

The Policy Uses and “Policy Validity” of Value-Added and Other Teacher Quality Measures

Douglas N. Harris
University of Wisconsin at Madison

March 17, 2008

Forthcoming in D. H. Gitomer (Ed.), *Measurement Issues and the Assessment of Teacher Quality*. Thousand Oaks, CA: SAGE Publications.

Abstract

In order to improve the quality of the nation’s teachers, the public education system has long relied on requirements and rewards for formal teacher education, experience, and other traits—the “credentials strategy.” However, policymakers and some prominent educators are increasingly embracing a radical overhaul—an “accountability strategy”—that largely ignores these traits and instead rewards teachers’ measured contributions to student results. In this chapter, I review evidence about the statistical validity of a variety of teacher quality measures. I also argue that this alone is not enough to determine how and when teacher quality measures should be used. Instead, I outline the criterion of “policy validity,” which is determined by statistical validity as well as the specific *functions* that the teacher quality measures serve in the education system and the *costs* of producing both the measures and the underlying teacher qualities they represent. Based on this framework and the most recent value-added-based research, I find that neither the traditional credentials strategy nor a simple value-added-based accountability strategy is likely to best address the teacher quality problem. Instead, as in the private sector, a valid policy approach requires using multiple strategies and measures that provide signals of teacher effectiveness and formative and summative assessments that facilitate and encourage improvement. Specifically, the evidence suggests that teachers be rewarded not for their graduate degrees, but for a combination of experience, certain types of professional development, and teacher and school performance. More generally, improving the quality of teachers will require a comprehensive strategy that few current or proposed policies provide.

Acknowledgements: Many of the studies from which the ideas in this chapter have grown were funded by the U.S. Department of Education, Institute for Education Sciences and the National Board for Professional Teaching Standards (NBPTS). In addition to this support, I gratefully acknowledge comments by Robert Floden, Drew Gitomer, David Monk, and Tim Sass and participants in research presentations at the 2007 Educational Testing Service Invitational Conference in San Francisco and at the University of Colorado at Boulder. The views expressed here are solely those of the author who is of course responsible for all remaining errors.

Introduction

The U.S. education system has long tried to maintain and improve the quality of its teaching workforce by requiring and rewarding specific teacher credentials, especially teacher experience and certain types of education. Teachers are prepared in university-based colleges of education that require state government approval. Most of these graduates are then certified to teach, so long as they pass knowledge and/or pedagogy competency tests. Teachers are then hired and compensated based on their years of experience and whether they obtain graduate degrees in education. Not all teachers have high levels of education or experience, but requiring and rewarding these teacher credentials remains the nation's dominant teacher quality strategy.

This traditional "credentials strategy" and the results it has produced have long been criticized.¹ The systems of public education in general, and human resource management in particular, are seen as less efficient than those used in the private sector. Also, the outcomes of these perceived inefficiencies have been seen as inadequate. The concern is not only that the country's low standing in international test score comparisons and perceived slow growth in national test scores (Harris & Herrington, 2006), but that faithfully executing the traditional strategy seems to have no consistent and measurable impact on teachers' effectiveness in raising student test scores. "Education production function" studies by economists have generally found little or no systematic relationship between teacher credentials and student outcomes (Goldhaber & Brewer, 2000; Hanushek, 1986).² Levine (2006) reaches essentially the same conclusion in his more recent study and critique of teacher education. The fact that the traditional strategy seems

to have little positive impact on student test scores reinforces larger concerns about the country's educational competitiveness.

These research findings, combined with new political pressures, have led to calls for an alternative strategy—teacher quality through accountability.³ Two major shifts toward this accountability strategy have occurred in recent years: one in the early 1990s as state governments began to introduce and expand “school report cards” and the second, in 2001, with the passage of the federal *No Child Left Behind* (NCLB) law that expanded student testing and the use of school report cards as the basis for interventions in schools whose students are not making adequate test score progress. While focused on holding schools accountable, rather than individual teachers, these pressures were clearly designed to “trickle down” and influence what teachers do. In many ways, this form of accountability has succeeded in its goals. There is ample evidence that teachers made changes in instructional time and practices in order to meet accountability objectives (e.g., Booher-Jennings, 2005). In states such as Florida that have long used strong accountability systems, the pressures have apparently permeated schools even more deeply, influencing basic human resource management practices (Rutledge, Harris, Ingle, & Thompson, in press) and policies related to student discipline (Figlio, 2003). There is of course considerable debate about whether these changes represent genuine improvements, and many of the same studies showing intended change also indicate unintended consequences (Booher-Jennings, 2005), but there is no questioning that many of the intended effects, such as greater focus on math and reading, have occurred.

The accountability movement has, however, left the traditional credentials strategy largely intact. While efforts have been made to provide alternative routes to

certification, the overwhelming majority of teachers still graduate from university-based teacher education programs, still receive state-sanctioned certification, and still receive compensation based on degrees and experience. Rather than replace the traditional strategy, accountability has simply been layered on top of it.

But perhaps not for long. Many school districts are now experimenting with compensation systems that aim to provide direct incentives for individual teachers, paying them based on the test scores of their own students and therefore implicitly reducing the weight given to university-based teacher education and experience. The federal government is also encouraging these “merit” or “performance” pay approaches through a voluntary pilot program, the Teacher Incentive Fund (TIF), and new bills in Congress promise to expand these new pay systems in persistently low-performing schools. (Full disclosure: I am on the federal technical working group that advises TIF districts on their plans.) There are also proposals to use student test scores as a primary basis for making teacher tenure decisions (Gordon, Kane, & Staiger, 2006). Depending on the outcomes of these proposals, the development, implementation, and adoption of similar merit pay policies could greatly expand and even replace the traditional teacher quality strategy altogether.

Yet, the debate about both the credential and accountability teacher quality strategies has been confused in a number of ways. The first issue is the way in which empirical evidence has been translated into policy recommendations. Information about statistical validity is far from sufficient to make such recommendations. Therefore, in the second section, I outline the criterion of “policy validity,” or the validity of the *use* of a teacher quality measure in education policy. The use of a teacher quality measure is valid

in this policy sense if the weight given to the measure (in determining who becomes a teacher and how teachers develop their instructional practices) is proportional to the statistical validity and costs associated with the measure. This is an admittedly vague definition, especially in comparison to the related cost-effectiveness concepts methods typically used by economists. However, I will show later that it does prove useful in interpreting the research evidence, and that the more typical and concrete cost-effectiveness frameworks are impractical here because of the complex nature of teacher quality and notable gaps in research.

The debate about teacher quality has also been confused by methodological problems and inconsistent findings in the research studies testing the validity of the teacher quality measures. One of the main methodological problems has been the “selection bias” caused by both the non-random assignment of teachers to education and of teachers to students (Harris & Sass, 2006). In the past several years, researchers have begun to develop and use “value-added modeling” to address these problems and more clearly identify the credentials of effective teachers. In the third section, I discuss the basic logic and assumptions of value-added and provide a summary of recent evidence about the credentials of effective teachers from value-added models, as well as some experiments and quasi-experiments. I also compare the results of these studies to the older generations of education production function studies, which remain influential in the current debates about teacher quality strategies. All of the research discussed in this paper focuses on student achievement scores on standardized assessments as the outcome of interest, mainly because there is little evidence about the causal effects of teachers on other important outcomes such as student motivation, creativity, and socialization.

In addition to using value-added to identify the credentials of effective teachers, a second potential use of value-added, consistent with the accountability strategy, is to directly measure the effectiveness of individual teachers. By analogy, the credentials approach amounts to measuring how much rain has fallen by the number of clouds in the sky. The accountability strategy would simply put an empty glass on the sidewalk and measure the rain directly—without getting “lost in the clouds.” In the fourth section, I revisit the assumptions of value-added and show that this rain metaphor is more problematic than it appears at first. The assumptions of value-added are, at best, untested and, at worst, simply wrong, and this has significant implications for using individual teacher value-added measures for accountability. I also summarize recent empirical evidence about the statistical validity of value-added for accountability.

In the final section, I interpret the recent evidence from value-added studies in terms of policy validity. I conclude that neither the credentials strategy nor the accountability strategy alone is the best approach to improving teacher quality. Just as there is little support for the long-standing tradition of giving higher salaries for having graduate degrees, so too is there little support for the other extreme—compensating and firing teachers substantially on the basis of value-added. Instead, what is needed is a coherent and comprehensive strategy that serves all of teacher quality functions well and uses resources effectively and efficiently. I provide some specific suggestions at the very end of the paper.

Introduction to Policy Validity

There is no single way to interpret the empirical evidence about teacher quality measures and translate it into clear policy recommendations. Decisions inevitably turn on difficult trade-offs and judgments driven as much by philosophy as evidence. However, there are some basic elements that are arguably required of any useful decision-making framework related to teacher quality policy. Below, I discuss three such elements—statistical validity, functions, and costs—and combine these into a framework that I refer to as “policy validity.” I begin below by discussing the different general functions and then show how these functions are intertwined with the meaning of statistical validity.

General Function #1: Predicting or “signaling” those teachers who are most likely to be effective. A signal is some quality of a person that indicates that they are likely to help the organization meet its objectives and can therefore be used to predict future behavior and effectiveness. For example, suppose that a researcher observes that teachers with academic degrees from particular university programs tend to be more effective than others. A degree from this institution therefore serves as a signal regarding the likely effectiveness of future candidates.⁴ Such signals could also be useful at earlier stages in the teacher pipeline. For example, leaders of teacher education programs might observe that teachers who have an average grade point average (GPA) in their freshman and sophomore years are more likely to go on to graduate and become successful teachers. Teacher education programs might therefore use GPA as a signal and as a basis for admissions. Of course, to be good signals, there must be some relationship between the signal and the contribution to the organizations’ objectives.

Signals can also be the basis for compensation. As noted earlier, teacher education and experience are used as the basis for teacher compensation in public schools. Using signals in this way, however, can be criticized because compensation occurs *after* services have been rendered—that is, teachers are paid for work they have already done. Thus, it would only seem logical to compensate teachers primarily on the basis of a signal if what the teachers did—how well they performed—were extremely difficult to measure.⁵

General Function #2: Improving teacher effectiveness either through formative assessments or summative assessments tied to incentives. While signals are useful for identifying potentially effective workers for hiring purposes, it is also important that organizations develop the skills and effort of their employees after they are hired. There are two ways that a teacher quality measure can aid in this process. First, the measure might suggest ways in which teachers could improve. For example, if professional development appears to improve teacher effectiveness, then a teacher who is performing poorly in a particular area (e.g., teaching fractions) might be encouraged to pursue additional professional development to improve skills in this content area. This would be considered a “formative” use of the measures.

An alternative use of teacher quality measures is simply to determine who is performing well—a “summative” assessment—in order to identify the teachers for hiring, promotion, additional compensation, or dismissal. While summative assessment does not provide advice to teachers about how to improve, it does provide incentives that might induce teachers seek out paths to improvement or, in the face of repeated low-performance, to leave the profession. That teachers often leave the profession well

before retirement age is often viewed as a problem, but departures of low-performing teachers are likely to improve outcomes if these teachers can be replaced by more effective ones. It is therefore clear that formative and summative assessments are interrelated. There must be paths to improvement as well as incentives, formal or otherwise, to follow those paths.

Multiple measures are also necessary to carry out the functions because a teacher quality measure that is useful as a signal may not be useful to improve effectiveness and vice versa. To be useful as a signal, the measure must be a strong predictor of teacher effectiveness (i.e., explain a high degree of the variance in effectiveness). In contrast, to be useful for improving effectiveness, the measure can have more modest statistical explanatory power, so long as it is alterable. Suppose that teachers with particular personality traits tend to make more effective teachers. In this case, the trait might serve as a useful signal but, if personality traits are essentially fixed or at least difficult to systematically alter, then they are useless for the purpose of improving an individual teacher's effectiveness. It also turns out, as discussed later, that a measure that is valid for formative assessment might not be valid for summative assessments.

The larger point is that whether an estimate is statistically valid, indeed the very meaning of statistical validity, depends on what type of conclusion one is trying to draw. Further, in trying to improve teacher quality, there are many types of conclusions or functions that are of interest.⁶ The multiple functions and meanings of statistical validity help to explain why policy validity is a useful term and concept. Policy validity takes into account the specific types of inferences—signaling and improvement—that are important with regard to teacher quality. It also accounts for an additional factor of great

interest to policymakers—policy costs. Later in this chapter, I discuss how economists conceptualize and measure costs and provide some back-of-the-envelope calculations about the costs of some aspects of teacher quality policies. In the two sections that follow, I discuss value-added as a source of evidence about the validity of different teacher quality measures.

From Education Production Functions to Value-Added

The impetus for the current reconsideration of the credentials strategy for improving teacher quality, as noted earlier, comes partly from evidence from “education production function” (EPF) studies which suggest that some key teacher credentials—education, experience, certification and so on—are not closely related to teacher effectiveness. Researchers have been well aware of the questionable validity of these studies, and resulting potential for “selection bias,” caused by the reliance on data from a single point in time. Nevertheless, these were some of the best studies available at the time and researchers and policymakers were right to take them seriously. Now that research methods and data have advanced, however, it is important to revisit the selection bias problems and show how new research methods potentially address them.

There are two main parts to the selection bias problem in this context. First, teachers are non-randomly assigned to students. For example, some teachers are systematically assigned to teach students who have lower initial achievement. Using data from a single point in time, these teachers will appear less effective. This problem can be partly corrected by controlling for student socio-economic status (SES) with measures such as student race and income, but recent evidence suggests that variation in initial

student achievement is only partly captured by these SES measures (Harris & Sass, 2005).

The second problem is that teachers are non-randomly assigned to some of their credentials. Most notably, some teachers may be systematically more likely than others to obtain additional education. If the least effective teachers obtain more education to address their deficiencies then, even if teacher education helps, it may appear falsely that teacher education makes teachers *worse*. Alternatively, better teachers might feel more confident about their potential ability to obtain more education. This would have the opposite effect, making it appear that education made teachers better when, in fact, teachers with more education were better to start with. The net effect of these different forms of selection bias is difficult to determine, but accounting for them is clearly important.

With colleague Tim Sass, I have also described a second generation of EPF studies that address non-random selection by controlling for a single previous student test score (Harris & Sass, 2006). Thus, if differences across students, such as their access to school resources in the past, affect their propensity for learning in the future, then controlling for a previous test score should account for the differences in past inputs. Put differently, with two scores, we can focus on the change in learning over a very specific time period, one or two years, and more reasonably assume that the change in test scores are due to what happened in the school during that time period.

The above “gain score” model partly addresses the non-random assignment of teachers to students as long as all students who have the same initial test score have an equal probability of making large gains in the next year. But, while this is a more

reasonable assumption than the one made in the point-in-time EPF model, it is still problematic. Some students who have the same test score at a point in time may also have different expected rates of learning growth. For example, consider a student who was far behind when starting kindergarten, but who has learned at a fast rate, compared with a student who started kindergarten with a high level of achievement but has been learning at a slow rate. These two students might have the same score at a point in time, but the expected rate of growth in the future is apparently higher for the first student and teachers assigned to that student will be at an advantage. Also, even if the students have the same expected rate of learning, the gain score model still fails to address the second selection bias problem—non-random assignment of teachers to teacher preparation. Therefore, the results from the gain score models are still potentially invalid, in all senses of the term.

Value-added models have the potential to address both selection bias problems. Intuitively, these models start with the average achievement gain of each teacher's students over several years and then adjust these averages based on how much the teachers' students would have been expected to learn given their growth trajectory over a long period of time as well as the resources those students received in other grades. By including data from multiple years and, in effect, controlling for this adjusted average rate of student learning, the researchers can identify the effect of each teacher by calculating for each teacher the average deviation from the students' expected learning trajectory. Teachers whose students systematically "beat expectations" have above-average value-added. By doing a better job of measuring how much we can expect each student to learn, we can better address the first selection bias problem described above.

Value-added also addresses the non-random assignment of teachers to some forms of teacher education. Just as with the student assignment problem, the teacher assignment problem can be addressed by using each teacher as his/her own control group. Consider a teacher who takes part in a professional development course. With a value-added model, we could measure the teacher's effect on students before the professional development and then measure it again afterwards. If teacher effectiveness improves, controlling for other factors like teacher experience, then it is reasonable to conclude that the professional development was the cause—and this is true even if the teachers were non-randomly assigned to the professional development.⁷ It would therefore appear that value-added is a significant advance in identifying the effects of individual teachers and teacher credentials.

Some of the most important early work on value-added can be found in Hanushek (1979) and Boardman and Murnane (1979). For more recent discussions, which discuss theoretical issues within the context of current data availability, see Harris (in press), Harris and Sass (2005), and Todd and Wolpin (2003).

Some Assumptions of Value-Added

While value-added has some important advantages over traditional EPF and gain score models, this new approach is still based on some important assumptions. I discuss some of the key assumptions below and return later to consider their implications.

*Assumption (1). Differences in the likelihood that students will make achievement gains can be accounted for by taking into account students' past (and future) growth.*⁸ I mentioned earlier that an advantage of value-added is that it accounts for the fact that

teachers teach different types of students who have different propensities to make achievement gains. More specifically, the student's propensity to make achievement gains has to be *fixed*. For example, some students have parents who consistently make them do their homework and expect them to go to college. However, some children might experience a divorce or other change in circumstances that influences his/her ability to learn. While these time-varying changes are not accounted in value-added models, this does not necessarily bias the estimates of individual teacher value-added. The measured effect of a teacher would only be biased if a teacher were systematically assigned to students who have positive or negative "shocks" in their learning propensities (e.g., divorce of parents). Family changes are not uncommon in the overall student population, but such students might not be systematically assigned to certain types of teachers, with specific types of credentials.

Assumption (2). 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. Some researchers, for example, 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 each 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.⁹ The significance of the assumptions about the test scale, as well as the adjustments that might be made to account the assumptions, are currently being explored by a number of researchers, though there is little evidence to report at present.

Assumption (3). Teachers are equally effective with all types of students. A high value-added teacher is one whose students learn at a faster rate than one would expect given their growth rates in other years. These deviations from the expected growth rate are then averaged and adjusted for all of the teacher's students. Because all of the students are averaged together, an implicit assumption is that teachers are equally effective with all groups of students. To see the problem with this, suppose that the opposite were true and that some teachers were effective with slow-achievement-growth students and other teachers were effective with fast-achievement-growth students. Further, suppose that all teachers were assigned only to students with whom they were most effective. In this case, all teachers would appear equally effective.

Now, suppose that some teachers were assigned to students with whom they were ineffective and, as a result, their value-added scores decrease. These same teachers who

were judged effective above will now appear ineffective simply because they were assigned to a different group of students. 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. This example is an extreme case, but illustrates the general problem with assuming that all teachers are equally effective with all students.

The degree to which this assumption poses a problem depends on the type of statistical validity that is of interest. While the differences in the ways that teachers are assigned to students are problematic for the sake of summative assessment, these value-added estimates would still be useful for teachers in trying to improve their performance (formative assessment).

Assumption (4). Student test data are missing at random. Given the complexity implied by the above discussion, it comes as no surprise that 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 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.

The above discussion focuses on the basic nature of value-added models and their underlying assumptions. Below, I discuss some important recent findings regarding value-added. First, I discuss the findings of “value-added modeling for program

evaluation” (VAM-P). While not referred to specifically above, VAM-P is used to identify the correlations and effects of teacher credentials, such as teacher test scores and teacher professional development. I go on to discuss “value-added modeling for accountability” (VAM-A), which is used to identify the effectiveness of each individual teacher (ignoring teacher credentials). I also show later why the assumption violations discussed above are much more significant for VAM-A than VAM-P and therefore why value-added should be used more cautiously in trying to evaluate individual teachers.

VAM-P Research on the Credentials of Effective Teachers

The previous section shows why the results from value-added models are more likely to yield statistically valid estimates of the credentials of effective teachers—for the simple reason that the two forms of non-random selection are addressed with greater care than in EPF and gain score studies. Below, I summarize recent studies that have used value-added to examine the credentials of effective teachers. I also discuss some of the difficulties in identifying such credentials, beyond the questionable assumptions mentioned above.

Before discussing these findings, it is important to distinguish between two types of teacher credentials: those that vary over time and those that are fixed. Above, I discussed teacher personality as an example of a fixed characteristic. 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 analysis, 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.¹⁰ In contrast, it is easier to determine 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.

Findings from VAM-P Regarding the Credentials of Effective Teachers

Based on Harris and Sass (2006), I am aware of 28 studies of the effects of teacher education and experience on teachers' contributions to student achievement, using either the gain score approach, value-added, or experimental methods. Table 1 summarizes the results from these studies, dividing them into two categories, based on the methods used. 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.

Table 1 includes the VAM-P studies together with a very small number of "related" studies that address the issues of non-random selection using data where students and teachers are actually or apparently randomly assigned to one another (these address only one form of selection bias). For the reasons stated in the previous section, there are reasons to trust the validity of the value-added studies more so than the "gain scores" studies shown in the middle column. Some studies find a positive and

statistically significant relationship between the teacher credential and teacher effectiveness, as indicated in the “Pos/Sign” category. Other studies find either an insignificant relationship or (rarely) a negative and significant one, which are indicated by “Insignif., Neg.” Note that only one of the studies (Harris & Sass, 2006) includes all of the teacher credentials in the Table 1.

Table 1

*Summary Results of Value-Added and Earlier Related Studies
(based on review by Harris & Sass (2006))*

<i>Teacher Credentials</i>	<i>Gain Score Studies</i>		<i>Value-Added or Related</i>	
	<i>Pos/Sign</i>	<i>Insignif., Neg.</i>	<i>Pos/Sign</i>	<i>Insignif., Neg.</i>
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

Finding (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.

As discussed below, the fact that most forms of formal teacher education are unrelated to teacher value-added does not mean that the same is true of *all* forms of formal education.

It is also important to emphasize that the measures of teacher education vary considerably across studies. Some look only at whether teachers have degrees from schools of education, while others consider training in particular subject areas and/or consider teacher effectiveness in specific school subjects. Gain score studies matching specific undergraduate majors with specific subjects (e.g., the effects of teachers who were math majors on student math achievement) are more likely to find positive effects than those looking at the level of education or whether the degree came from a school of education, but the results are inconsistent even in these cases. Also, note that nearly all of the studies on formal teacher education focus only on the quality of the signal teacher education provides, not whether it improves teacher effectiveness. (In the case of undergraduate teacher education, this is simply a consequence of the fact that it is not time-varying.)

Wayne and Youngs (2003), in a previous summary of evidence using a broader variety of research methods, conclude that: (1) there is a relationship between teachers' mathematics coursework and student mathematics gains in high school, but no such effects are apparent in elementary grades or other subjects; and (2) mathematics certification is related to students' math scores gains, but there is insufficient evidence with regard to other subjects and grades. However, given the apparent sensitivity of the results to research methods, their inclusion of studies that do not adequately account for non-random selection is somewhat problematic.

Finding (2). There is some evidence that “pedagogical content knowledge” is associated with teacher effectiveness. This finding is based on our analysis of the Florida data (Harris & Sass, 2006) and is only partially evident from the table. We used VAM-P to study the time-varying effects of teacher professional development where it is possible to compare each teacher’s effectiveness before and after the education takes place. Specifically, we found that “content” professional development was in some cases positively associated with teacher value-added.¹¹ I interpret this as pedagogical content knowledge because professional development, unlike some courses in undergraduate education, rarely focuses on content alone, so content-oriented professional development is likely to include a mix of pedagogy and content.

Two related findings are noteworthy. First, we found what appears to be a lagged effect of teacher professional development; that is, it can take several years for professional development to produce higher teacher value-added. This is important as it suggests that studies looking for immediate effects will understate the long-term impact. Second, when studying the signal effect of different types of undergraduate teacher education, we found that undergraduate courses mixing content and pedagogy were sometimes positively associated with teacher value-added.¹² Thus, pedagogical content knowledge appears to be the only case in which a particular type of formal education is useful both as a signal and as a path to improvement.

Finding (3). 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, 2007). This suggests that teachers, as well as other workers, learn not only through formal coursework, but also learn by doing—through their own trial-and-error.

Finding (4). 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. While only two studies have considered teacher test scores with value-added and related methods, these have yielded more mixed results. Clotfelter, Ladd, and Vigdor (2005) find a positive relationship, while Harris and Sass (2006) find no effect.¹³ Both studies focus on college entrance exams (as opposed to certification tests) and on the potential of these scores to serve as valid signals of teacher effectiveness.

Finding (5). Various forms of teacher certification, including NBPTS, are inconsistently associated with teacher value-added. The research my colleagues and I have produced in Florida have not focused on state certification for various reasons, but we have studied a new type of teacher certification: the National Board for Professional Teaching Standards (NBPTS) (Harris & Sass, 2007a). In short, we find that NBPTS certification inconsistently identifies more effective teachers. Results from other similar studies find that NBPTS teachers are more effective than others and more so than those who attempt NBPTS certification but fail to pass (Goldhaber & Anthony, in press), but these, too, are inconsistent.

NBPTS also involves a substantial amount of time in the preparation of materials and studying for the required tests. Therefore, it is also plausible that NBPTS, whether or

not it is an accurate signal of teacher value-added, still improve teacher effectiveness. In this case, the results are more consistent: none of the value-added studies suggest that teachers improve as a result of going through the process. This is not necessarily a criticism of NBPTS, because directly improving teaching is not the main purpose of the certification, but it is noteworthy because it highlights how a single measure might be useful for one purpose but not another.

While setting aside some important methodological issues that arise even in the most advanced studies,¹⁴ this review summarizes the latest research from value-added and gain score studies. The next section provides a similar review for a very different type of teacher quality measure.

VAM-A Research on Individual Teacher Effectiveness

The traditional strategies for improving teacher quality focus on credentials such as teacher education and experience, discussed in the previous section. An alternative strategy is to measure teacher effectiveness directly—measuring the rain by putting a glass on the sidewalk, so to speak. Below, I discuss important findings regarding the validity of value-added for accountability (VAM-A) that inform the feasibility and usefulness of this alternative strategy.

Findings about Individual Teacher VAM Measures (VAM-A)

Finding (1). *Value-added varies considerably across teachers.* 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.

This conclusion is important because it suggests that, even though few teacher *credentials* are systematically associated with student learning, the teachers themselves clearly do matter—and do vary. While this conclusion might not seem very surprising, remember that earlier studies by Coleman, Hanushek and others focused only on measurable credentials that ultimately appeared to be unrelated to teacher effectiveness, which led some to conclude that teachers did not generally matter. While it remains difficult to identify specific credentials that are important, these value-added results suggest that students do better—perhaps substantially so—with some teachers compared with others.

The apparent variation in teacher effectiveness is partly driven by problems with the assumptions of value-added as well as issues such as measurement and estimation error. Nevertheless, the general finding that teacher effectiveness varies is perhaps obvious enough from other types of studies. Studies on student “tracking,” for example, suggest that the instruction received by lower-track students is more likely to emphasize memorization compared with the higher order thinking skills emphasized in higher-track classes (Ogbu, 2003). In this case, the quality of instruction appears to vary across classrooms and teachers, so we would expect value-added to vary as well.

Finding (2). 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 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 (2007b), 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 literature that specifically addresses the relationship between subjective and objective assessments of school teachers. Some studies have examined the relationship between teachers' student 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” type 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 mid-sized school district and reach two main conclusions: (1) there is a positive correlation between the subjective and objective measures; and (2) that this correlation holds even after controlling for teacher experience and education levels, which are currently the primary bases for determining teacher compensation. We find similar results from our analysis of a separate mid-sized school district in Florida (Harris & Sass, 2007b). While the comparison of principal evaluations of teachers with teacher value-added measures cannot be viewed as a validity check per se, it does suggest that value-added measures provide useful information.

Based on the above findings—that teacher effectiveness varies and that this measure is correlated with other credible measures—the news on value-added reinforces the potential use of value-added for accountability. This is not the case with two other findings below.

Finding (3). Teacher value-added scores are imprecise. There are several sources of error that make estimates of teacher value-added imprecise. This a natural outgrowth of the fact that value-added focuses on changes in student achievement over time, which compounds the measurement error in the achievement scores. Also, by their nature, value-added models involve estimating large number of parameters (one performance measure per teacher) with relatively few observations per teacher, reducing reliability. Kane and Staiger (2001) provide an excellent discussion of the types of errors involved with VAM estimation and their impact on grade-level school performance measures. The problems they describe are worse when data are disaggregated to the individual teachers, as is the case with the VAM-A models of interest here.

Finding (4). Individual teacher value-added changes considerably over time—i.e., the measures are not “stable.” Intuitively, we would expect that the actual effectiveness of each individual teacher changes little from year to year. Teachers might gradually improve over time, as suggested by the earlier discussion of evidence on teacher experience,¹⁵ but it is unlikely that they will jump from the bottom to the top of the performance distribution. It is even less likely that teacher rankings on value-added should drop significantly from year to year.

Yet, the results suggest that teacher value-added estimates are indeed unstable. Koedel and Betts (2005) find 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—many dramatically so. It is intuitively implausible that actual teacher effectiveness is this erratic over time. The low reliability and stability of value-added measures therefore reinforce the need to proceed with some caution in using value-added for accountability. This instability may or may not be a severe problem in VAM-P. If the instability is due entirely to random fluctuations, then this probably introduces no bias in the estimates of teacher credential effects, but only makes it more likely that estimates will be statistically insignificant.

Revisiting the Assumptions

Earlier, I discussed the assumptions of value-added, which apply to both VAM-A and VAM-P. While violations of these assumptions are problematic for both model types, the assumptions of any value-added model are more problematic for VAM-A. There are two reasons for this: First, achieving valid estimates is always harder in VAM-A than VAM-P because there are fewer student test score observations to work with for each relevant estimate. Second, many violations of the assumptions may arise only for a small number of teachers or occur randomly so that the problems “wash out” when considering the entire sample of teachers. But when looking at the value-added of individual teachers, this means that the value-added scores for some, perhaps many teachers, will be biased. For example, most teachers may be equally effective with all students or systematically assigned to students with whom they are most effective (see Assumption (3)), but what about the other teachers? What is true on the average will not be true everywhere and

this means that a large number of teachers are likely to be inaccurately evaluated when using value-added measures for accountability.

The importance of the assumptions, as well as the earlier problematic empirical results, might seem to suggest that VAM-A should not be used. These limitations must be balanced, however, against the main advantages of VAM-A—that it provides a more direct measure of effectiveness and sends a message about the priorities of the school system. Another way to see the difference between VAM-A and VAM-P measures is to observe that while the “noise” tends to wash out in VAM-P and yields fairly precise estimates of the effects of teacher credentials, these effects are small and explain little of the total variation value-added; in contrast, the VAM-A measures are imprecise, but they imprecisely measure what is of greatest interest.

Also, all of the assumptions of VAM-A apply to other methods of using student test scores to evaluate teachers. For example, suppose we were to evaluate teachers simply on the level of achievement at the end of the school year, or even on the simple gains from the previous year. In these cases, all the assumptions discussed above—that teachers are equally effective with all types of students, etc.—are still required. These simpler and more common uses of student test scores also require an additional assumption: Instead of Assumption (1), evaluating teachers based on test score level or gains requires the far less plausible assumption that that *there are no differences in the likelihood that students will make achievement gains* and therefore there is no need to make any adjustments to account for non-school factors. This is a fundamental flaw as it means that these measures clearly attribute to the school, and to teachers, causes of low achievement that are clearly outside of school control—especially family factors that are

well known and powerful influences over student achievement (e.g., Coleman, 1966; Harris, 2007). This same assumption is commonly made with “school report cards” and in defining school success under the federal NCLB, even in states that use “growth curve” models, though extensive discussion of this topic is outside the scope of the present study.¹⁶

This leads to two important conclusions. First, the assumptions of VAM-A may be somewhat problematic, but there are clearly fewer problematic assumptions involved with VAM-A than with more common approaches to using test scores. This highlights the fact that VAM-A, and its assumptions, should be judged not by whether it meets all the desirable statistical properties—it does not and no amount of research will change this. Rather, VAM-A should be judged relative to the realistic alternatives. When compared with other ways of using student test scores to assess teacher performance, VAM-A stacks up reasonably well.

The other alternative—using VAM-P to identify the credentials of effective teachers—has problems of its own. The fact that VAM-P allows random deviations from the assumptions to wash out when measuring the credentials of effective teachers is an advantage, but these credentials explain little of the variation in teacher value-added and therefore are only modestly helpful in identifying the most effective teachers. These are inherent trade-offs that policymakers must take into account when deciding how to use the various measures.

Policy Validity: Interpreting the Evidence

One possible way to interpret the above evidence on teacher quality measures is through the economics method of cost-effectiveness analysis; that is, measuring the effects and costs of various options and recommending the set of options that provides the greatest “bang for the buck.” This framework, however, is impractical for trying to draw policy conclusions from evidence about teacher quality measures. First, there is no agreement on the specific effects and costs of any of the measures. In the case of effects, the previous section shows that this is a result of the variation in results across studies. While it might be possible to establish reasonable ranges for the effects, there is almost no evidence about the costs of the measures with which to compare them. Second, cost-effectiveness analysis tends to assume that the policy options are independent of one another; yet, as I show below, there are multiple functions that the measures play in the process of improving teacher quality and that these are interrelated with one another. In short, teacher quality policy is just too complex to be boiled down to a few simple numbers. Policy validity therefore incorporates the cost-effectiveness concept with the specific complex issues that arise in improving teacher quality, taking into account the limits in the research evidence.

After discussing specific policy functions that might be served by teacher quality measures, I present some preliminary evidence about the costs of the measures. Finally, I use this new framework to draw conclusions from the evidence on VAM-A and VAM-P discussed in the previous sections.

Specific Functions of Teacher Quality Measures

To make the functions of teacher quality measures more concrete, it is worth going beyond the general categories of signaling and improvement discussed earlier. Table 2 includes four specific functions, broken into two categories depending on the “weight” given to the measure—that is, the degree to which the measure is relied upon to carry out the function and influence the quality of the teacher workforce. Recommending that teachers reach a minimum level of some teacher quality measure obviously involves less weight than dismissing teachers from their jobs who are below the minimum. Other cases are less clear. For example, requiring teachers to meet a particular threshold gives significant weight to the teacher quality measure, but not necessarily more than compensating teachers by the measures.

Table 2

General and Specific Functions of Teacher Quality Measures

<i>General Function</i>	<i>Specific Function with low “weight”</i>	<i>Specific Function with high “weight”</i>
Signal	Characteristic is recommended as a condition of hiring	Characteristic is required as a condition of hiring
Improvement	Characteristic is required for experienced teachers to maintain certification	Characteristic is the primary basis for compensation or dismissal (firing)

In theory, a teacher quality measure can serve multiple functions. Based on the evidence discussed earlier, pedagogical content knowledge and experience appear to be two cases that serve both the signaling and improvement functions. However, these

appear to be rare exceptions. Many measures (e.g., personality and undergraduate education) are essentially fixed and can therefore serve only as signals. Even in these cases, it is not entirely clear that the signaling quality is sufficient to justify excluding potential teachers without observing their actual performance. This means that a combination of measures, each serving its own function in a multi-faceted strategy, is the best general approach to improving teacher quality.

Cost of Teacher Quality Measures

The focus so far on statistical validity and functions highlights the potential benefits of teacher quality measures. In other work, I have emphasized the fact that the costs of education programs are just as important as their effects (Harris, 2008). In this case, some teacher credentials are much more costly to produce than others.

There are general types of costs that need to be distinguished. Economists define costs as the value of a resource in its next best use—the “opportunity cost.” For example, for each hour a teacher spends instructing students, the teacher could have been working in some other job, spending time with her family, or some other personally valuable activities. Using the hour to teach children therefore comes at cost in terms of these opportunities foregone. Further, the value of this time can be measured in terms of the compensation paid by educational systems to personnel because compensation is assumed to reflect the opportunity cost of personnel time.¹⁷

The fact that compensation can be used to measure the economic cost of personnel is intuitive from a practical decision-making perspective. When district administrators are considering a new program that would involve hiring more teachers,

they will look at the budget and consider whether resources can and should be made available. In the case of teachers, the budgetary, or accounting, impact is often similar to the opportunity costs discussed above, but this is not always the case. Some budgetary costs over-state opportunity costs, e.g., if a state government decides to provide additional funding for school construction, it may pay for the schools over a 20-year period, but the value of the school, and thus its opportunity cost, is likely to last a half-century or more. Conversely, some economic costs may not show up in accounting costs. A notable example is that the teacher time spent in a graduate course does not show up directly in the school district budget, though the teacher could have spent that time developing lesson plans, correcting homework assignments and so on. When there is a difference between budgetary and opportunity costs, it is recommended that researchers use opportunity costs because this includes all costs to society (Levin & McEwan, 2001).

The most costly teacher quality measure is almost inarguably the master's degree in teacher education, which involves nearly 1,000 hours of teacher time in class and completing assignments.¹⁸ At an hourly rate of \$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 100 times larger than some professional development programs.¹⁹ 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 measures are used as the basis of compensation programs, as is typically the case in public (and most private) schools, then the above costs 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 20 years, then this could cost the school district \$60,000 over the teacher career—three times more than the costs of teacher time mentioned above. The example of compensation also highlights the fact that the specific functions and costs of the measures, like the functions and meanings of statistical validity, can be intricately related.

Functions, Costs, and Policy Validity

Having introduced the basic elements of policy validity—statistical validity, functions, and costs—it is now possible to discuss the relationships between the elements and to draw some general conclusions about valid policy uses of specific measures.

Table 3 provides a qualitative assessment of statistical validity. Each of the general functions—signaling and improvement—correspond to different types of policy inferences or conclusions and therefore involves a separate analysis of statistical validity. (Recall that statistical validity depends on the type of conclusion one is trying to make.) While all of the credentials have evidence regarding the signal, only experience and professional development have evidence about improvement. In the case of undergraduate education, the lack of evidence has to do with the nature of undergraduate education and policies that require teachers to have these degrees before entering the classroom on a full-time basis. In the case of graduate education and teacher test scores, it is possible for changes to occur while teachers are in the classroom, but I am aware of no evidence that has considered such effects (indicated in Table 3 by “[no evid.]”). The last column provides a qualitative assessment of the costs. Again, the purpose here is to

identify teacher quality measures that have high validity and low costs which, in this framework would mean that it is reasonable to give them greater weight in policy uses.

Table 3 suggests that teacher experience warrants the greatest weight of all the measures, given its moderate-high statistical validity and low costs.²⁰ As noted in Table 1, experience is consistently found to improve teacher value-added, especially in the early years on the job.

Table 3

Statistical Validity and Costs

<i>Teacher Characteristic</i>	<i>Stat. Validity: Signal</i>	<i>Stat. Validity: Improvement</i>	<i>Costs</i>
Undergrad. Educ.	Low	[no evid.]	High
Graduate Educ.	Low	[no evid.]	High
Prof. Develop.	Low-Moderate	Low-Moderate	Varies
Experience	Low-Moderate	Moderate	Low
Teacher Test Scores	Low	[no evid.]	Very Low

Teacher value-added is not included in Table 3 because the nature of statistical validity is somewhat different. While it is true by definition that a teacher whose value-added has increased has also improved, the larger question is whether teacher value-added measures can be used to facilitate improvement, e.g., by providing useful information to teachers and motivating them to get better or informing school leaders about whether to continue employing low-performing teachers and rewarding high-performing ones. Unfortunately, there is little direct evidence on this issue and, given the problems identified in the discussion of VAM assumptions and VAM-A findings, value-added currently deserves only a “low-moderate” rating for statistical validity.

The costs of value-added are also relatively low. Certainly, there are upfront costs to creating a data system that can provide the detailed data necessary to make the calculations, but these costs are spread across a large number of teachers. Suppose it costs \$100 million to create such a data system in a state and that the state has 100,000 teachers. In this case, the cost is only \$1,000 per teacher and some of these costs occur only once.²¹ The cost per teacher of standardized tests such as PRAXIS is also quite low.

In the introduction to this chapter, I described the traditional credentials strategy to improving teacher quality. Table 4 compares this current state of affairs with what the evidence from Table 3 seems to suggest about policy valid uses for the two general functions. Notice that, in the signaling category, the current weight given to the various signals is higher than the policy valid weights in all cases except for teacher experience and professional development. The reason for this is simply that the statistical validity of the other measures is weak, and some are costly, so that the best overall approach to the signaling function is to avoid filtering out too many teachers, because doing so is likely to result in a lot of effective teachers being excluded. This is somewhat less true of teacher experience and for this reason the current use of experience as a basis for compensation seems reasonable in terms of policy validity.

Conversely, Table 4 suggests that greater emphasis should be given to approaches that improve teacher quality, including measuring teacher performance (through VAM-A and perhaps other approaches) and by providing paths to improvement through professional development. Note that VAM-A is essentially excluded from current uses because it is currently used only in a small number of states and districts and, even in those locations, it is given little weight.

Table 4

Weight Given to Teacher Quality Measures

<i>Teacher Credential</i>	<i>Signals</i>		<i>Improvement</i>	
	<i>Current Policy</i>	<i>Policy Valid</i>	<i>Current Policy</i>	<i>Policy Valid</i>
Undergrad. Educ.	Moderate	Low	---	---
Graduate Educ.	Mod.-High	Low	Mod.-High	Low
Prof. Develop. Experience	Low	Low-Mod.	<i>Moderate</i>	<i>Moderate</i>
Teacher Test Scores	<i>Mod.-High</i>	<i>Mod.-High</i>	<i>Mod.-High</i>	<i>Mod.-High</i>
VAM-A	Moderate	Low	---	---
	---	Moderate	---	Moderate

Corruptibility and Other Possible Objections

There are a variety of possible objections to the conclusions drawn in the previous discussion. One of the most important is that the implicit assumption so far that statistical validity influences the appropriate weight given to any measure in policy, but that the weight *does not* influence statistical validity. In reality, giving considerable weight to any measure can “corrupt” and reduce its value as a tool in policy.

Accountability based on student standardized tests is frequently criticized, for example, because it sometimes leads teachers to teach students how to answer particular types of test questions, rather than helping students truly understand the content. If the truly lowest-value-added teachers carry out this form of test preparation more than the truly high-value-added teachers, and if test preparation succeeds in raising student scores, then the resulting teacher value-added scores will be corrupted—that is, they will inaccurately measure the genuine contributions to actual student learning made by teachers.

Teacher credentials are also potentially corruptible. Labaree (1997) describes how students in K-12 can “succeed in school without really trying” and argues persuasively that students’ efforts to make high marks makes the entire education system worse. It is reasonable to expect that this same phenomenon applies to teacher education, especially graduate education where a large percentage of teachers take classes mainly because they are required to do so in order to move into school administration or to obtain a higher salary. While some useful learning certainly takes place in these programs, these motives are obviously not conducive to genuine learning.

Nevertheless, while all measures are corruptible, it might be reasonable to conclude that measures of teacher performance are more corruptible than those of credentials, based on the simple fact that performance measures directly affect instructional practice. Paying teachers based on their degrees might corrupt teacher education, but have little negative influence on classroom instruction.

There are two other possible objections that warrant brief mention. First, this policy validity framework ignores the relationship between teacher quality measures and other student outcomes and intermediate outcomes such as teacher retention. As indicated earlier, value-added models are based on student achievement scores, which are imperfect both in the breadth of student knowledge, skills, and outcomes they cover. Much has been written on this subject, and even a cursory review is beyond the scope of this analysis. I would only add here that I am aware of no evidence about the relationship between student test scores and other student outcomes that would lead to substantially different conclusions than those reached based on student achievement and value-added methods.

Finally, VAM-A is given the most attention here as a direct measure of teacher performance. Earlier, I mentioned evidence that principals' subjective evaluations are positively (and statistically significantly) related to teachers' success in raising student test scores (Harris & Sass, 2007b; Jacob & Lefgren, 2005). In addition, structured principal evaluations and peer evaluations, might also be considered as a part of teacher quality policies.

Conclusions about Policy Validity

The ideal teacher quality measure is one that has high statistical validity, can be produced at low costs, and serves multiple functions. While it is obvious that no teacher quality measure lives up to this standard, this does not mean that the imperfect measures discussed above should not be used at all. Clearly, there has to be some strategy for improving the nation's teaching workforce and such a viable strategy almost certainly must be based on some combination of the measures considered here. The question is not so much whether the measures should be used, but how?

I argue that neither the traditional credentials strategy, nor the alternative strategy provides a sufficient answer to this question, and that a policy valid approach would involve a melding of these general strategies. As shown in Table 4, experience and some types of professional development deserve to be given considerable weight, both as signals and as a means of improvement. Further, because even these measures are only proxies for teacher effectiveness, it is also worth giving weight to more direct measures of teacher effectiveness, such as value-added to student achievement.

This analysis also suggests recommendations for what not to do. First, the master's degree is given too much weight in the traditional strategy. Instead of paying teachers based on the master's degree, perhaps it would make more sense to let schools and districts use the degree as one basis for promotion and taking steps up the career ladder, e.g., to the "master teacher" level. One might say that such a proposal would have the same effect as paying teachers based on the degree, because master teachers earn more money, but there are two important ways in which this is not the case: (a) all teachers who get the master's degree would not necessarily be promoted; and (b) master teachers have different responsibilities and part of the logic here is to require the degrees only when it seems plausible that the additional knowledge would contribute to the additional responsibilities. Clearly, a master teacher ought to have more, and more diverse, teaching skills than the average teacher and the master's degree might help to provide those additional skills, even if it does not improve individual teacher value-added.²²

Going to the other extreme, and focusing mainly on value-added, would be equally problematic. The unconfirmed assumptions and problematic empirical findings regarding individual teacher value-added (VAM-A) suggest that it, too, should be given only modest weight. The glass on the sidewalk simply does not measure rainfall as well as we might think. One option then is to combine value-added with principal and peer evaluations to develop a complete picture of teacher performance when making decisions about hiring, promotion, compensation, and/or dismissal. Given the limitations of the credentials strategy, experimentation with more direct measures of performance is certainly warranted. At the same time, proposals such as using value-added as the

primary basis for teacher tenure decisions (Gordon, Kane, & Staiger, 2006) arguably gives more weight to value-added measures than seems justified at present.²³

The discussion of various policy options also highlights their interconnectedness. There would likely be less pressure to compensate teachers based on performance if teachers could be more easily dismissed on the basis of low performance. Likewise, if a viable accountability system could be established, this would reduce (but not eliminate) the need to identify specific credentials of effective teachers and the priorities given to various credentials in decisions such as hiring. In one of our studies, we found that school principals report that “caring” is the most important attribute in teachers when hiring teachers (Harris, Rutledge, Ingle, & Thompson, 2006). This may be entirely rational because, if teachers were not caring, then there may not be enough other factors to motivate teachers to perform well after tenure. A shift toward accountability might also give principals greater motivation to become instructional leaders, as teachers would have greater incentive to improve and, in the case of subjective evaluations, to listen and respond to what the principals suggests. We could therefore expect that a change in external accountability policies would influence a wide variety of internal human resource policies as well.

The point here is less to criticize the traditional teacher quality strategy and more to show how this strategy impacts schools and teachers and therefore how alternative strategies might do more to improve schools. Such an effort is no doubt complex and requires a range of considerations regarding the functions of teacher quality measures, the various goals of education, and the larger policy context. The purpose of this study has been bring order to that complexity and provide direction for the next generation of

teacher strategies that, given the current widespread concern with the present strategy, is already well on its way to being formed. If there is one clear conclusion from this discussion, it is that the general shift toward an accountability strategy appears warranted, but it is also possible to go too far and create new failed policies that, rather than facilitating innovation and success, only serve to reinforce the limitations of the status quo.

References

- 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(2), 231-268.
- 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.
- Figlio, D. (2003). Testing, crime, and punishment. Working Paper. Gainesville: University of Florida.
- Goldhaber, D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, 22(2), 129-145.
- 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: The Brookings Institution.
- Greenwald, R., Hedges, L., & Laine, R. (1996). The effect of school resources on student achievement. *Review of Educational Research*, 66(3), 361-396.
- Hanushek, E. A. (1979). Conceptual and empirical issues in estimating educational production function issues. *Journal of Human Resources*, 14(3), 351-388.
- Hanushek, E. (1986). The economics of schooling. *Journal of Economic Literature*, 24, 1141-1177.
- Harris, D. N. (2007). "High flying schools, student disadvantage and the logic of NCLB," *American Journal of Education*, 113(3), 367-394.
- Harris, D. N. (2008). *New benchmarks for interpreting effects sizes: Combining effects with costs*. Unpublished manuscript, University of Wisconsin at Madison.

- Harris, D. N. (in press). Education production functions: Concepts. In B. McGaw, P. L. Peterson, E. Baker (Eds), *International encyclopedia of education*. Oxford, England: Elsevier.
- Harris, D. N., & Adams, S. (2007). Understanding the level and causes of teacher turnover: A comparison with other professions. *Economics of Education Review*, 26, 325-337.
- Harris, D. N., Handel, M., & Mishel, L. (2004). Education and the economy revisited: How schools matter. *Peabody Journal of Education*, 19(1), 36-63.
- Harris, D. N., & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2), 209-238.
- Harris, D. N., & Rutledge, S. (2007). *Models and predictors of teacher effectiveness: A review of the evidence with lessons from (and for) other occupations*. Unpublished manuscript, University of Wisconsin at Madison.
- 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 2006 conference of the American Education Research Association, San Francisco, CA.
- Harris, D. N., & Sass, T. (2005, March). *Value-added models and the measurement of teacher quality*. Paper presented at the 2005 conference of the American Education Finance Association, Louisville, KY.
- Harris, D. N., & Sass, T. (2006, March). *Teacher training and teacher productivity*. Paper presented at the 2006 annual meeting of the American Education Finance Association, Denver, CO.
- Harris, D. N., & Sass, T. (2007a). *The effects of NBPTS-certified teachers on student achievement*. Unpublished manuscript, University of Wisconsin at Madison.
- Harris, D. N., & Sass, T. (2007b, March). *What makes a good teacher and who can tell?* Paper presented at the 2007 annual meeting of the American Education Finance Association, Baltimore, MD.
- Harris, D. N., Taylor, L. Albee-Levine, A., Ingle, W. K., & McDonald, L. (2008, January). *The resource cost of standards, assessments and accountability*. Paper presented at the 2008 Workshop Series on State Standards, National Academy of Sciences, Washington, DC.
- Hoxby, C. (2002) The cost of accountability. In W. M. Evers & H. J. Walberg (Eds.), *School accountability*. Stanford, CA: Hoover Institution Press.

- Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (Working Paper #11463). Cambridge, MA: National Bureau of Economic Research.
- Kane, T., & Staiger, D. (2001). *Improving school accountability measures* (NBER Working Paper #8156). Cambridge, MA: National Bureau of Economic Research.
- Koedel, C., & Betts, J. R. (2005). *Re-examining the role of teacher quality in the educational production function*. Unpublished manuscript, University of California at San Diego.
- Labaree, D. (1997). *How to succeed in school . . . without really trying*. New Haven, CT: Yale University Press.
- Levin, H., & McEwan, P. (2001). *Cost-effectiveness analysis* (2nd ed.). London: Sage Publications.
- Levine, A. (2006). *Educating school teachers*. Washington, DC: The Education Schools Project. Retrieved October, 25, 2007, from http://www.edschools.org/teacher_report.htm
- Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student assessment: Evidence from Cincinnati. *Peabody Journal of Education*, 79(4), 33-53.
- Murnane, R.J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger.
- Ogbu, J. U. (2003). *Black American students in an affluent suburb: A study of academic disengagement*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Rivkin, S. G., Hanushek, E., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, 73(2), 417-458.
- Rutledge, S. Harris, D. N., Ingle, W., & Thompson, C. (in press). Certify, blink, hire: An examination of the process of teacher hiring. *Leadership and Policy in Schools*.
- 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(3), 247-256.
- Todd, P. E. & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), F3-33.

Wayne, A. J., & Youngs, P. (2003). Teacher credentials and student achievement gains. *Review of Educational Research*, 73(1), p.89-122.

Notes

¹ Some might refer to the “credentials strategy” instead as an “inputs-based” strategy and the “accountability strategy” as an “output-based” strategy. The credentials strategy also might be viewed as “process-oriented,” as reflected in teacher tenure rules and formal, low-stakes evaluations of teachers that are required in most schools and school districts.

² Others have reached different conclusion even when reviewing the same evidence. Greenwald, Hedges and Laine (1996) indicate that “school resources are systematically related to student achievement and that these relationships are large enough to be educationally important” (1996, p.384). Nevertheless, the relationships are inconsistent, a point reinforced by the more recent review in the present study.

³ One might argue that accountability incorporates elements of the accountability strategy, e.g., NCLB includes requirements that schools employ highly qualified teachers. While the term accountability is no doubt used in different ways by different people, I define it here as “test-based” and other forms of outcomes-focused policies, which clearly distinguishes the credential and accountability strategies.

⁴ When a signal is used by an organization in this way it is considered an act of “screening.”

⁵ This chapter does not address whether teachers should be paid more based on the specific subjects and grades they teach. Basic economic theory suggests that teachers should be paid more in fields where teacher supply is lower. For example, it is widely believed that teachers of math and science have greater opportunity costs than other teachers. Supply in specific teaching jobs is also affected by the characteristics of jobs, e.g., school location, student discipline problems, and differences in jobs by subjects. In addition to math and science, many schools have difficulty finding qualified special education students and this is at least partly due to state and federal regulations that limit the instructional options available to these teachers and impose considerable administrative burdens.

⁶ There are two general aspects of statistical validity that apply in all cases: bias and precision. Again, however, the meaning of “bias” depends on what one is trying to estimate. Another reviewer of an earlier version of this paper put this point differently, noting that “statistical validity is a property of an inference not of a test” (Floden, October 23, 2007, personal communication).

⁷ There are still some ways in which the professional developments effects might be biased in this value-added framework. For example, if teachers are assigned to professional development in each period based on achievement growth of students in previous periods, then the effects are still biased. Also, there may be other unmeasured factors that influence the teacher’s productivity while the professional development was taking place, but this would not bias the estimated effect unless teachers happened to be assigned to professional development based on unmeasured time-varying teacher credentials. This is possible. Some teachers do experience drops in productivity, e.g., when they have personal crises such as divorce from a spouse or the birth of a child, and these same teachers are less likely to take part in professional development for the same reason that it reduces their effectiveness—they have less time to devote to it. However, for this to substantially affect the estimated effect of teacher education, a substantial proportion of teachers have to experience such time-varying changes that influence both their classroom effectiveness and their likelihood of receiving professional development. This seems unlikely, though I am aware of no evidence that would shed significant light on the issue.

⁸ Value-added modeling accounts for student growth both before the teacher taught the student and afterwards. For example, if we were studying the effect of a fourth grade teacher, the student’s average rate of growth would be estimated would account for student learning in third and fifth grade, as well as fourth.

⁹ 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.

¹⁰ In some ways, the distinction between fixed and time-varying credentials just reiterates the distinction made earlier between signals and improvement, but there is a subtle difference. Signaling and improvement have to do with the function 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 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 the text.

¹¹ Other forms of professional development are combined into an “other” category, due to the limited distinctions made in the data. All of our results, including this one, tended to show greater statistical significance in middle school math. This is probably partly because, compared with reading, math achievement is determined more by specific math courses (there are few “reading” classes per se) and because, compared with elementary school, teachers are more likely to specialize in specific subjects; they therefore have more student observations from which the estimates can be made. In this particular case, the effect was also significant in some cases for high reading and math.

¹² The effects are only statistically significant in elementary and middle school math.

¹³ This may be because the researchers in this study control 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.

¹⁴ In addition to the twin non-random assignment problems, there are two other factors that may make it inherently difficult to identify the true credentials of high value-added teachers. The first is that the value-added approach necessarily requires very few “degrees of freedom.” Consider the simple case of a difference in mean value-added scores between two groups of teachers. Obviously, the more teachers in the respective groups, the more likely it is that any difference between them would be statistically significant. However, with VAM-P, we are identifying the effects of teacher credentials by comparing each teacher to herself. As discussed later, each teacher’s value-added is very imprecisely estimated and this problem is naturally compounded when analyzing changes in teacher value-added. Precision of course improves after combining the within-teacher differences across teachers in the analysis, but perhaps not enough to identify statistically significant effects even when such effects really exist.

In Harris and Sass (2007a), we approach this problem by estimating the models with and without teacher fixed effects. In the latter case, our assumptions are more similar to those of the gain score models discussed earlier. As expected, the statistical significance is considerably higher without teacher fixed effects, but this might come at the expense of some estimation bias. In about half the cases, the point estimates for the effects of teacher experience and professional development were similar (same sign and similar magnitude), implying that the gain score model may not introduce enough bias to change the general conclusions.

A second methodological issue is that the linear regression models, including value-added, may not capture the complexity of how teacher credentials combine to produce student achievement. For example, it may be that no characteristic really matters by itself, but only when it is combined with others. In addition it may be that there may be many contrasting ways to be an effective teacher. For example, some school principals in one of our studies indicated that some intelligent teachers have dull personalities, which makes them less effective in motivating students (Harris, Rutledge, Ingle, and Thompson, 2006). So, suppose that there are two types of teachers: those who are intelligent and unenthusiastic and those who are enthusiastic but less intelligent. Further, suppose that both intelligence and enthusiasm are positively associated with teacher effectiveness, but that very few teachers have both traits. In this case, a standard value-added analysis of these two groups of teachers will conclude—falsely—that neither enthusiasm nor intelligence is important. This is just one example and the larger point is that the importance of certain traits may be dependent on other traits.

¹⁵ In VAM-P studies, teacher experience is generally accounted for directly. This is not the case in VAM-A studies where interest lies primarily in determining which teachers are most effective rather than why.

¹⁶ While the specifics vary across states, the general idea is that schools will be judged not on whether their students are making AYP, but whether individual students are on track to being proficient. While in some sense, this is a positive move toward value-added modeling, most of the same problematic assumptions remain. Even with the growth curve analysis, all students still have to be proficient by 2014 in order to avoid sanctions. Thus, the fundamental problem of holding educators responsible for factors outside their control remains firmly intact. It is only the intermediate measures of school performance, between now and 2014, which actually change.

¹⁷ The assumption that personnel compensation provides a reasonable measure of opportunity requires some explanation. It is assumed that for-profit firms compensate their workers based closely on the individual's contribution to production. For-profit firms will not pay a worker more than she contributes, for doing so would reduce profits. At the same time, for-profits cannot underpay the worker for risk of losing the worker to another organization. Because for-profits compete for workers with non-profits and governments, it is reasonable to use the compensation paid to governmental and non-profit workers as a measure of their opportunity cost because these workers have opportunities in for-profit firms.

¹⁸ This calculation was made as follows. Suppose that master's degree requires 10 semester-long courses, each of which meets three hours per week for 15 weeks and requires an equal amount of time outside the classroom: (10 courses) x (15 weeks) x (6 hours) = 900 hours.

¹⁹ Harris and Sass (2007a) report that NBPTS certification requires roughly 200 hours of work. Professional development programs vary widely.

²⁰ The meaning of "costs of experience" requires some clarification. One might argue that the costs are actually high because it takes a teacher a full year in the classroom (at full salary) to gain a year of experience and, on top of that, teachers salaries increase with experience. There are two reasons why this intuition is not quite correct and why it is more reasonable to consider the costs of experience to be low: (1) experience is "naturally occurring" so that most learning from experience would occur regardless of whether it is rewarded through salary; and (2) the main activity going on during a school year is that teachers are teaching students and it is difficult to separate how much of the teacher's time is going toward teaching per se versus teacher learning—but surely the benefits received by students from the teaching are substantially greater than zero, thus substantially reducing the potential "cost of experience."

²¹ Harris, Taylor, Albee, Ingle, and McDonald (2008) and Hoxby (2002) discuss the costs of accountability systems, including the cost of data systems. Harris et al. argue that it is difficult to attribute the costs.

²² Using the master degree for other purposes might address the corruptibility problem by eliminating the least effective master's degree programs and allowing remaining programs to focus on the skills that master teacher teachers need.

²³ Gordon, Kane, & 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).