models succeed in accounting for this, teachers may vary in how much they contribute to learning of different types of students.

---

[13] This is not the only assumption required regarding the properties of the student achievement tests. For example, there is also an implicit assumption that the content of the tests is constant over time.

[14] The advantage of ordinal scales is that they require less restrictive assumptions, although they do throw out potentially useful information.

To see the problem more clearly, suppose that some teachers were effective with low-achievement students and other teachers were effective with high-achievement students. Further, suppose that all teachers were assigned only to students with whom they were most effective and that, in such a situation, all teachers appear equally effective in their value-added score. Now, suppose instead that some teachers were "mis-assigned" to students with whom they were ineffective, and as a result, their value-added scores decrease. These same teachers who had been judged effective above will now appear ineffective simply because of the assignment process. This is problematic because teachers cannot control which students they are assigned to, and it would be difficult to argue that these mis-assigned teachers are really less effective than the others.

The above example is an extreme case, intended to illustrate the potential problem created for value-added if teachers are not equally effective with all students. Lockwood and McCaffrey (this volume) conclude that differential effects explain less than 10 percent of the variation in overall teacher effects. Therefore, what seems like a potential issue in theory may not be significant in practice.

The above five assumptions do not represent an exhaustive list of assumptions that apply to all value-added models, though they are arguably the ones that are considered to be potentially most problematic.[15] Other assumptions vary depending on the model specification. Harris and Sass (2005) test a variety of these assumptions. It is also important to point out that these assumptions may be interrelated so that violating one assumption might compound, or offset, the

[15] Another assumption is that student test data are missing at random. The data requirements for value-added are significant and that data will be missing for a large portion of the students, due to absenteeism, mobility across schools, and data processing errors. Missing data do not bias the results so long as they are missing at random, though missing data significantly diminish the reliability of the estimates. This is a strong assumption and is especially likely to be a problem in high-poverty schools where absenteeism and mobility are high and test-taking rates are lower. It is therefore a significant question whether valid value-added estimates can be made in schools with high mobility.

impact of violations in other assumptions. Research at present is mainly focused on testing individual assumptions, which is often quite complicated by itself.

Statistical Properties of Teacher Value-Added

It is possible that all of the assumptions of value-added models are violated, but that the violations are not so severe that they have a practical impact on whether teachers would be rewarded or punished in an accountability system. Conversely, all of the assumptions might hold, but the models might still not have the statistical properties necessary for particular types of policy uses. This section explores other empirical findings regarding value-added that are relevant to understanding their usefulness for accountability.

Teacher value-added is positively correlated with other measures of teacher effectiveness. Teacher value-added can be viewed as an objective measure of teacher effectiveness in the sense that the method of calculating it is the same for all teachers and is not filtered through the subjective preferences and beliefs of a supervisor or other evaluator. There is a long history of research studying the relationships between subjective and objective measures of worker productivity as well as the implications of this relationship for employment contracts. As noted by Harris and Sass (2007c) and Jacob and Lefgren (2005), this research suggests that there is a positive, but arguably weak, relationship between subjective and objective measures. There is also a limited amount of literature that specifically addresses this issue. Some studies have examined the relationship between teachers' students' test scores and their principals' subjective assessments (e.g., Milanowski, 2004; Murnane, 1975). All of these studies find a positive and significant relationship despite differences in the degree to which the observations are used for high-stakes personnel decisions.

Some more recent studies have utilized longitudinal data to estimate gain scores models that partly address the selection bias issues described earlier (Medley & Coker, 1987; Peterson, 1987, 2000). Also, Jacob and Lefgren (2005) used value-added models to study two hundred teachers in a midsized school district and reached two main conclusions: there is a positive correlation between the subjective and objective measures, and this correlation holds even after controlling for teacher experience and education levels, which are currently the primary bases for determining teacher compensation. Harris & Sass (2007c) found similar results from an analysis of a separate midsized school district in Florida. In addition to asking for their overall subjective assessments, they asked principals how well teachers contributed to student achievement so that could determine how much of their subjective assessments reflected teachers' contribution to outcomes other than achievement. With a simple correlation of 0.7, the results suggest that achievement is probably the main objective of these principals, but also that other outcomes such as motivation and socialization likely explain the modest size of the correlation between the two measures.[16] For this reason, the comparison of principal evaluations of teachers with teacher value-added measures cannot be viewed as a validity check per se, but it does suggest that value-added measures provide useful information.

Value-added measures have been replicated in a randomized control trial. Recall that Kane and Staiger (2008), in their study of the Los Angeles School District, were able to nearly replicate random assignment-based estimates of teacher effectiveness with non-experimental value-added estimates. Conducting an experiment of this sort is inherently difficult which makes Kane and Staiger's work especially impressive. However, there are some limitations that make it difficult to

[16] A related issue is that school principals in the study had some access to some of the same data as the researchers and their assessments of teachers' contributions to student achievement might have been direct reflections of this. On the other hand, the principals had, at most, access to simple student achievement gains and not the value-added measures described here.

view this as a validation of teacher value-added measures. First, it is unclear how principals were assigning teachers before the experiment took place. If they were assigning teachers in effectively random ways, then the "experiment" is really no different from what was already happening and their results could not be interpreted as evidence in support of value-added.[17] On the other hand, if principals were tracking students and non-randomly assigning teachers to different types of teachers, then the results here are significant and reinforce the potential of teacher value-added.

Based on these findings—that teacher value-added is correlated with principal evaluations and has been replicated in a random assignment experiment—the news on value-added reinforces the potential use of value-added for accountability. This is not the case with the following two findings.

Teacher value-added scores are imprecise. A pre-requisite for any performance measure to be useful is that different teachers obtain different scores. Sanders and Horn (1998) and Rivkin, Hanushek, and Kain (2005), for example, find considerable differences between the most and least effective teachers, based on value-added results. However, it is important to consider to what degree that this reflects variation in actual performance.

Kane and Staiger (2001, 2002) provide one of the best and most well known discussion of the types of errors in value-added. Using data from North Carolina, they concluded that only about half the variation in grade-level achievement gains is due to "persistent" differences between schools—that is, to differences that could plausibly be attributed to factors under the control of the schools. This is noteworthy given that their analysis was conducted at the grade level where classrooms are grouped together and where the amount of imprecision is therefore likely to be

[17] Their findings regarding value-added specifications would still be valid even if principals had been randomly assigning teachers and students to begin with. Each specification makes different assumptions, as the earlier discussion highlights, and the goal is to get as close to the experimental estimates as possible.

better than for individual teachers. Reinforcing this point, they showed that the persistent component of grade-level gains was considerably smaller in schools with fewer students. Other researchers have shown that teacher value-added scores are imprecise enough that, by the usual standards of statistical significance, it is only possible to clearly distinguish very-low-value-added teachers from very-high-value-added teachers (Jacob & Lefgren, 2005). This is a problem for policies that intend to make high-stakes decisions based on the measures, except perhaps if those decisions only pertain to rewards for very high performers and punishments (e.g., rejection of tenure) for very low performers. It is also a difficult problem to address because the number of students per teacher is essentially uncontrollable; therefore, this is not a problem that will ever be solved by changing the model specification.

Some of the "other non-persistent" variation identified by Kane and Staiger (2001, 2002) is driven by measurement error. Boyd et al. (this volume) explain how to account for measurement error and show that the impact of measurement error is greater in value-added models than in cross-sectional models because value-added models, by definition, involve changes in achievement over time. While accounting for measurement error reduces the observed variation in teacher value-added scores, it certainly does not eliminate it which means there are still meaningful differences in teacher performance.[18]

Individual teacher value-added is unstable over time. Intuitively, we would expect that the actual effectiveness of each teacher changes little from year to year. Teachers might gradually improve over time, as suggested by the earlier discussion of evidence on teacher experience, but it is unlikely that they will jump from the bottom to the top of the performance distribution. It is even

[18] Boyd et al. focus more on the impact of measurement error on the estimated impacts of programs Harris (forthcoming) distinguishes between value-added for accountability (VAM-A), which is the focus of the present study, and value-added for program evaluation (VAM-P). For examples of the latter, see the later discussion of evidence on the relationship between teacher value-added and teacher credentials.

less likely that true teacher rankings on value-added should drop significantly over a short period of time, except perhaps in cases such as divorce or other significant change in teachers' family status or health.

Some of the earliest evidence on this topic, however, suggested that teacher value-added is much more unstable than this intuition would suggest. Koedel and Betts (2007) found that only 35 percent of teachers ranked in the top fifth of teachers on teacher value-added one year were still ranked in the top fifth in the subsequent year. This suggests that 65 percent of teachers actually got worse relative to their peers over a short period of time—some dramatically worse. Stability appears somewhat higher in studies by Aaronson et al. (2007) and Ballou (2005), but this may be due solely to the fact that, in contrast to Koedel and Betts who divided teachers in five groups, these other two studies divided teachers into only four groups, making it less likely that changes in groups would be observed. Overall, these results are remarkably similar across studies.

McCaffrey, Lockwood, and Sass (this volume) make an important contribution to this literature by showing that the vast majority of this instability is due to measurement error. Once measurement error in test score gains has been accounted for, the degree of stability, as measured above by the percentage of teachers staying within the same quartile from one year to the next, doubles or triples so that 50-90 percentage of teachers remain in the same performance group.

Like Koedel and Betts (2007), McCaffrey, Sass, and Lockwood also find that stability greatly improves when student and school effects are omitted. This is a predictable result because there is mobility off teachers across schools that changes the basis of comparison each year. There is much less variation in the entire pool of teachers in a school district or state, and therefore less change who each teacher is being compared with. The increased stability would seem to suggest that the unobserved heterogeneity of the average teacher's students (which the student fixed effects

are supposed to account for) is unstable over time. Without student fixed effects, unobserved heterogeneity, which itself is unstable, shows up in the teacher value-added.

The remaining instability may be due to genuine changes in teacher effectiveness over time, which value-added measures are intended to capture, or to violations in the assumptions. For example, as noted earlier, value-added models assume that accounting for past achievement is sufficient to account for past resources. If instead as Rothstein (this volume) suggests, teachers are assigned based on unobserved time-varying student characteristics, and these unobserved characteristics (or the process of non-random assignment) change over time, then this might generate "false" instability. Also, if each teacher's value-added did vary considerably across student groups, then year-to-year changes in assignment of students to teachers, combined with differential impacts, would reflect true changes in teacher value-added that are larger than the above intuition alone might suggest. I revisit some of these statistical properties later as many other policy approaches require similar assumptions and have similar statistical properties.

Purposes and Costs of Teacher Value-Added

The policy validity framework outlined by Harris (forthcoming) includes not only statistical validity, a topic covered in the previous two sections, but the purposes of the measures. This means that the above discussion of statistical properties has little meaning without specifying the types of conclusions one wishes to draw. The implicit assumption above is that teacher value-added is intended to create signals that (potentially) provide information about which teachers contribute the most to student achievement. Therefore, teacher value-added tells us how well teachers are performing overall, but tells us nothing about how they might improve.

The definition of the purpose is also relevant to the third piece of the framework—cost. Here, I consider both the standard opportunity cost definition as well as budgetary costs. The costs

of making teacher value-added calculations, or any other statistical adjustments to student tests scores, are small. If we assume the tests will be administered with or without the value-added, then the only additional cost is limited to hiring some expert staff or consultants who are knowledgeable about value-added to make the calculations. An additional cost is explaining the meaning of the calculations to educators—this is far from trivial, as the measures can otherwise be misunderstood or not taken seriously. Also, since the purpose here to hold teachers accountable, it is arguably also necessary to include not just the costs of value-added measures themselves but the costs of the related accountability mechanisms. Harris et al. (2008) show that the budgetary costs of programs teacher merit pay plans can be quite high.[19] Other accountability policies based on teacher value-added, such as use in tenure decisions, would require few resources of any kind.

The fact that there are many ways in which teacher quality measures might be used in policy makes it difficult to generalize about policy validity. However, the framework does suggest that if the goal of education is to raise student achievement, then teacher value-added is a plausibly cost-effective option: it focuses on the outcome of interest (achievement), has some desirable statistical properties for creating signals of effectiveness, and has policy uses that involve little cost.

Policy Validity of Teacher Credentials

To make any fair judgment about teacher value-added for teacher accountability, it is necessary to compare it with other policy options for improving the quality of instruction. As Harris (2008a) points out, the number of possible options (or what he calls "far substitutes") is quite large, so I focus on several options that are often discussed in the context of the value-added debate (teacher credentials) as well as others that represent alternative uses of student test scores.

[19] Harris et al. (2008) describe these as largely budgetary, rather than opportunity, costs because merit pay is a cash transfer, unless it changes either the types of teachers who enter the profession or the amount of effort and time they exert. Such changes are possible, but evidence remains unclear.

One of the most widespread policy traditions for improving teaching is to reward credentials—experience, certification, and formal education. As Harris (forthcoming-a) discusses, there may be patterns such that teachers with particular credentials tend to have higher value-added. In this sense, teacher credentials may be useful as signals of teacher effectiveness. One of the key difficulties in comparing the policy validity of teacher value-added with that of teacher credentials, however, is that credentials also serve as a path to improvement. For example, teachers who obtain a master's degree could potentially raise their value-added as a result.

It important to distinguish between two types of teacher credentials: those that vary over time and those that are fixed. Teacher personality is an example of a relatively fixed characteristic and is often measured in teacher selection instruments such as the Teacher Perceiver. Undergraduate education is another example because very few teachers are in the classroom full time before they have their degrees. Other forms of teacher education, such as graduate training and professional development, change over time. The distinction between fixed and time-varying credentials is important partly because it highlights what can be learned about the policy validity of different types of measures. For a characteristic that is fixed in nature, or one that might vary but is only measured at a single point in time in a particular data set (e.g., undergraduate education), we can only hope to learn whether the measure is a good signal of teacher effectiveness. We cannot know in this case whether the quality of the signal is due to some unmeasured characteristic of teachers that is correlated with the measured characteristic, or whether improving one's standing on the fixed measure actually causes teacher improvement.[20] In contrast, it is easier to determine the

[20] In some ways, the distinction between fixed and time-varying credentials reiterates the distinction made earlier between signals and improvement, but there is a subtle difference. Signaling and improvement have to do with the function that the measures serve, whereas the fixed versus time-varying distinction has to do with the type of data that are available to the researcher. Credentials that are fixed in the data can only be used to study the usefulness of the

causal effects of alterable and time-varying credentials, such as teacher experience and professional development, because individual teachers can be compared before and after the change takes place, using the VAM-P method mentioned above. Value-added models are useful for identifying the causal impacts of time-varying teacher credentials for the same types of reasons they are useful for accountability: they account for selection bias.[21]

Based on Harris and Sass (2007a), I am aware of twenty-eight studies of the effects of teacher education and experience on teachers' contributions to student achievement, using either the gain score, value-added, or experimental methods. Table 1 summarizes the results from these studies, dividing them into two categories based on the methods used. For reasons explained by Harris and Sass (2007a), as well as above in the discussion of value-added assumptions, the value-added and related types of studies are probably more valid than the gain score studies.[22] Note that the numbers in the table add to a number considerably larger than 28 because many of the studies have estimates of more than one teacher credential.

Some studies find a positive and statistically significant relationship between the teacher credential and teacher effectiveness, as indicated in the Positive/Significant category. Other studies find either an insignificant relationship or (rarely) a negative and significant one, which are indicated by Insignificant/Negative. Note that only one of the studies (Harris & Sass, 2007a) includes all of the teacher credentials in Table 1.

---

measures as teacher quality signals, whereas time-varying credentials can be used to study both signaling and improvement. Some examples of this distinction are given in the discussion later in this chapter.

[21] The first form of selection bias—the non-random assignment of students to teachers—was already discussed above. Harris and Sass (2007a) describe a second form which involves the non-random assignment of teachers to credentials.

[22] Table 1 includes the studies together with a very small number of related studies that address the issues of nonrandom selection using data where students and teachers are actually or apparently randomly assigned to one another (these address only one form of selection bias).

Table 1
Summary Results of Value-Added and Earlier Related Studies
(based on review by Harris & Sass (2007a)

Teacher Credentials	*Gain Score Studies*		*Value-Added or Related*	
	Pos/Sign	*Insignif., Neg.*	*Pos/Sign*	*Insignif., Neg.*
Undergraduate	5	4	1	2
Graduate	3	10	3	6
Prof. Develop.	0	1	2	1
Experience	7	8	8	1
Test score	5	2	1	1

Most measures of formal teacher education, especially graduate education, appear unrelated to teacher value-added. In the gain scores studies, 8 of the 23 estimates of the effects of teacher education (undergraduate, graduate, and professional development) suggest that some aspect of teacher education is positively associated with teacher effectiveness. The same finding holds for 6 of the 15 value-added or related types of estimates that have studied teacher education. Most of the remaining studies find statistically insignificant associations between education and teacher effectiveness. Harris and Sass (2007a) provide evidence that certain types of teacher professional development (those providing pedagogical content knowledge) lead to improvement in teacher effectiveness.

Teacher experience is consistently positively associated with teacher effectiveness, at least for the first several years. Roughly half of the gain score studies found a positive effect of teacher experience. The effects are overwhelmingly positive in the value-added and related studies, making teacher experience the characteristic that is most clearly related to teacher effectiveness. These results for teacher experience are consistent with evidence on worker experience in other occupations (Harris & Rutledge, forthcoming). This suggests that teachers, as well as other

workers, learn not only through formal coursework, but also by doing—through their own trial and error.

Teacher test scores are inconsistently associated with teacher value-added. The gain score studies in Table 1 suggest that teacher test scores are consistently positively related with teacher effectiveness. Only two studies have considered teacher test scores with value-added and related methods, but these have yielded more mixed results. Clotfelter, Ladd, and Vigdor (2005) find a positive relationship, whereas Harris and Sass (2007a) find no effect.[23]

Various forms of teacher certification, including NBPTS, are inconsistently associated with teacher value-added (Clotfelter et al., 2005; Goldhaber and Anthony, forthcoming; Harris & Sass, 2007c). A recent extensive review of this evidence has concluded that National Board teachers have higher value-added than others (National Research Council, 2008). NBPTS is an especially interesting credential, however, because it highlights clearly the distinction between the signaling and improvement purposes. The above studies of NBPTS consider not only whether NBPTS is a good signal of value-added, but whether the process of certification increases teacher value-added. While improvement is arguably not the main purpose of NBPTS, it is plausible that such impacts might arise because NBPTS involves over 200 hours of work by teachers, more than many professional development programs. None of the studies suggest, however, that NBPTS has any impact on value-added.

Costs of Credentials

The most costly teacher quality measure is almost inarguably the master's degree in teacher education, which involves nearly a thousand hours of teacher time spent in class and completing

[23] This may be because the researchers in this study controlled for a wide variety of other factors such as coursework. If teacher candidates with greater cognitive ability are more likely to take certain types of college courses, then this may make the effect of cognitive ability look smaller than it is.

assignments.[24] At $20 per hour, the degree costs at least $20,000 in teacher time alone. This time commitment is five times as long as the time commitment of NBPTS certification and perhaps one hundred times larger than some professional development programs.[25] And these figures ignore the costs of the programs themselves—faculty salaries, university classroom space, and so on. If these were added, the direct costs would only grow.

When the credentials are used as the basis of compensation programs, as is typically the case in public (and most private) schools, the costs just listed may be dwarfed by the budgetary costs of additional salaries. If a teacher with a master's degree earns $3,000 more per year than a teacher without the degree, and the teacher stays for twenty years, this could cost the school district $60,000 over the teacher's career—three times more than the costs of teacher time just mentioned.

Teacher Value-Added Versus the Alternatives

While the arguments for teacher value-added are often framed in terms of a comparison with the teacher credential strategy, the more obvious alternatives to teacher value-added are other uses of student tests scores. In this section, I consider school value-added and formative uses of student assessments.[26] This is followed by a comparison of teacher value-added accountability with the credentialing policy and with the other two uses of student test scores.

Policy Validity of School Value-Added

The same general method described above for teacher value-added can be used to measure school value-added and has several advantages. First, school administration and teacher teamwork

[24] This calculation was made as follows: suppose that master's degree requires ten semester-long courses, each of which meets three hours per week for fifteen weeks and requires an equal amount of time outside the classroom: 10 courses x 15 weeks x 6 hours = 900 hours.

[25] Harris and Sass (2007a) report that NBPTS certification requires roughly two hundred hours of work. Professional development programs vary widely.

[26] The words "formative uses of student assessments" reflect the fact that student tests can be designed as "formative assessments," but state standardized tests do not fall into that category. Instead, we are talking here about using tests that are designed to be summative, but which could be put to formative uses.

are captured as part of the calculation, so we need no longer make Assumption #1. Further, because administration and teamwork are captured in school value-added, they can be rewarded. Moving from the teacher to the school level is another way, in addition to comparing teachers across schools, to avoid pitting teachers against one another.

An additional advantage in terms of statistical validity is that there are roughly 10 times as many students per school as per teacher. This goes far toward addressing the imprecision problem with teacher value-added. Note also that school value-added measures appear to be less sensitive to violations of the assumption regarding test scaling (compare the results in Ballou (this volume) who studies teachers with those of Briggs, Weeks, & Wiley (this volume) who studies schools).

Another problem with teacher value-added is that it can be calculated only for a small percentage of teachers—those who teach for several consecutive years in tested grades and subjects. School value-added solves part of this problem, e.g., by still allowing us to measure value-added even for teachers who teach tested grades for only a year or two. It does not solve other aspects of this problem. For example, the gym and music teachers will still contribute little if anything to student achievement and this is no less true with school value-added.

There are two important disadvantages, however. School value-added accountability is subject to the free rider problem. If the whole school is rewarded or punished based on school value-added, then the incentives for effective performance, both within the classroom and in teamwork, may be weak. On the other hand, there is evidence noted above that principals evaluations of teachers are correlated with teacher value-added and there is anecdotal evidence that "everyone knows" who the high-performers are. To the degree that this is true, the school-level incentives could create considerable pressure for low-performing teachers to improve and for principals to hire and retain those who contribute more to school value-added.

Also, measuring school value-added is not as simple as it might seem. With teacher value-added, we account for the selection bias of students to teachers by the fact that students switch teachers regularly and because these changes are essentially required by the basic structure of schooling, i.e., with rare exceptions, the teacher in charge of any given subject changes at the end of each year and this is true for nearly all students, so there is no reason to be concerned that only certain types of student changes teachers. School value-added is identified from movement across schools. While many of these moves are also structurally required by educational policy (i.e., the movement of students from elementary to middle school), a large proportion also occur through moves between schools during the school year, which are clearly non-random. Also, students change schools much less frequently than they change schools overall, so that the large number of student observations contributes less than proportionally to statistical power.[27]

School value-added would almost certainly be a more accurate measure of school contributions to student achievement than the current federal and state accountability systems that reward only the level of proficiency and therefore do not account for the large role of family and community factors. While actual school performance is embedded within such measures, the fact that it is confounded with other, perhaps more powerful forces, means that any school effort to improve student test scores is much less likely to show up in higher school grades. Instead, such systems primarily reward schools for attracting students from advantaged backgrounds, and pushing out students from less advantaged backgrounds. The so-called "growth models" approved by the U.S. Department of Education in some pilot states do little to fix the problem. In short, these models measure whether students are learning fast enough that they could eventually reach

[27] Some have also questioned whether the selection bias problem can be accounted for in school value-added where there is so few students changing schools in any given year; however, note that *all* students eventually change schools when they switch move from elementary to middle school, and from middle to high school. Further, the schools are non-nested so that middle schools have students from multiple elementary schools and so on.

proficiency. This means that schools serving low-performing student are expected to get these students to learn at a *faster* rate than high-performing. This is unrealistic and creates perverse incentives as the proficiency model itself.

Policy Validity of Formative Data Uses

Another quite different use of test scores would involves giving the student data, including specific test topics (sometimes called "strands") to teachers without calculating teacher or teacher value-added. While state standardized tests are not "formative assessments" in the way this term is typically used, using the tests in this way does constitute a "formative use." School districts are increasingly using state tests this way, some through the adoption of additional "quarterly assessments" that measure student progress throughout the school year (Burch and Hayes, 2007).

The advantage of this approach is that it provides useful information to teachers about how they and their students are doing, information specific enough to help teachers improve. Teacher and school value-added, in contrast, provide only signals, which are important for creating incentives but insufficient to drive improvement. This formative use of the data is also potentially inexpensive if it involves only providing state standardized tests data in disaggregated form rather than additional tests such as quarterly assessments. As with other uses of student test scores, providing professional development to teachers to help them interpret and respond to the test results is likely to be an important and more costly part of the process.

Teacher Value-Added Versus the Alternatives

The first comparison of interest is between teacher value-added accountability and teacher credentials. The evidence presented in Table 1 is indicative of the widespread perception that teacher credentials neither signal teacher effectiveness nor improve it, and this partly explains why a focus on teacher performance—particularly, contribution to student test scores—have generated

so much interest. If we are interested in maximizing student achievement, then measuring teachers' contributions directly, rather than relying on indirect signals of effectiveness such as credentials, would seem like a better approach.

But the comparison is more difficult than it appears. The first difficulty that arises is that many teacher credentials serve both signaling and improvement purposes. To the degree that university degrees and certification can be viewed as signals of teacher performance that are explicitly rewarded through hiring decisions and the single salary schedule, it appears that teacher value-added stacks up well compared with teacher credentials in the policy validity framework. However, teacher credentials also provide a path to improvement that value-added does not. While one could argue that improvement is difficult, and that the key to improving teaching is make sure that the better teachers are hired and retained, it is hard to argue that schools do not also need to create a system and culture of improvement, or that training options are unnecessary to facilitate such improvement. In this regard, teacher value-added and teacher credentials serve different purposes.

As noted earlier, signals of effectiveness and paths to improvement are interrelated with one another. Signals can be used in accountability to provide incentives that can induce teachers to seek out opportunities to improve. Teacher value-added and credentials are specifically interrelated in that, under the present system, the motivation to obtain credentials offers little motivation to genuinely improve. Rather, teachers have an incentive to do as little as possible to obtain their credentials. Within the context of a university course or group professional development program, this incentive is hardly conducive to genuine improvement. In *How to Succeed in School Without Really Learning,* Labaree (1997) argues that students' efforts to make high marks makes the entire education system worse, by focusing student attention on getting good grades rather than learning

the material. It is reasonable to expect that this same phenomenon applies to teacher education, especially graduate education, where a large percentage of teachers take university courses mainly because they are required to do so in order to move into school administration or to obtain a higher salary. Thus, one reason the credentials may seem largely unrelated to teacher value-added (see Table 1) is that teachers are getting less out of the credentials than they would if the incentives were set up differently. If teachers sought out credentials on their own, in order to improve their performance (e.g., value-added), then they would not only be more likely to seek out the best credentials, but also more likely to put forth the kind of effort that would make the credentials useful. Thus, the fact that credentials seem unrelated to teacher value-added is not a reason to eliminate credentials, but it is a reason to reform the incentive structure that drives them.

Perhaps the more direct comparisons come from other non-mutually exclusive uses of student achievement scores. Since teacher value-added accountability is arguably the most controversial policy under consideration here, let us consider a situation in which intensive school-level accountability is already in place and teachers already have access to strand or topic-level scores. In this situation, the best case scenario is that teacher value-added accountability brings all the benefits of its advocates, eliminating the free rider problem that remains with school-level incentives and indirectly improving teacher credentials.

A worst case scenario is that teacher value-added accountability would reinforce all the negative unintended consequences of the current system, turning education into one large game of teachers pressuring principals to give them the students who yield high teacher value-added scores and teachers instructing students primarily in how to answer particular types of test questions rather than imparting genuine long-term learning. This worst case scenario is plausible if the value-added

measures really have low statistical validity and reflect behaviors that are unrelated to true performance.

A middle ground between these extremes is that teacher value-added might turn out to be superfluous if these alternatives were adopted. If school-level incentives already provide significant pressure on teachers to improve and if the achievement data were provided to teachers in a way to facilitate improvement, then have no additional impact. McCaffrey and Hamilton (2007) provide some evidence that this middle ground is likely. Studying samples of school principals who recently received information about their teachers' value-added, they found that most principals did not use the information to change their decision-making. The possible impact of teacher value-added accountability is therefore far from clear.

Conclusion

A great deal of attention has been paid recently to the statistical assumptions of value-added models, and many of the most important papers are contained in the present volume. It is intuitively plausible, given that teaching is "loose coupled," that the effects of administration and teacher teamwork are small relative to the teachers' own direct impacts on their own students. I am aware of no convincing evidence on this (Assumption #1) or on whether previous achievement is sufficient to account for previous school resources (Assumption #2). The idea that student fixed effects might account for the non-random assignment of students to teachers (Assumption #3) is more clearly rejected (Kane & Staiger, 2008; Rothstein, this volume). The assumption that achievement tests are interval-scaled (Assumption #4) is also apparently rejected (Ballou, this volume), though it does appear reasonable to ignore variation in teacher impacts across student types (Assumption #5; McCaffrey, Sass, & Lockwood, this volume).

Violations in the assumptions (especially Assumption #2) may explain why Rothstein (this volume) finds that current teacher assignment predicts past achievement gains (Rothstein, this volume). This is problematic, as is the fact that teacher value-added measures are so imprecise that it is only possible to make clear distinctions between the very highest and very lowest level of teacher value-added by traditional statistical standards (Ballou, 2002). This imprecision partly explains why teacher value-added is so unstable over time (Betts & Koedel, 2007), although evidence from McCaffrey, Sass, & Lockwood (this volume) suggests that this instability problem may be largely solvable with adjustments for measurement error.

Notwithstanding all of its problems, Kane and Staiger (2008) find that some value-added models can replicate teacher performance when teachers and students are randomly assigned. This evidence might not be as supportive of value-added as they seem, however. The school principals in the Kane and Staiger study, like some of those studied by Monk (1987), may have been assigning teachers in essentially random ways before the study was conducted. The real concern is what happens in the large percentage of schools where assignment is non-random and where teacher value-added measures are most likely to be biased measures of teachers actual performance. School administrators might well change the way they assign teachers in response to teacher (or school) value-added accountability.

There is also evidence that teacher value-added is positively correlated with principals' own confidential assessments of teachers (Harris & Sass, 2007c; Jacob & Lefgren, 2005). The problem in these studies is that there is insufficient data about what principals know about the value-added of their teachers. Their confidential assessments may partly reflect value-added-like information that they already have. Harris and Sass try to account for this problem by gathering information about how much weight they give to student achievement in their assessments of teachers, but they

have insufficient evidence about how much information each principal has and how they use it to inform their views.

For all its problems, teacher value-added seems to stack up well compared with teacher credentials for the purpose of signaling teacher effectiveness. This is hardly surprising given that assumed goal of education here is to raise student achievement. We would certainly expect that the achievement scores themselves would provide a better indication of teachers' contribution to achievement than, for example, whether teachers participated in particular training programs. But, even if we accept that raising achievement is the only goal, this is still an unfair comparison because credentials serve another important purpose in the educational system—to provide a path to improvement. This is less a defense of the credentialing status quo than of the need for credentials such as formal education. Teacher value-added accountability could help improve credentials by limiting participation to those who are most serious about improving their skills and reduce the direct and indirect costs among those teachers who would otherwise choose not to obtain the degrees.

It is still not a sure bet, however, that teacher value-added accountability would improve student achievement. School value-added accountability could create substantial pressure on schools that may filter down to individual teachers and formative uses of the data could translate these pressures into improved teacher-level performance. Teacher value-added accountability might add nothing new to the mix, as suggested by McCaffrey and Hamilton (2007) who find that school principals pay greater attention to the school accountability information that were already being provided with, and held accountable for. Moreover, school-level incentives could be more productive in facilitating cooperation and collaboration among teachers.

Unfortunately, we know very little about the potential of value-added accountability to improve student achievement. One the most important future steps is to carry additional experiments, especially experiments in which the random assignment is clearly contrasted with situations where students are highly non-randomly assigned. If value-added performs seems to have high statistical validity even in these studies, then this would suggest that it may not matter how principals assign teachers to students and one of the major impediments to value-added would be removed. More realistically, it is likely that the value-added will continue to have some problems in the area of statistical validity and reliability and the issue will come down to whether the measures themselves are better than the alternatives.

Given both the plausibility of teacher value-added as a tool in accountability, and the many unknowns about its true impacts, I have argued elsewhere that the appropriate policy response is to provide funds to both facilitate district experimentation with reasonable policy options and to make sure that those policies are rigorously evaluated (Harris, 2008b). The need to evaluate existing teacher value-added accountability policies—as well as alternatives such as school value-added and formative data uses—is in fact the most important step that researchers need to take. All the talk of assumptions, statistical validity, and perverse incentives will be moot in the end if the impacts of teacher value-added accountability are less positive than the alternatives.

References

Aaronson, D., Barrow, L., Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25: 95–135.

Ballou, Dale (2002). Sizing up test scores. *Education Next* 2(2), 10-15.

Ballou (this volume). Test scaling and value-added measurement. *Education Finance and Policy*.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, 52, 113–121.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, 42, 231–268.

Briggs, D., Week, J. & Wiley, E. (this volume). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*.

Burch, P. & Hayes, T. (2007). *Accountability for sale: The K-12 testing industry, district contracting, and NCLB.* Madison, WI: University of Wisconsin.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2005). *Teacher-student matching and the assessment of teacher effectiveness.* Unpublished manuscript, Duke University, Durham, NC.

Coleman, J. (1966). *Equality of educational opportunity* (Report OE-38000). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Evans, H. (this volume). Value-added in English Schools. *Education Finance and Policy*.

Feng, L. (2005). *Hire today, gone tomorrow: The determinants of attrition among public school teachers*. MPRA Paper 589, University Library of Munich, Germany.

Figlio, D.N. (forthcoming). Testing, crime, and punishment. *Journal of Public Economics*.

Figlio, D.N. & Kenny, L.W. (2007). Individual teacher incentives and student performance, *Journal of Public Economics*, 91, 901-914.

Fryer, R.G. & Levitt,S.D. (2004). Understanding the Black-White Test Score Gap in the First Two Years of School. *Review of Economics and Statistics* 86(2), 447-64.

Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education*, 59, 185-198.

Goldhaber, D., & Anthony, E. (in press). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics.*

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper 2006–01). Washington, DC: Brookings Institution.

Hanushek, E. A. (1979). Conceptual and empirical issues in estimating educational production function issues. *Journal of Human Resources, 14,* 351–388.

Harris, D. N. (2007). High flying schools, student disadvantage and the logic of NCLB. *American Journal of Education, 113,* 367–394.

Harris, D. N. (2008a). *New benchmarks for interpreting effects sizes: Combining effects with costs.* Unpublished manuscript, University of Wisconsin at Madison.

Harris, D.N. (2008b). Breaking the logjam on teacher value-added. Education Week. Downloaded May 10, 2008 from www.edweek.org/ew/articles/2008/06/18/42harris-com_web.h27.html.

Harris, D. N. (forthcoming-a). Education production functions: Concepts. In B. McGaw, P. L. Peterson, & E. Baker (Eds.), *International encyclopedia of education*. Oxford, UK: Elsevier.

Harris, D.N. (forthcoming-b). The policy uses and "policy validity" of value-added and other teacher quality measures. In D. H. Gitomer (Ed.), *Measurement Issues and the Assessment for Teacher Quality*. Thousand Oaks, CA: SAGE Publications.

Harris, D. N., & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education, 112*, 209–238.

Harris, D. N., & Rutledge, S. (forthcoming). Models and predictors of teacher effectiveness: A review of the evidence with lessons from (and for) other occupations. *Teachers College Record*.

Harris, D. N., Rutledge, S., Ingle, W., & Thompson, C. (2006, April). *Mix and match: What principals look for when hiring teachers.* Paper presented at the annual meeting of the American Education Research Association, San Francisco.

Harris, D. N., & Sass, T. (2005). *Value-added models and the measurement of teacher quality*. Paper presented at the annual conference of the American Education Finance Association, Louisville, KY.

Harris, D.N. & Sass, T. (2007a). *Teacher training, teacher quality, and student achievement.* National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper #3. Washington, DC: Urban Institute.

Harris, D.N. and Sass, T. (2007b). *The effects of NBPTS-certified teachers on student achievement*. National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper #4. Washington, DC: Urban Institute.

Harris, D. N., & Sass, T. (2007c). *What makes a good teacher and who can tell?* Paper presented at the summer workshop of the National Bureau of Economic Research, Cambridge, MA.

Harris, D. N., Taylor, L., Albee, A., Ingle, W. K., & McDonald, L. (2008, January). *The resource cost of standards, assessments and accountability.* Paper presented at the National Academy of Sciences Workshop Series on State Standards, Washington, DC.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (NBER Working Paper #11463). Cambridge, MA: National Bureau of Economic Research.

Kane, T.J. & Staiger, D.O. (2001). *Improving school accountability measures* (NBER Working Paper #8156). Cambridge, MA: National Bureau of Economic Research.

Kane, T.J. & Staiger, D.O. (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures" Journal of Economic Perspectives, Vol. 16, No. 4, pp. 91-114.

Kane, T.J. & Staiger, D.O. (2008). *Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates.* Paper presented at the National Conference on Value-Added Modeling, Madison, WI, April 22-24, 2008.

Koedel, C. & Betts, J.R. (2007). *Re-examining the role of teacher quality in the educational production function.* Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

Labaree, D. (1997). *How to succeed in school without really learning.* New Haven, CT: Yale University Press.

Lee, V.E. & Burkham, D.T. (2002). *Inequality at the Starting Gate*. Washington, DC: Economic Policy Institute.

Levin, H.M. (1991). Cost-effectiveness at a quarter century. In M.W. McLaughlin and D.C. Phillips (Eds.), *Evaluation and education at quarter century* (pp.189-209). Chicago: University of Chicago Press.

Levin, H., & McEwan, P. (2001). *Cost-effectiveness analysis* (2nd ed.). London: Sage.

Lockwood, J.R. & McCaffrey, D. (this volume). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*.

McCaffrey, D., Sass, T., & Lockwood, J.R. (this volume). The intertemporal stability of teacher effects. *Education Finance and Policy.*

McCaffrey, D & Hamilton, L. (2007). *Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project.* Santa Monica, CA: RND Corporation.

Medley, D.M. & Coker, H. (1987). The Accuracy of Principals' Judgments of Teacher Performance. *Journal of Educational Research,* 80(4), 242-247.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student assessment: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal,* 88, 2, 166-187.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R.J., & Cohen, D. (1986). Merit pay and the evaluation problem: why most merit pay plans fail and a few survive. *Harvard Educational Review,* 56, 1-17.

National Research Council (2008). *Assessing Accomplished Teaching: Advanced-Level Certification Programs*. Milton D. Hakel, Judith Anderson Koenig, and Stuart W. Elliott, Editors, Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards, National Research Council.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven, CT: Yale University Press.

Ogbu, J. U. (2003). *Black American students in an affluent suburb: A study of academic disengagement.* Mahwah, NJ: Lawrence Erlbaum.

Peterson, K.D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, 24(2), 311-317.

Peterson, K.D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2d ed.). Thousand Oaks, CA: Corwin Press.

Podgursky, M. & Springer, M. (2007) Teacher performance pay: A Survey. *Journal of Policy Analysis and Management*. 24(4), 909-949.

Rice, J.K. (2002). Cost analysis in education policy research: A comparative analysis across fields of public policy," In Henry M. Levin and Patrick J. McEwan (Eds), *Cost-effectiveness in educational policy* (pp.21-35), Larchmont, NY: Eye on Education.

Rivkin, S. G., Hanushek, E., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73,* 417–458.

Rothstein, J. (this volume). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*.

Rothstein, R. (2004). *Class and Schools*. New York: Teacher's College Press.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVASS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247–256.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113,* F3–F33.