A Measured Approach

Value-Added Models Are a Promising Improvement, but No One Measure Can Evaluate Teacher Performance

ILLUSTRATED BY PAUL ZWOLAK

By Daniel Koretz

Suppose you and I teach fifth grade—as I did many years ago—but we teach in very different settings. Our students are different: perhaps yours enter fifth grade with lower levels of achievement, or you have more students with limited proficiency in English. Their previous teachers were not similar: perhaps those who taught my students were more skilled. On average, my students have more highly educated parents than yours. Our schools have different levels of resources, and the peer culture and community support for education are dissimilar. But our students do have one thing in common: at the end of the school year, our students will take the same achievement tests, and policymakers would like to use their scores to judge how effective we both were. How fairly can that be done, given our very different situations?

The education policy community is abuzz with interest in value-added modeling as a way to estimate the effectiveness of schools and especially teachers—even those with very different students, in very different settings. Value-added approaches are widely believed to be superior to the common alternatives as a way of estimating the performance of schools and teachers. But just how well do value-added models serve this role? There is no doubt that value-added models are superior in some important ways, but they are no silver bullet. Value-added models provide important information, but that information is error-prone and has a number of other important limitations. Moreover, these methods are still under development, and the various approaches now in use do not always paint the same picture. Value-added estimates can be an important part of an evaluation of teachers and schools, but they are not sufficient by themselves for this purpose.

Although there has been intense discussion of the strengths and limitations of the value-added approach among research-

Daniel Koretz is professor of education at Harvard University. He founded and chairs the International Project for the Study of Educational Accountability and is a member of the National Academy of Education. His research focuses on the effects of high-stakes testing, including effects on schooling and the validity of score gains, and the design and evaluation of test-focused educational accountability systems. Previously, he taught emotionally disturbed students in public elementary and junior high schools. *He wishes to thank Daniel McCaffrey, J.R. Lockwood, and Laura Hamilton—colleagues with whom he worked on RAND's evaluation of value-added modeling—for their helpful comments on an earlier draft.*

ers, too little discussion has taken place in the education policy community. This may stem from the tremendous technical complexity of most value-added approaches, which render them seemingly incomprehensible to most people, or from policymakers' hope for a relatively simple way of evaluating teachers and schools, or both. Yet without this discussion, we are not likely to use value-added modeling in an appropriate and productive way. This article describes some of the key issues raised by value-added modeling and concludes with some suggestions for its use. Many of the issues are similar regardless of whether schools or teachers are evaluated, and I touch on both, but I focus especially on the evaluation of teachers.

How Value Added Improves on the Status Quo

Most test-based accountability programs in the United States have used one of three approaches for evaluating student achievement. *Status models* are based simply on the scores of a group at one time. For example, the average performance of a school's fourth graders, or the proportion of fourth graders who exceed a standard such as "proficient," can be compared with an expected level or with the results from other schools. *Cohort-to-cohort change models* are based on the change in statistics such as these over time. For example, the percentage of fourth graders considered proficient this year can be contrasted with the comparable statistic from last year to see which schools have attained an expected degree of improvement. The federal education law, No Child Left Behind (NCLB), is a hybrid of these two approaches. For most schools, NCLB functions as a status model: in any given year, the performance of the school is compared with the state's annual measurable objective for that year. However, the objective increases every year (on its way to the goal of 100 percent proficient by 2014), which creates pressures similar to that found in a cohort-to-cohort change system. In addition, NCLB's safe harbor provision is a true cohort-to-cohort change approach.

In contrast to both of these, *value-added models* (VAMs) are based on the growth individual students achieve during a year of schooling. If I were still a fifth-grade teacher, a status model would evaluate me based on my students' performance at the end of this year, and a cohort-to-cohort change model would judge me based on the difference between the end-of-year scores of my fifth graders this year and those I had the year before. Under a VAM, I would be rated on the basis of my students' gains during their year with me; I would be evaluated favorably if they showed more growth than whatever comparison policymakers decided to use (which might be the average of other teachers in my district or state, or some pre-established amount), even if my students' performance when entering my class was so weak that their scores at the end of fifth grade remained low.

Unfortunately, the term "value added" is used to represent two very different quantities. The first is students' total growth—how much their achievement increased, *for whatever reason*, during their fifth-grade year with me. The second is how much *my efforts* contributed to that growth—how much "value" I added. Because many factors other than teachers' work contribute to (or impede) growth, these two quantities can be quite different. I'll use the term *value added* to refer to both for now, but I'll return to this distinction later.

In test-based accountability systems, value-added approaches offer three very important advantages compared with status models and cohort-to-cohort change models. First, at least in theory, VAMs measure the right thing, which neither status nor cohort-to-cohort change models do. A sensible accountability system, for teachers or for any other professionals, holds people accountable for what they can control. Teachers should be held accountable for what they contribute to their students' growth, not for the accumulated knowledge and skills (or lack thereof) that students bring with them to the first day of class.

The education policy community is abuzz with interest in value-added modeling as a way to estimate the effectiveness of schools and especially teachers. Value-added models provide useful information, but that information is error-prone and has a number of other important limitations.

While adjusting for students' achievement levels when they enter the grade would be a clear and important improvement over cohort-to-cohort change and status models, it is not enough to get us a true estimate of "value added." The ideal is to adjust not only for students' prior achievement levels, but rather for their *expected growth trajectories*. To better understand this, let's go back to the example of fifth grade, and let's add the condition that you and I are equally effective teachers. This time, let's assume that, for whatever reason, you are given a class of high achievers, with very few students reading below grade level and many reading several years above grade level. In contrast, I draw—as I did in actuality, many years ago—a class with many very poor readers, some several full years below grade level. (So far below, in fact, that many still struggled with decoding and read letter by letter.) Would these two groups gain reading skills at the same rate if they had equally effective teachers? Should I be judged less effective than you if my students gained less in reading skills during the fifth grade than yours? Most experienced teachers, I suspect, would say no. To make the comparison truly fair, one would want the system to adjust for *differences in the growth* that these two very dissimilar groups would show during fifth grade if they were given equally high-quality schooling.

The achievement level of students when they enter a grade

reflects the cumulative effects of many factors, both educational and not. Some of these factors will persist after the students enter your class and will tend to push them toward a growth trajectory similar to that which they showed before. Some of these are characteristics of the students themselves, such as disabilities, health conditions, and simple differences in aptitude. Some are characteristics of their families or communities. For example, my own children attended school in a neighborhood in which many parents either hired tutors or retaught material themselves if their children encountered difficulties (as I did when my son encountered difficulties with his mathematics homework)—which increased their children's rate of growth and gained the schools some credit they did not actually deserve. The combined effects of these influences make some students much easier to teach than others. I have taught in settings ranging from special education elementary school classes to doctoral-level university courses, and this variation in students has been striking in every class I have taught.

Therefore, some current VAMs try to adjust for differences in students' expected growth trajectories by taking into account several years of prior achievement, not just scores from the year before entry to a class. By evaluating several years of scores, the models indirectly take into account persistent noneducational factors that influence students' rate of growth, and some approaches also incorporate some of these factors directly into the model.

This brings us to the second main advantage of VAMs: they can do a substantially better job than status models or cohort-to-cohort change models of controlling for differences among students that would otherwise be confounded with the effects of teaching. Currently, there is a great deal of argument among experts about how well VAMs do this—how close they come to estimating the value added by teachers rather than just estimating student growth. For reasons that I will explain below, we cannot be confident that value-added models pare away all of the growth attributable to other factors in order to reveal the pure effects of teaching. Nonetheless, in general, VAMs do a better job of adjusting for other influences on achievement than do the typical status or cohort-to-cohort approaches.

The final major advantage of VAMs is that they reveal substantial differences among classrooms and schools in students' performance. We all have known superb teachers and teachers who are struggling, so it is reasonable to expect a measure of student performance to show substantial variations. Test scores show great variation among schools, but research has often found that after adjusting for factors such as background characteristics, relatively little variation—implausibly little, some observers would say—remains. In contrast, VAM estimates often show the sizeable differences among teachers and schools that many would expect.*

Difficulties in Using Value-Added Models for Accountability

Applying VAMs to the evaluation of schools and teachers is not straightforward, and some of the issues debated by experts, while important, seem simply impenetrable to most people other than statisticians and psychometricians. This in itself is a drawback, as it's certainly preferable for educators, parents, policymakers, and the like to understand how their teachers and schools are being evaluated. Fortunately, many of the most important complications can be reduced to the following six simple questions, each of which I'll briefly discuss: (1) What are we measuring?

Value-added models can do a better job than the alternatives of controlling for differences among students that would otherwise be confounded with the effects of teaching. But we cannot be confident that value-added models pare away all of the growth attributable to other factors in order to reveal the pure effects of teaching.

(2) How do we measure it? (3) How precise can we be? (4) How certain are we about how to model gains? (5) How well do we adjust for other influences on achievement growth? (6) How does score inflation affect value-added models?

1. What are we measuring?

It is essential to keep in mind a warning offered by some of the progenitors of achievement testing more than half a century ago: standardized achievement tests can only measure a subset of the critically important goals of education. First, they measure only achievement, not motivation, curiosity, creativity, and the ability to work well in groups. Second, most testing systems measure achievement in only a subset of the subject areas with which we should be concerned. Third, within the tested subject areas, they measure only a subset of the important knowledge and skills. Some important outcomes are very difficult or impractical to test with standardized, externally imposed tests. The information yielded by standardized tests can be tremendously valuable, but it is nonetheless seriously incomplete, and therefore scores taken alone cannot provide a comprehensive evaluation of perfor-

* In the current context of NCLB, another advantage is that most value-added models take into account every student's progress. In contrast, NCLB and most state accountability systems focus primarily on the percentages of students reaching a proficient standard, which renders progress by most students—those well below or well above the standard—invisible and unimportant. As I explain in my new book *Measuring Up* (see chapter 8), this is only one of many serious drawbacks of reporting student achievement only in terms of performance standards. However, this advantage is not inherent to VAMs. There is no reason why cohort-to-cohort change or status models need to focus on the percentages of students reaching a standard rather than on the performance of all students.

mance. (To better understand this concern, see the sidebar from *Measuring Up* on page 22.)

Far from circumventing this problem, value-added models may exacerbate it. The VAMs we use today require that growth in achievement be cumulative across grades. We want to know how far a student has progressed in learning mathematics by the end of grade 4, so that we can evaluate how much her knowledge has increased by the end of grade 5. This requires *vertically scaled* tests: tests that place performance in adjacent grades on a single scale.† The more dissimilar the content of instruction is from grade to grade, the less plausible this approach is. Vertically scaled tests are commonplace in reading comprehension and certain areas of mathematics, but they may not be practical in science or social studies, even in the elementary and middle grades. More subtle, but also important, is that using VAMs may constrain what we test within a subject as well. The more grade-specific the important content in one subject is, the less practical it becomes to build defensible vertically scaled tests. Therefore, reliance on VAMs may encourage focusing on a subset of important subjects and narrowing the focus within subjects to the material most amenable to vertical scaling.

2. How do we measure achievement?

Although many people believe that tests are direct and simple measures of achievement, they are anything but. A test is only a small sample from a large "domain" of knowledge and skill, and performance on the tested sample—the test score—is only valuable to the extent that it provides a good estimate of mastery of the entire domain. (These issues are explored in the sidebar on page 22.) Constructing a test entails a long series of decisions, both substantive and technical. Some of these decisions, such as the choice of a mathematical model for creating a scale, are arcane, but they matter: they can substantially affect the estimates of gains that are provided by value-added models. I'll give three examples.

The first is the selection of content. Consider middle school mathematics. In many middle schools, there is considerable tracking in mathematics, and there are likewise curricular differences among schools. Some seventh graders are studying algebra, while others are still focused on arithmetic. Suppose you and I are equally effective seventh-grade math teachers. You are teaching a class in which a good deal of time is devoted to algebra, while I am teaching one focused primarily on arithmetic. Suppose also that our state uses a test that focuses on basic skills. What will value-added models say about us? You lose: much of the progress you make with your students will not be captured by the test because it does not include algebra. The technical term you may see for this is *dimensionality*. Most tests measure multiple aspects or dimensions of performance, although they provide a summary score combining all of them. The closer the mix of tested dimensions is to your curriculum, the more effective you will seem.

The second testing issue is scaling: deciding on a set of numbers to represent performance. Most value-added approaches assume an *interval scale*, such that any given increment, say, 20 points, means the same improvement in achievement at any level of the scale (so, for example, an increase from 120 to 140 represents the same amount of growth as an increase from 200 to 220). Most people don't give this concern much thought, since most of the measures we use in daily life, such as pounds, feet, and temperature, are interval scales. Unfortunately, test scores do not have this handy property: we would like an interval scale, but most of the time we don't know whether we have one. We can't be confident that, for example, an increase from 500 to 540 on the SAT mathematics test represents the same amount of gain as an increase from 700 to 740. Worse, different scales do not necessarily agree in this regard. A high-achieving student and a low-achieving student who appear to have gained the same amount on one scale may show different amounts of growth on another.

For many practical purposes, this uncertainty does not matter much. For example, it has been shown that many of the commonly used scales correspond reasonably well in this respect, provided that the comparison is restricted to one grade and year, and to students who are not dramatically different in performance. However, it clearly can matter with VAMs. For example, some scales will show the performance of high achievers and low achievers diverging as they progress through the grades, while others show the reverse, and yet others show the two groups keeping pace with each other. This creates a distressing uncertainty in the results of value-added models when the groups compared start out at substantially different levels of achievement. (I'll return to this at the end, when I offer some suggestions about using VAMs sensibly.)

The final example is the timing of testing. Most states test once a year, near (but not at) the end of the school year. Therefore, the growth attributed to a teacher excludes the final weeks or months of the school year and includes both the final period in the previous year (with the previous teacher) and summer vacation. Particularly given evidence that students show different patterns of growth or loss during the summer, these problems of timing are worrisome.‡ Although there are some statistical simulations suggesting that the effects of this less-than-optimal timing are usually not great, the jury is still out, and there may be some circumstances in which this is an appreciable source of bias in the ranking of teachers or schools.

3. How precise can we be?

Years ago, fresh out of graduate school, I wrote testimony for a congressional committee in which I referred to the "margin of error" in my estimate of the impact of a program the committee was considering terminating. This angered the chair of the committee, who glowered at the person giving the testimony—unfortunately, my boss—and said, "What is this 'margin of error' stuff? Doesn't it mean that you don't know what the hell you're talking about?" Well, in a sense, yes, although he was overstating the problem. While the chair wanted certainty, no one could honestly give it to him: all statistical estimates are subject to some uncertainty or imprecision, and this includes test scores and the results of models that use them. Terms such as "margin of error" or the more specific "standard error" are just our tools for quantifying how much imprecision remains.

To start, we have to distinguish between *error* and *bias*. In

† A few value-added models loosen this requirement slightly, but these exceptions do not contradict the points made here. There are also statistical approaches for estimating the value added by individual teachers that are not based on prior growth in the same subject area, but we are not considering those here.

‡ For more on summer learning loss, see "Keep the Faucet Flowing" in the Fall 2001 issue of *American Educator*, online at **www.aft.org/pubs-reports/american_educator/fall2001/faucet.html**.

educational testing, as in most of quantitative science, "error" has a narrower meaning than it does in common parlance. If you buy a cheap bathroom scale, it may simply be inconsistent, so that your weight seems to be different each time you step on it, but not systematically too high or too low. This inconsistency is error. On the other hand, your bathroom scale could be systematically wrong, so that it consistently tells you that you are lighter than you really are. In educational testing, this systematic inaccuracy is called *bias*, not error. If a student's score is consistently too low, as may happen in the case of students not fully proficient in English, that would constitute bias; but if a student's score is sometimes too low and sometimes too high, that would be error.

Even if they are entirely unbiased, estimates based on test scores inevitably entail error. In fact, both bias and error are concerns when value-added models are used to evaluate teachers or schools. I'll discuss error here and return to bias a bit later.

Error is of two analogous types that have different sources: *sampling error*, which is more familiar to most people, and *measurement error*.* *Sampling error* stems from the selection of particular individuals from whom data will be collected. In the case of educational accountability, sampling error arises because a teacher is given a different sample of students every year, and, as one teacher put it in a study years ago, "there are good crops and bad crops." Your scores—and your apparent "effectiveness"—will fluctuate as a result of these differences in samples. These fluctuations are particularly pronounced for small groups

* I provide a more thorough explanation of bias, measurement error, and sampling error in *Measuring Up*.

Measuring Up

What Educational Testing Really Tells Us

Educational testing is ubiquitous in America, and its importance is hard to overstate. Tests have a powerful influence on public debate about many social concerns, such as economic competitiveness, immigration, and racial and ethnic inequalities. And achievement testing seems reassuringly straightforward and commonsensical: we give students tasks to perform, see how they do on them, and thereby judge how successful they or their schools are.

This apparent simplicity, however, is misleading.

Test scores do not provide a direct and complete measure of educational achievement. Rather, they are incomplete measures, proxies for the more comprehensive measures that we would ideally use, but that are almost always unavailable to us. There are two reasons for the incompleteness of achievement tests. The first, which has been stressed by careful developers of standardized tests for more than half a century, is that these tests can measure only a subset of the goals of education. Some goals, such as the motivation to learn, the inclination to apply school learning to real situations, the ability to work in groups, and some kinds of complex problem solving, are not very amenable to large-scale standardized testing. Others can be tested, but are not considered a high enough priority to invest the time and resources required. The second reason for the incompleteness of achievement tests—and the one that I will focus on here—is that even in assessing the goals that we decide to measure and that can be measured well, tests are generally very small samples of behavior that we use to make estimates of students' mastery of very large domains of knowledge and skill.

The accuracy of these estimates depends on several factors, one of the most important being careful sampling of content and skills. For example, if we want to measure the mathematics proficiency of eighth graders, we need to specify what knowledge and skills we mean by "eighth-grade mathematics." We might decide that this subsumes skills in arithmetic, measurement, plane geometry, basic algebra, and data analysis and statistics, but then we would have to decide which *aspects* of algebra and plane geometry matter and how much weight should be given to each component (e.g., do students need to know the quadratic formula?). Eventually, we end up with a detailed map of what the test should include, often called "test specifications" or a "test blueprint," and the developer writes test items that sample from it.

But that is just the beginning. The accuracy of a test score depends on a host of often arcane details about the wording of items, the wording of "distractors" (wrong answers to multiple-choice items), the difficulty of the items, the rubric (criteria and rules) used to score students' work, and so on. The accuracy of a test score also depends on the attitudes of the test takers—for example, their motivation to perform well. It also depends, as we shall see later, on how schools prepare students for the test. If there are prob-

because there is less opportunity for the characteristics of individual students to cancel each other out. Thus, the smaller the group, the greater the sampling error, and the greater the uncertainty in the group's test scores—or in the estimates of value added based on them.

Measurement error is different: it affects even the score of a single student and reflects inconsistencies from one instance of measurement to another. Students who take the SAT multiple times, for example, generally see a fluctuation in their scores from one time to the next because of measurement error. As explained in the sidebar (below) from *Measuring Up*, there are three primary sources of measurement error: the selection of specific test items in constructing the test, fluctuations in the student's performance from day to day, and inconsistencies in scoring.† Some states and districts now take measurement error into account when reporting scores, telling parents that the best estimate of a student's performance falls within a range surrounding her obtained score.

The score reports used in accountability systems are subject to both measurement error and sampling error. As a result, one can't take the precise score obtained for a school or classroom at face value. Rather, the score is an estimate, and the true value lies within a band of uncertainty that surrounds the estimate obtained. (This is no different from the polls you see in the news-

(Continued on page 26)

† *Reliability* is a function of error: a perfectly reliable score would be error-free (in most cases, an impossibility), while a completely unreliable score would represent nothing but error.

lems with any of these aspects of testing, the results will provide misleading estimates of students' mastery of the larger domain.

A failure to grasp this fact is at the root of widespread misunderstandings—and misuses—of test scores. It has often led policymakers astray in their efforts to design productive testing and accountability systems. By placing too much emphasis on test scores, they have encouraged schools to focus instruction on the small sample actually tested rather than the broader set of skills the mastery of which the test is supposed to signal.

To make the principles of testing concrete, let's construct a hypothetical test. Suppose that you publish a magazine and have decided to hire a few college students as interns to help out. You receive a large number of applicants and have decided that one basis for selecting from among them is the strength of their vocabularies. How do you determine that? Conversations with them will help, but may not be sufficient because they are not uniform: a conversation with one applicant may afford more opportunities for using advanced vocabulary than a conversation with a second one. So you decide to construct a standardized test of vocabulary.* You would then confront a serious difficulty: although many teachers and parents may find this fact remarkable in the light of their own experience, the typical adolescent has a huge working vocabulary. Clearly, you will have to select a sample of words to put into your test. In practice, you can get a reasonably good estimate of the relative strengths of applicants' vocabularies by testing them on a small sample of words, if those words are chosen carefully. Assume you will use 40 words, which would not be an unusual number in an actual vocabulary test.

The box below gives the first few words from three lists that you could use to select words for your test.

A	B	C
siliculose	bath	feckless
vilipend	travel	disparage
epimysium	carpet	minuscule

Which list would you use? Clearly not list A, which comprises specialized, very rarely used words. Everyone would receive a score of zero or nearly zero, and that would make the test useless: you would gain no useful information about the relative strengths of their vocabularies. List B is no better. Everyone would obtain a perfect or nearly perfect score. Therefore you would construct your test from list C, which comprises words that some applicants would know and others not.

In this example, the fact that a test is merely a sample of a larger domain is clear. But is sampling always as serious a problem as it is in this contrived example? For the most part, yes.† The tests that are of interest to policymakers, the press, and the public at large entail substantial sampling because they are designed to measure sizable domains, ranging from knowledge acquired over a year of study in a subject to cumulative mastery of material studied over several years.

Returning to the vocabulary test: what would have happened if you had chosen words differently, while keeping them at the same level of difficulty? To make this concrete, assume that you selected all three of the words shown in list C, and that I was also constructing a vocabulary test, but I dropped *feckless* and used *parsimonious* instead. For the sake of discussion, assume that these two words are equally difficult.

What would be the impact of administering my test rather than yours? Over a large enough number of applicants, the average score would not be affected at all, because the two words in question are equally difficult. However, the scores of some individual students would be affected. Even among students with comparable vocabularies, some would know *feckless* but not *parsimonious*, and vice versa.

This illustrates one source of *measurement error*, which refers to inconsistency in scores from one measurement to the next. To some degree, the ranking of your applicants will depend on which words you select from list C, and if you tested applicants repeatedly using different versions of your test, the rankings would vary a little. Another source of measurement error is the fluctuation over time that would occur even if the items were the same. Students have good and bad days. For example, a student might sleep well before one test date but be too anxious to sleep well another time. Or the examination room may be overheated one time but not the next. Yet another source of measurement error is inconsis-

* People incorrectly use the term *standardized test*—often with opprobrium—to mean all sorts of things: multiple-choice tests, tests designed by commercial firms, and so on. In fact, it means only that the test is uniform: that is, that all examinees face the same tasks, administered in the same manner, and scored in the same way. The motivation for standardization is to avoid irrelevant factors that might distort comparisons among individuals.

† There are tests that are not samples of a larger domain. For example, a teacher may want to know whether her class has mastered the list of vocabulary words presented in the past week. She would not be trying to draw any conclusions about students' overall vocabularies, and she would be happy indeed if most students got most of the words right.

tencies in the scoring of students' responses.

Obviously, it's important to try to keep measurement error to a minimum—and that's why test developers are so concerned with *reliability*. Reliable scores show little inconsistency from one measurement to the next—that is, they contain relatively little measurement error. Reliability is often incorrectly used to mean "accurate" or "valid," but it properly refers only to the consistency of measurement. A measure, including a test, can be reliable but inaccurate—such as a scale that consistently reads too high.

So when all is said and done, how justified would you be in drawing conclusions about vocabulary from the small sample of words on your test? This is the question of *validity*, which is the single most important criterion for evaluating achievement testing. In public debate, and sometimes in statutes and regulations as well, we find reference to "valid tests," but tests themselves are not valid or invalid. Rather, inferences based on test scores are valid or not. A given test might provide good support for one inference, but weak support for another. For example, a well-designed end-of-course exam in statistics might provide good support for inferences about students' mastery of basic statistics, but very weak support for conclusions about mastery of mathematics more broadly. The question to ask is: how *well supported* is the conclusion?

None of the preceding is particularly controversial. These fundamentals of testing may not be well known outside the testing community, but inside that community they are widely agreed upon. The next and final step in this hypothetical exercise, however, is contentious indeed.

Suppose you are kind enough to share with me your test of 40 words. And suppose I intercept every single applicant en route to taking your test, and I give each one a short lesson on the meaning of every word on your test. What would happen to the validity of inferences you might want to base on your test scores?

Clearly, your conclusions about which applicants have stronger vocabularies would now be wrong. Most students would get high scores, regardless of their actual vocabularies. Students who paid attention during my mini-lesson would outscore those who did not, even if their actual vocabularies were weaker. Mastery of the small sample of 40 words would no longer represent variations in the students' actual working vocabularies.

This last step—teaching the specific content of the test, or material close enough to it to undermine the representativeness of the test—illustrates the contentious issue of *score inflation*, which refers to increases in scores that do not signal a commensurate increase in proficiency in the domain of interest. Inflation of scores in this case did not require any flaw in the test, and it did not require that the test focus on unimportant material. The 40 words were fine. My response to those 40 words—my form of test preparation—was not.

In real-world testing programs, issues of score inflation and test preparation are far more complex than this example suggests. So let's set aside our vocabulary test and take a closer look at what I believe should be a very serious concern among educators and policymakers: how to prepare for tests.

Test preparation has been the focus of intense argument for many years, and all sorts of different terms (like "teaching the test" and "teaching to the test") have been used to describe both good and bad forms. I think it's best to ignore all of this and to distinguish instead between seven different types of test preparation: (1) working more effectively, (2) teaching more, (3) working harder, (4) reallocation, (5) alignment, (6) coaching students, and (7) cheating.

The first three are what some proponents of high-stakes testing want to see. Clearly, if educators find ways to work more effectively—for example, developing better curricula or teaching methods—students are likely to learn more. Up to a point, if teachers spend more time teaching, achievement is likely to rise. The same is true of working harder in school, although this can be carried too far. For example, it is not clear that depriving young children of recess, which some schools are now doing in an effort to raise scores, is effective, and in my opinion it is undesirable regardless. Similarly, if students' workload becomes excessive, it may interfere with learning and may also generate an aversion to learning. But if not carried to excess, these three forms of test preparation can be expected to produce real gains in achievement that would appear not only in the test scores used for accountability, but on other tests and outside of school as well.

At the other extreme, cheating is unambiguously bad. But what about reallocation, alignment, and coaching? All three can produce real gains, score inflation, or both. Reallocation refers to shifting instructional resources—classroom time, homework, parental nagging, whatever—to better match the content of a specific test. A quarter century of studies confirm that many teachers reallocate instruction in response to tests. And some studies have found that school administrators reassign teachers to place the most effective ones in the grades in which important tests are given.[1]

Is reallocation good or bad? Does it generate real gains in achievement or score inflation? This depends on what gets more emphasis, *and what gets less*. Some reallocation is desirable and is one of the goals of testing programs. For example, if a ninth-grade math test shows that students do relatively poorly in solving basic algebraic equations, one would want their teachers to put more emphasis on such equations. The rub is that devoting more resources to topic A entails fewer resources for topic B.

Scores become inflated when topic B—the material that gets less emphasis as a result of reallocation—is also an important part of the domain. If teachers respond to a test by de-emphasizing

material that is important to the domain but is not given much weight on the particular test, scores will become inflated. Performance will be weaker when students take another test that places emphasis on those parts of the domain that have been neglected.

Alignment is a lynchpin of policy in this era of standards-based testing. Tests should be aligned with standards, and instruction should be aligned with both. And alignment is seen by many as insurance against score inflation, but this is incorrect. Alignment is just reallocation by another name. Whether alignment inflates scores also depends on the importance of the material that is de-emphasized. And research has shown that standards-based tests are not immune to this problem. These tests are still limited samples from larger domains, and therefore focusing too narrowly on the content of the specific test can inflate scores.

Coaching students refers to focusing instruction on small details of the test, many of which have no substantive meaning. Coaching need not inflate scores. If the format or content of a test is sufficiently unfamiliar, a modest amount of coaching may even increase the validity of scores. For example, the first time young students are given a test that requires filling in bubbles on an answer sheet that is going to be scored by a machine, it is worth spending a very short time familiarizing them with this procedure before they start the test.

Most often, however, coaching students either wastes time or inflates scores. A good example is training students to use a process of elimination in answering multiple-choice questions. A *Princeton Review* test-prep manual urges students to do this because "it's often easier to identify the wrong answers than to find the *correct* one."[2] What's wrong with this? The performance gains generated depend entirely on using multiple-choice items. Of course, when students need to apply their knowledge in the real world outside of school, the tasks are unlikely to appear in the form of a multiple-choice item.

This example shows that inflation from coaching is in one respect unlike inflation from reallocation. Reallocation inflates scores by making performance on the test unrepresentative of the larger domain, but it does not distort performance on the material tested. (If I taught applicants the vocabulary words on your test, they would know those words—but their scores on the test would not be good estimates of their overall vocabulary knowledge.) In contrast, coaching can exaggerate performance on the tested material. In the example just given, students who are taught to use the process of elimination as a method for "solving" certain types of equations will know less about those types of equations than their performance on the test indicates.

So what distinguishes good and bad test prep? The acid test is whether the gains in scores produced by test preparation truly represent meaningful gains in student achievement. We should not care very much about a score on a particular test. What we should be concerned about is the knowledge and skills that the test score is intended to represent. Gains that are specific to a particular test and that do not generalize to other measures of the domain and to performance in the real world are worthless.

* * *

This brings me to a final, and politically unpalatable, piece of advice: we need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. They are likely instead to produce substantial inflation of scores and a variety of undesirable changes in instruction, such as excessive focus on old tests, inappropriate narrowing of instruction, and a reliance on test-taking tricks.

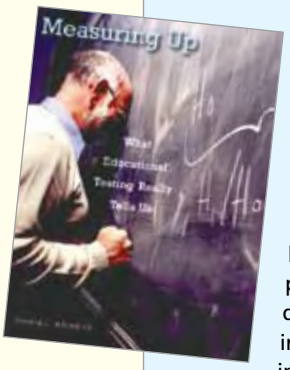
I strongly support the goal of improved accountability in public education. I saw the need for it when I was an elementary school and junior high teacher, many years ago. I saw it as the parent of two children in school. Nothing in more than a quarter century of education research has led me to change my mind on this point. And it seems clear that student achievement must be one of the most important things for which educators and school systems should be accountable. However, we need an effective system of accountability, one that maximizes real gains and minimizes bogus gains and other negative side effects. Even a very good achievement test will leave many aspects of school quality unmeasured. Some hard-core advocates of high-stakes testing disparage this argument as "anti-testing," but it is a simple statement of fact, one that has been recognized within the testing profession for generations.

So how should you use scores to help you evaluate a school? Start by reminding yourself that scores describe some of what students can do, but they don't describe all they can do, and they don't explain why they can or cannot do it. Use scores as a starting point, and look for other evidence of school quality—ideally not just other aspects of student achievement but also the quality of instruction and other activities within the school. And go look for yourself. If students score well on math tests but appear bored to tears in math class, take their high scores with a grain of salt, because an aversion to mathematics will cost them later in life, even if their eighth-grade scores are good.

Sensible and productive uses of tests and test scores rest on a single principle: don't treat "her score on the test" as a synonym for "what she has learned." A test score is just one indicator of what a student has learned—an exceptionally useful one in many ways, but nonetheless one that is unavoidably incomplete and somewhat error-prone.

–D.K.

This sidebar was adapted from Daniel Koretz's new book, *Measuring Up: What Educational Testing Really Tells Us*. Detailed but nontechnical, the book addresses the common misunderstandings and misuses of standardized tests, and offers sound advice for using tests responsibly. To learn more, go to **www.hup.harvard.edu/catalog/KORMAK.html**. *Measuring Up*,



Endnotes

1. For a good overview of some of the most important research on teachers' and principals' responses to testing, see Brian M. Stecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education*, ed. Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, CA: Rand, 2002), **http://www.rand.org/pubs/monograph_reports/MR1554**.

2. Jeff Rubenstein, *Princeton Review: Cracking the MCAS Grade 10 Math* (New York: Random House, 2000), 15.

(Continued from page 23)

paper: they are usually reported with a "margin of error" of plus or minus a few percentage points, which is their band of uncertainty.) This inevitable error is one of several reasons why no single measure should be used to make an important decision. Even if a measure is entirely unbiased, any single test score may be too high or too low, sometimes by a considerable amount.

Error affects all accountability approaches—status, cohort-to-cohort, and value-added models. There is still disagreement among experts about the precise amount of error in different VAMs, but there is no doubt that it is a serious problem indeed, particularly when the model is applied to individual teachers (since they have a limited number of students, the sample size is small, and sampling error is large). To rank teachers based on VAMs, we would need very small errors, and research to date suggests that we cannot yet reach that threshold. We may be able to identify some teachers whose students show higher- or lower-than-average gains, but it does not seem that we can be much more precise than that. For example, if one wanted to rebuke or intervene with teachers in the bottom decile in terms of growth or reward those in the top decile, we would often select the wrong teachers.

There are two ways to lessen this problem (although there is no way to eliminate it entirely). One is to add more data, which one might do by combining each teacher's or school's results from several years (e.g., instead of just looking at my value added this year, you could average my results from this year plus the last two years). A second is an analytical approach, which brings us to uncertainties about how we should estimate growth.

One of the biggest failures of education policy in recent years has been the failure to adequately evaluate the accountability systems that were imposed on teachers and students. The movement toward value-added models exacerbates this because of serious gaps in our knowledge of their workings and effects.

4. How certain are we about how to model gains?

A variety of different statistical approaches are used to estimate value added. Most are highly complex, and while the differences among them seem extremely arcane, in this case, the old cliché really is true: the devil is in the details. The choice among methods can matter; it can influence, sometimes substantially, how a school or teacher is rated. And yet, other than the experts, few people understand how these models work or what the implications of the various choices are. Let's look at a handful of the more important technical issues.

One important issue is how to deal with the uncertainty caused by sampling error. All teachers will sometimes appear more or less effective than they really are because of sampling error, and substantially incorrect estimates will be much more common among teachers with smaller classes (or schools with smaller enrollments). One approach ignores the fact that these errors are worse in small groups and takes each group's estimate at face value. The alternative approach, called a "random effects model," compensates for the uncertainty by "shrinking" the estimates for each teacher or school back toward the average teacher or school, with more shrinkage for the groups with fewer students. In the aggregate, the latter approach seems preferable, because it compensates for small samples, puts large and small groups on the same footing, and reduces the number of instances in which a teacher or school is inappropriately rewarded and sanctioned because of sampling error. For individual teachers or schools, however, this approach is not necessarily fair. For example, if you happen to be an exceptionally effective teacher but have a small class, a random effects model will assume that the atypically rapid growth of your students reflects sampling error and will shrink it. Therefore, random effects models reduce one type of error but increase another: the probability of missing truly effective or truly ineffective teachers.

Another issue pertains to the persistence of the effects of teachers. Value-added models ask the question: how much has the year with you added to students' growth *given what prior experience contributed*? To answer that question, one first has to estimate those prior contributions, and the different ways in which various VAMs do this can affect how teachers are rated. Suppose you receive a group of students who had highly effective teachers the previous two years, and suppose that the students score very well at the end of your year with them too. To calculate your value added, one has to somehow subtract what the students would have known at the end of your year, given their prior experience. The more the effects of that prior good teaching persist, the less credit you deserve for the students' strong performance at the end of the year. One of the most common models, the "layered model," assumes that the impact of good or bad teaching persists forever without any lessening at all. (As a teacher, I find this hard to accept; I could only wish that everything my students learned persisted without any deterioration.) Other models, however, allow for an erosion of prior teachers' effects over time, giving you more credit (or blame) for the performance of students at the end of their year with you. Decisions about how to handle persistence can clearly influence how individual teachers or schools are rated.

Another choice is how to deal with missing data. All value-added models require longitudinal data, that is, data that track individual students over time. However, some students—and in some districts or schools, many students—do not have complete data. Their data may be missing for all manner of reasons: their

families moved, they were truant, they were assigned to a special class, and so on. What is important is that the students whose data are missing are often unlike those whose data are complete. Worse, we generally know only enough to discern that these students are different; we do not know enough about them to adjust for the effects of leaving them out of the calculation. Some of the VAMs can handle missing data, provided that the problem is not too severe, but it remains an open argument just how serious this problem has to be before it substantially biases estimates for some teachers.

Apart from the first of these issues, all of these are matters of bias, not error. For example, if we overestimate persistence, we will introduce a bias by systematically over- or underestimating the impact of teachers depending on the effectiveness of those who preceded them.

5. How well do we adjust for other influences on achievement growth?

To provide an unbiased estimate of the effects of teaching, value-added models must remove the impact of other influences on achievement growth. Teachers often express concern that the models now used will not do this well enough to be fair. For example, many teachers find that their effectiveness varies with the characteristics of their students. I certainly have; my style and methods of teaching work much better with some types of students than with others. If these effects are large, value-added models would have to take them into account.

Teachers are right to be concerned. On the positive side, a recent study* found that in one context, effectiveness as estimated by a value-added model was similar to true effectiveness measured by an experiment, but there are a number of reasons why we cannot in general assume that this is true.

One potentially important source of bias in the evaluation of teachers is called "interference." Suppose you want to evaluate the impact of providing after-school math tutoring, and you do this by randomly dividing students from a school into two groups and giving tutoring to only one of them. You give both groups a math test at the end of the year, and you use the difference in scores between the groups to evaluate the impact of the tutoring. This sounds like an ideal evaluation—a true experiment. The problem is that the tutored and untutored students interact with each other: they attend the same math classes, they may study together, and so on. This is interference: the effects of tutoring seep into the untutored control group, leading to a biased estimate (in this case, too low) of the impact of tutoring. Interference is a potentially severe problem in using VAMs to evaluate teachers because teachers are embedded in schools, and there are many sources of interference that could bias estimates for individual teachers. Interference could arise not just because of the instruction of other teachers, but because of administrative arrangements, peer effects, and so on. For example, in some secondary schools, teachers in subjects other than math and English have been instructed to incorporate more math and writing into their classes, which makes the value seemingly added by math and English teachers dependent in part on the other teachers their students are assigned to. For this reason, some researchers have warned that with the value-added models we have now, the effects of teachers cannot be entirely separated from those of the school context.

Apart from interference, there is an ongoing, intense debate about how well VAMs control for other factors that influence achievement growth, such as students' backgrounds. The adequacy of the models is likely to vary, depending on the context (for example, the degree to which students with similar characteristics attend the same classrooms and schools) and the methods used. The more similar the contexts in which two teachers work, the less these other factors come into play, and the closer a value-added model will come to an estimate of the teachers' impact. But in real-world situations in which the contexts of teaching vary markedly (even within a single school), research tells us that we can't assume that the results of our models give us a sufficiently unbiased estimate of the effects of teaching.

6. How does score inflation affect value-added models?

In this era of test-based accountability, one of the biggest problems confronting testing programs is *score inflation*: increases in test scores that are larger than the actual gains in learning they are thought to represent. Research has shown that score inflation is widespread and that it can be very large. Some studies have found score gains that are three to five times as large as they should be, and others have found large score gains that were not accompanied by any meaningful improvements at all. Score inflation results in both an illusion of progress and misleading comparisons of schools and teachers, both of which are detrimental to students. (A more detailed explanation of score inflation, as well as a discussion of the grey area between good instruction and inappropriate test prep, are included in the sidebar from *Measuring Up* on page 22.)

VAMs do nothing to address the problem of score inflation. There may be ways that policymakers can lessen this problem, such as relying on multiple measures, setting more realistic targets, and strengthening the role of human judgment in the evaluation of teachers and schools, but simply switching from status or cohort-to-cohort change models to a value-added approach will not do the trick.

Where Do We Go from Here?

For all the uncertainties and concerns about the use of value-added models, there is no question that they are in some important ways superior to the status and cohort-to-cohort change models that have dominated test-based accountability in the United States for the past 30 years. I believe that most people working in this area would agree with me that we should continue to look for appropriate ways to incorporate value-added modeling into accountability systems in order to capitalize on that superiority.

At the same time, to use value-added models sensibly, we can't treat them as a silver bullet. We need to find ways to use VAMs that take into account both their limitations and the uncertainties we still have about their functioning and impact.

First, we must recognize that value-added modeling remains

(Continued on page 39)

* S. Cantrell, J. Fullerton, T. J. Kane, and D. O. Staiger, *National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment* (National Board for Professional Teaching Standards, June 11, 2008, draft).

Value Added

(Continued from page 27)

a work in progress, a project that is in its adolescence in some respects and its infancy in others. Despite several years of intense work by a number of researchers, we still confront many uncertainties about the statistical and psychometric aspects of value-added models—that is, about the pros and cons of various ways of conducting the analyses and about the limitations of the results. There has been very little research on the practical effects of using VAMs—for example, how teachers' instructional responses compare with those under status or cohort-to-cohort change models. For the time being, using value-added models requires that we choose among alternative approaches with only limited information about the effects that our choices may have on the ratings of teachers or schools, or on the education experienced by students.

Second, we must accept the fact that value-added models, taken by themselves, are not an adequate measure of overall educational quality. Like any other measure based on standardized tests, VAMs provide a valuable but incomplete view of students' knowledge, skills, and dispositions. Because of the need for vertically scaled tests, value-added systems may be even more incomplete than some status or cohort-to-cohort systems. Value-added-based rankings of teachers are highly error-prone. And value-added modeling does nothing to address the interrelated, core problems of an excessive focus on standardized test scores in an accountability system: undue narrowing of instruction, inappropriate test preparation, and the resulting inflation of test scores.

Finally, we have to accept that even within the range of outcomes assessed by the tests used in VAMs, they cannot be counted on to give us true estimates of teachers' value added as opposed to students' overall growth (which has many causes). Although VAMs generally do much better than status and cohort-to-cohort change models in removing the confounding effects of other influences on achievement, we cannot assume at this stage that they will always do this as well as they would have to in order to be trustworthy measures of teachers' effectiveness.

How can we use VAMs in a way that takes these limitations into account and is nonetheless productive? Given the pending reauthorization of NCLB, this is a pressing question. However, given the uncertainties I have described, it should be no surprise that there is no consensus about this. I can only offer my own suggestions:

1. Consider using value-added models rather than cohort-to-cohort or status approaches where appropriate—for example, in elementary school reading and mathematics. But do not let the particular requirements that VAMs impose lead to further narrowing of the accountability system. How much science high school students learn is very important, and if we can't address that with a value-added system, we should address it in some other way.

2. If VAMs will be used, state tests must be constructed from the ground up to be appropriate for this purpose—that is, to support a vertical scale that allows for sensible comparisons from one grade to the next. Efforts to graft VAMs onto grade-specific tests and standards are bad practice.*

3. Use VAMs only with full recognition of the imprecision they entail. Don't pretend that the estimates of teacher or school effectiveness are more precise than they really are. To lessen the impact of this imprecision, add more data, ideally from more years of testing and from other sources entirely. And do not make the consequences of the scores more substantial than the level of precision warrants.

4. Use VAMs primarily to compare classes or schools that start at fairly similar levels of performance. For a number of reasons, comparisons of growth become less and less trustworthy as the initial difference between groups becomes larger. (One reason is that, as explained earlier, the difference between 120 and 140 may not be the same as the difference between 200 and 220.)

5. Don't use test scores as the sole focus of the accountability system. Research in many other fields shows that using too narrow a set of outcomes in an accountability system generates undesirable behavior and distortions in the measured outcome. Evaluations have shown that in the case of test-based accountability systems, these distortions can be severe indeed. VAMs do nothing to lessen this problem.

6. And finally: evaluate, evaluate, and evaluate more. By this, I do not mean testing students more; I mean evaluating the accountability programs themselves. One of the biggest failures of education policy in recent years has been the failure to adequately evaluate the accountability systems that were imposed on teachers and students. We have done enough research to show that the systems do not work as we would like, but we have not done enough to guide the development of better systems. The movement toward VAMs only exacerbates this problem because of the remaining serious gaps in our knowledge of their workings and effects. We need ongoing, independent evaluations to help guide midcourse corrections. For example, we should evaluate the imprecision in value-added estimates, inconsistencies across alternative approaches, the extent of score inflation and other possible biases, and the effects on educational practice and student learning. Our children deserve no less. ☐

* If we want to measure growth well, we will need to put aside standards-based reporting entirely and go back to more traditional scales. This would have other benefits, as the recent change to standards-based reporting was in many respects a bad decision. This is discussed at some length in *Measuring Up*.

For Further Reading

For those interested in reading more about VAM, two sources written for nontechnical audiences are the following:

RAND Corporation. 2004. *The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. Research Brief.* Santa Monica, CA: RAND Corporation. **http://www.rand.org/pubs/research_briefs/RB9050/index1.html**.

Braun, Henry. 2005. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Princeton, NJ: Educational Testing Service. **http://www.ets.org/Media/Research/pdf/PICVAM.pdf**.

A much more detailed but still relatively nontechnical source, which includes discussion of many of the points made here and which was the basis for the RAND research brief noted above, is:

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND Corporation. **http://www.rand.org/publications/MG/MG158**.