Center for American Progress

FLICKR/OLD SHOE WOMAN

Adding Value to Discussions About Value-Added

A New Framework for Talking About Teacher Effectiveness

Raegen Miller December 2009

Center for American Progress

Adding Value to Discussions About Value-Added

A New Framework for Talking About Teacher Effectiveness

Raegen Miller December 2009

The Center for American Progress thanks The Bill & Melinda Gates Foundation for generously providing support for this report.

Contents

Introduction

The quality of the U.S. teacher workforce is under the microscope, and rightly so. Teachers represent the most important school-based resource determining students' academic success,[1] and a shortage of graduates with knowledge and skills necessary to drive innovation or to command premium wages in a global economy threatens the nation's economic prosperity.[2] Moreover, children from low-income families and children of color are disproportionately assigned to the least effective teachers,[3] a finding that helps explain yawning gaps between average educational outcomes of groups defined by family income or ethnicity. Broad improvements in teacher quality will thus serve the strategic goals of raising student achievement overall and reducing disparity in achievement between groups.

Past initiatives to improve teacher quality offer two general lessons. First, simplistic responses—across-the-board raises, more stringent licensure requirements, mandated professional development—are extremely expensive, utterly ineffective, or both. Only policies that tightly link incentives to desired results stand a chance of being effective and affordable. Clearly, making such links requires a robust approach to assessing teachers' impact on outcomes of interest, especially student achievement. Second, teachers must be involved in crafting and implementing policies aimed at improving their instructional potency.[4] These lessons together highlight the need for a language of productivity calibrated to education.

The need for such a language would probably astonish Adam Smith, the father of modern economics. Smith wrote, "There is one sort of labour which adds to the value of the subject upon which it is bestowed,"[5] and he was not writing about teaching. Smith was concerned with manufacturing instead. And the conventional language used to discuss productivity today— especially the term "value-added"—is well-suited to that sector of the economy. In elementary and secondary education, however, the use of the term value-added has proved problematic. Although widely embraced by researchers and policymakers to denote estimates of teachers' productivity, typically referred to as effectiveness, the term value-added "sends chills down the spine" of most teachers union officials.[6]

Why such a visceral reaction? The usual explanation trots out a series of concerns with estimates of teacher effectiveness, but as important as these concerns are, they miss a crucial point: The actual term value-added may be partly to blame. This paper unpacks this new and complementary explanation, and it constructs an alternative to the term value-added better suited to conversations—especially ones involving teachers—about the use of estimates of teacher effectiveness in education policies.

New terminology, of course, does not address legitimate concerns about estimates of teacher effectiveness derived from student achievement data. Accordingly, this paper goes on to offer a conceptual framework for appreciating these concerns, and specific guidance for addressing them. Combined, the new terminology, the framework, and the guidance form a set of tools that may be put to good use immediately, especially in states planning to apply for competitive Race to the Top funds.

A failure to communicate

Inability of various stakeholders to communicate frankly and directly around the concept of teacher effectiveness is a roadblock to promising reforms. Part of the problem is the sheer novelty of such communication.[7] The chief finding of the 1966 Coleman report—that family background is the best predictor of student achievement—shielded teacher effectiveness from scrutiny for many years.[8] This paradigm began to shift with the 1983 publication of *A Nation at Risk* and the subsequent standards movement.[9] The emergence of datasets linking information on student achievement to teachers over time, notably in Tennessee, allowed social science researchers to begin explaining within-school variation in student achievement with estimates of teacher effectiveness furnished by advanced statistical methods.

Applying the name value-added to these estimates made perfect sense, especially to labor economists, and policymakers and others basically adopted the term by default. Because the methods make an explicit gesture toward fairness by controlling statistically for characteristics of students or schools that affect student achievement, especially students' prior levels of academic achievement, nearly everyone using the term may actually feel as though they are bending over backwards to respect teachers. Unfortunately, the very term value-added and close relatives of it—such as "teacher effects"—may be objectionable to teachers.

What's in a name?

That names matter is a principle well understood in the world of commerce. "Patagonian toothfish," for example, sounds strange and menacing, but when sold under the tamely exotic name "Chilean sea bass," a surge in consumption nearly destroyed the fishery.[10]

The conventional nomenclature for productivity centers on the term value-added, and in manufacturing and many other settings, the productivity of an individual worker or a team can be described sensibly and succinctly as a matter of adding value. A given amount of labor produces a clearly defined output with some market value, subtracting from which the costs of labor and other inputs yields, literally, a value-added. And while value-added indicators of productivity ignore potentially important nuances of performance such as collaboration, they offer a reasonable and well-accepted basis for aligning workforce policies and strategic goals. In education, by contrast, such acceptance has been harder to come by, perhaps in part because the term value-added strikes teachers as alienating and deceptive.

Alienation

The work of teaching is complex, but test-based accountability has focused attention on one dimension of practice: A teacher's ability to produce gains in student achievement. Statistical techniques associated with the term value-added offer useful information in this regard, but the very term value-added, when used publicly to describe a teacher's strength on one dimension of practice, devalues the other dimensions by implication. A survey conducted by Public Agenda offers evidence that teachers may resent the focus on test scores conveyed by the term value-added. Across outlooks and generations, three of every four teachers feel that test scores are less important than a lot of other measures.[11]

Should teachers just "get over" the term value-added? It seems unlikely that they will, since more than half of them operate in grades, subjects, or specialties devoid of the test scores necessary to estimate their individual contributions to student learning.[12] Even if these teachers take no offense at being omitted from testing regimes, it does matter to them that test-driven accountability has tended to pull resources away from un-tested subjects.[13] Thus, it is little wonder that groups representing large swaths of teachers take issue with value-added. There are many reasons why these groups may be hesitant to embrace some of the policies that value-added indicators of effectiveness make possible,[14] but it does not help these policies' cause if key stakeholders abhor the terms on which the policies pivot.

Crouching estimates, hidden uncertainty

Nor is it helpful that the term value-added belies a sense of bottom-line certainty that is not justified. The statistical methods behind value-added indicators of teachers' effectiveness—inscrutable to virtually all teachers—produce estimates, and even the most sophisticated estimates are subject to error, bias, and misinterpretation. This is especially concerning where measures of academic achievement are involved. [15]

The threat posed by error has many facets. A particularly troubling one is encapsulated by the finding that value-added estimates are sensitive to the choice of achievement test.[16] The reason is that tests include items representing a modest fraction of the content standards that guide a teacher's instruction in academic subjects as broad a fourth grade math or eighth grade English language arts, for example. Thus, one test may be heavy with items corresponding to standards that a teacher teaches to well, while another test has many fewer such items.

Concern that estimates may be biased rears its head because student-teacher matches are not random. Teachers tend to prefer to work in schools with better working conditions, and some combination of principal reasoning and parental jockeying usually informs the development of classroom rosters. Such dynamics tend not to be documented, much less incorporated into statistical models of effectiveness.[17] To the extent that the instructional challenges represented by classroom rosters are systematically harder or easier for a teacher than would appear to be the case based on information available to the model, his or her value-added estimate will be biased. That bias can run either way is cold comfort to risk-averse teachers.

Value-added estimates—like any estimate of teacher effectiveness—have to be taken with a grain of salt, as illustrated by this example. A group of students may enter ninth grade having scored below average on state mathematics tests in previous years. The students all wind up in the same science class, and in the same math class. As it happens, their science teacher is highly engaging, and the students work hard to master the material, even staying after school frequently to get extra help on the more mathematically oriented science topics. The hard work pays off, and the students find themselves thriving in science, and in mathematics. The latter success is not due to their mathematics teacher being particularly effective, however. Rather, the students' experience in science affords them the key mathematical knowledge and the persistence and confidence to succeed in mathematics despite having a decidedly mediocre mathematics teacher.

This hypothetical example highlights the danger of attributing students' achievement gains to a particular teacher.[18] Indeed, some empirical evidence of misattribution seems to militate against using value-added estimates for any purpose.[19] Yet policymakers and school officials live in a world where the entire enterprise of schooling is premised on the idea that teachers actually cause students to learn. Given that reality, it has to matter that value-

added estimates of effectiveness make better predictors of future teaching success than any other policy-relevant predictors.[20] Furthermore, as a matter of fairness, if the results of achievement tests matter for students, they should somehow matter for teachers.

Owing to the inescapable problems of imprecision, inaccuracy, and interpretation, it is reasonable, however, for teachers to expect that uncertainty inherent in indicators of their effectiveness be somehow present during discussions with them about it. Gaining widespread acceptance for these indicators among teachers will require more than improved terminology, but this fundamental barrier to communication should be dealt with head-on. Fortunately, terms that avoid alienating teachers or obfuscating uncertainty are readily available. A service-oriented profession other than teaching has shown the way.

Value-added, heal thyself

At a practical level, teachers and doctors do very different work, but at an abstract level, their work lines up in three ways. First, both professions entail complex work comprising a core function surrounded by complementary and ancillary ones. Doctors and teachers engage their clients with the aim of promoting health or learning, respectively, but they also provide counseling and advice, communicate with clients' families, and mentor less experienced colleagues.

Second—for both doctors and teachers—variation in the contexts in which they practice makes it difficult to rate their effectiveness fairly. Many factors outside of their control influence outcomes. Whether students or patients live in concentrated poverty, for example, affects the nature of the instructional or medical challenges that they present, on average. In either profession, indicators of effectiveness based on outcomes data have to account statistically for such factors, when possible.

Finally, both professions embrace a wide variety of practices, but some practices attract more attention around the issue of quality than others. In part, this is because relevant outcomes vary across practices in clarity, and in their susceptibility to measurement. In this sense, it may be more difficult to peg the effectiveness of an art teacher or psychiatrist than it is for a third-grade teacher or a cardiac surgeon.

Risk-adjusted mortality rates

The point of this narrow teacher-doctor analogy is that in medicine there already exists an operational vocabulary for value-added measures that avoids the pitfalls of the term value-added. Specifically, estimates that go by the name "risk-adjusted mortality rates" offer a public window on the effectiveness of cardiac surgeons (and hospitals) in New York State.[21]

The term "risk-adjusted" may not spell out the statistical procedures that yield the rates, but it does convey a sense that they are estimates, as opposed to un-adjusted raw rates. Deception averted. Furthermore, risk-adjusted explicitly signals that the estimates account for key aspects of the context in which cardiac surgeons practice such as patients' risk profiles, which include the nature and severity of cardiac illness, other current or prior illnesses, age, smoking history, and so on. The term "mortality rates," grim as it may seem, anchors

the entire phrase squarely in the bailiwick of cardiac surgeons: performing operations to improve the quality of patients' lives in the long-term, or to avert death in the short-term.[22]

Do all doctors like the sound of risk-adjusted mortality rates? Some may not, but surely the great majority of them can agree that risk-adjusted mortality rates speak to job performance on a dimension of practice of central interest to the patients and employers of cardiac surgeons. Moreover, the phrase risk-adjusted mortality rates casts no shadow on the quality of cardiac surgeons' work on other dimensions of practice such as bedside manner or supervision of interns. Nor does an ophthalmologist, for example, feel slighted for lack of a risk-adjusted mortality rate.

Context-adjusted achievement test effects

With the medical example as a guide, a teacher-friendly alternative to the term value-added almost writes itself: context-adjusted achievement test effects. The term "context-adjusted" conveys the idea that numbers under this banner are estimates, as opposed to some type of ranking or average based on raw test scores. More importantly, the term signals that some attempt to account for factors other than teacher skill and knowledge that affect student achievement. Clearly, a name is no place for full disclosure of the methodology involved, but it is a good place for a respectful signal of intent. The term "achievement test effects" pins the entire phrase to the dimension of practice at issue, the core duty of promoting learning.

By design, the phrase context-adjusted achievement test effects avoids alienating or deceiving teachers, but it has other virtues that may support its widespread use in discussions about teacher quality. First, the phrase can be abbreviated as CAAT Effects, or more memorably, CAATs. This paper focuses exclusively on Teacher CAATs, so "Teacher" is henceforth omitted for convenience. However, other productivity discussions may benefit from the obvious meaning of School CAATs, Fourth-Grade CAATs, or District CAATs. Second, the phrase is portable. It is not tethered to any particular state, contractor, or academic domain. And while honest about adjusting for context, CAATs is impartial as to the specific approach involved. This matters because the choice of statistical procedures used to estimate CAATs should reflect properties of the data available, and the intended uses of the estimates, by whatever name.

Decisions, decisions

The idea that CAATs should inform decisions about teachers is anathema to many. Take decisions about continued employment, for example. Teachers can be forgiven for worrying about the imposition of crude, unfair policies characterized by the phrase "sort and fire," given public education's history of half-baked, faddish, and fleeting reforms.[23] Such rhetoric highlights a need for a framework that facilitates discussion and enables the use of CAATs in ways that respect the gravity of the decisions they inform.

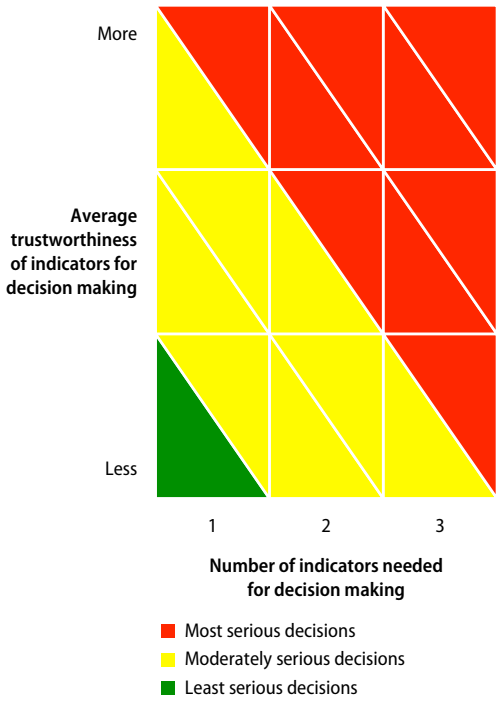
Two principles are especially useful in respectfully framing the use of information about teachers' effectiveness. First, the more serious a decision is, the more important it is that multiple indicators of effectiveness inform the decision. Second, the more serious a decision is, the more important it is that indicators of effectiveness be trustworthy.[24] Figure 1 illustrates a policy-analytic framework based on these principles. The gravity of a decision is represented by color. The green zone comprises low-stakes decisions; the red zone comprises the gravest decisions. Moderately serious decisions inhabit the yellow zone. The number of indicators of effectiveness available to inform a decision increases to the right.[25] A higher position on the vertical axis represents greater average trustworthiness of the indicators.

This framework offers a tool to facilitate new discussions about proposals to use CAATs to inform decisions about teachers. It does so by focusing attention on two questions that have nothing to do with CAATs. This may seem counterintuitive, but focusing on these questions may prevent parties from sinking into irreconcilable positions based on preconceived notions about CAATs.

FIGURE 1

Effectiveness-based decisions

A conceptual framework relating the gravity of a decision to be based on indicators of teacher effectiveness to the number and trustworthiness of the indicators



How serious is the decision relative to other ones?

Some of the decisions that reformers would like to be informed by CAATs are more serious than others. Views about the gravity of decisions about school placement, teaching assignment, and professional development, for example, are bound to vary with perspective and

experience, yet there may be a great deal of consensus around some decisions. Few would dispute, for example, that the decision to terminate a teacher's employment is a terribly serious one. In contrast, few would maintain that the decision to award a $600 bonus to especially effective teachers is anywhere near as serious.

Parties may disagree vehemently about how serious a given decision is, but that is OK. It is the process of surfacing agreement or disagreement that matters, in two ways. First, it gets the ball rolling with an opportunity for parties to demonstrate credibility. Weighing in on how serious a decision is requires no specialized knowledge about the properties of CAATs or other measures of effectiveness, and claiming ignorance would undermine one's credibility as an agent for any party in the discussion. Furthermore, nobody can credibly locate all decisions at the extreme end of a spectrum of seriousness. Second, the process supplies information that is highly relevant to how stakeholders should feel about CAATs.

Which other indicators of effectiveness will inform the decision, and how trustworthy are they?

There are many indicators of teachers' effectiveness, and which ones happen to be available should matter as should their trustworthiness. Vague checklists of teacher behaviors and classroom attributes, the cornerstone of perfunctory evaluation systems, are low-quality indicators of teacher effectiveness. Summaries of observation notes keyed by rubrics to descriptions of effective teaching can be valid and reliable indicators, provided sufficient investment is made in training observers, among other things. Research and development on assessment systems exploiting video technology may eventually yield trustworthy and relatively inexpensive measures of teacher effectiveness. Some sources of information, however, may be important politically despite being subjective—surveys of student and parent opinions about teachers' effectiveness, for example.

Disclosure around the number and caliber of indicators of effectiveness that will inform the decision in question sets the stage for a more dispassionate and informed conversation about the trustworthiness of CAATs than might otherwise ensue, as illustrated in the following vignette.

Test driving the framework

Suppose the trustees of Blue Briar Public Schools have empowered its superintendent to bargain with the teachers' union to fold CAATs into decisions about granting tenure to teachers who have completed a three-year probationary period. Instead of talks seizing up immediately based on preconceived ideas about CAATs, the framework facilitates the following initial exchange:

Superintendent's team: We'd like to start by talking with you about how serious the decision to grant tenure is. We think it's serious, but not as serious as the decision to terminate a teacher mid-year. After all, if a teacher does not get tenure, he or she remains employed until the end of the year, and has plenty of opportunity to line up another teaching job in another district for the next school year.

Teachers' team: We think tenure is serious, too. Getting tenure is a career milestone. Not getting tenure can be embarrassing or even depressing. In terms of seriousness, the tenure decision is way up there.

Superintendent's team: Currently, we base the decision to grant tenure on supervisors' evaluation reports. These reports are based on three formal observations of a teacher's teaching—as per contract—and the administration provides a one-day professional development seminar every summer for principals and other evaluators new to Blue Briar. It's important to make sure everyone's on the same page because our teaching standards are a little different than other ones out there.

Teachers' team: Yeah, we know. We helped write the standards, remember? What concerns us most is that principals get bullied by parents into being tougher on some teachers than others. We hear a lot of scuttlebutt about this, but it's hard to document.

Superintendent's team: Parents have every right to raise concerns with our principals, but we're certainly aware of the danger you've mentioned. It's why we take measures to protect the integrity of the evaluation process. And this brings us to the subject of context-adjusted achievement test effects or CAATs. We think it makes nothing but sense to combine CAATs with evaluation reports, in grades and subjects where it's possible, when making tenure decisions. Given your concern about principals' objectivity, we think you would be pleased to bring an additional source of information into the mix, especially one that may be more objective than anything else.

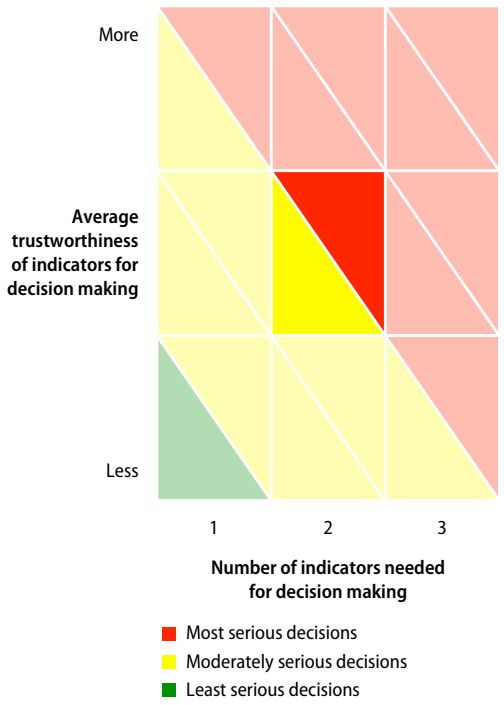
Teachers' team: We're not thrilled about adopting CAATs. The tenure decision is made toward the end of year three, so only two years' worth of test scores are available at that time, right? Not to mention, you're leaving out lots of teachers.

The superintendent's proposal has some distance to go before implementation, but the exchange situates concern with CAATs in a specific part of the framework, as depicted in Figure 2. The decision to grant tenure falls somewhere between most serious and moderately serious, and CAATs would be the second source of information to be brought

FIGURE 2

Effectiveness-based decisions

Hypothetical decision to include CAATs in tenure decision situated in a conceptual framework (center square with bold colors) relating the gravity of a decision to be based on indicators of teacher effectiveness to the number and trustworthiness of the indicators.



to bear on the decision. The other source, evaluation reports, is neither foolproof nor rubbish in Blue Briar. Thus, the key question about CAATs is whether they are reasonably trustworthy indicators of teacher effectiveness.

Whether CAATs are reasonably trustworthy should be the key question no matter how it comes up, but getting to it by way of the framework has two advantages. First, using the framework clarifies that increasing the trustworthiness of evaluation reports is a legitimate strategy for increasing the acceptability of a role for CAATs in the decision to grant tenure. This clarification arrests any misguided thoughts that somehow CAATs can supplant other indicators of teacher effectiveness, and it places an affirmative burden on the superintendent's team to bolster the trustworthiness of existing indicators of effectiveness.

Second, using the framework reveals that this specific use of CAATs proposed does not call for the very highest standard of trustworthiness. This revelation short-circuits the defensive tactic of tearing down the trustworthiness of CAATs. Reflexively rattling off a litany of problems with the statistical properties of CAATs is not constructive, and to some extent, the framework orients the teachers' team towards the notion that appropriate trustworthiness can be ensured. Suitable terminology and a framework that prevents discussions from falling into predictable ruts are fine and good, but a successful embrace of CAATs will always hinge on this notion.

Due diligence

The subject of the trustworthiness of CAATs tends to be approached from a technical standpoint, but such an approach is seldom edifying—even for the technically savvy. Furthermore, a descent into the psychometric properties of achievement tests and the jargon of statistical methods for estimating CAATs is not liable to further discussions that are, when push comes to shove, political ones. It would behoove advocates for the use of CAATs to ensure appropriate trustworthiness by heeding the following principles.

1. Make correct comparisons

One attraction of CAATs to policymakers and school officials is their ability to rank teachers on a dimension of practice that is strategically important. It turns out, however, that there are many ways to rank teachers, so the specific ranking scheme implied by a set of CAATs should be made clear. Are teachers to be ranked in relation to colleagues within their school, across their district, or even across their state? Do rankings span multiple grade-levels or subjects, or are they confined to a single grade-level and subject, fourth grade mathematics, for example?

These questions can be answered in plain English, and the answers had better make sense in light of the decision to be informed by CAATs. Preposterous schemes have a way of standing out when the underlying approach to ranking teachers is ill-conceived. It would not make much sense, for example, to use within-school rankings as a basis for offering bonuses to highly effective teachers who move to high-poverty schools. Nor would it make sense to use any ranking of teachers as the sole basis for evaluating teachers, since for the most part their practices should be stacked up against absolute standards.

2. Use moving averages

The second guiding principle is that rankings are more trustworthy when they are based on data from longer intervals of time. Weekly rankings of Major League Baseball players by batting average are quite volatile, for instance, but a ranking of cumulative seasonal batting averages provides a strong basis for deciding a player's fate with a team, or for awarding special bonuses for offensive production. Similarly, rankings of teachers by indicators

of effectiveness exhibit year-to-year volatility.[27] Real changes in effectiveness are part of the story, especially because teachers' effectiveness tends to rise with experience at least for a handful of years.[28] Nonetheless, disentangling the sources of volatility in yearly rankings of teachers is as hopeless an errand as doing so for weekly rankings of baseball players by batting average.

For many purposes, this principle can be applied to enhance the trustworthiness of CAATs. In Tennessee, for example, three-year moving averages of CAATs inform a variety of decisions in a multitude of local programs.[29] In the above discussion between superintendent and teacher team at Blue Briar, however, a three-year moving average is not available to inform tenure decisions for which data from two years, at most, are available. According to the principle, two is better than one. In any case, discussions about the potential use of CAATs to inform a decision should be grounded solidly in a sense of how well trustworthiness has been bolstered by using multiple years' data.

3. Use three bins

The third principle is that sorting teachers into bins by indicators of effectiveness is more error prone the more bins there are. In this sense, teacher effectiveness is no different than, for example, tire pressure. Given a cheap tire-pressure gauge, one can sort inflated tires into bins labeled 28 psi, 29 psi, 30 psi, and so forth. A few tires will go into the wrong bins because of imprecision inherent in a cheap gauge, but most will land in the bin best matching their unknown true pressure. If, however, bins were labeled 28.00 psi, 28.25 psi, 28.50 psi, and so forth, quite a few tires would go into the wrong bin.

It is tempting to make fine grained distinctions tantamount to using too many bins given a ranking of teachers by numerical estimates of their effectiveness. The problem is that engineering a gauge of teacher effectiveness is not a physics problem. The relevant science is much less settled, and instead of wading straight into this quicksand, it makes sense to formulate policies using as few bins as possible.

Fortunately, many decisions can be informed using just three bins. Some teachers are highly effective, some are chronically ineffective, and the rest are harder to characterize. Specific decisions tend to focus on one of these groups. Which teachers should be offered a bonus for transferring to a high-poverty school? Which teachers should be the focus of an extra thorough performance evaluation, perhaps leading to dismissal? In which teachers should a district invest scarce professional development resources? Strategic management of schools demands answers to these types of questions, so it makes little sense to sacrifice trustworthiness of CAATs by using them to parse teachers into more than three groups.[30]

Full steam ahead

The urgency of improving the quality of the teacher workforce is hard to overstate, but this is no license for recklessness. Teachers are right to be cautious about the use of student achievement data to inform workforce policies given the history of education reform, but there is a point at which caution becomes negligence. Doctors understand this. Despite swearing the Hippocratic Oath, "First, do no harm," doctors are faced with a moving bar as to what constitutes the appropriate standard of care and what constitutes harm. And this dynamic varies across practices at a pace dictated by the publication of relevant scientific research papers.

The bar is moving in education, too. Comparable student achievement data is ever more widely available. Data systems linking information on students and teachers are improving at a rapid clip. And a growing body of research highlights the benefits and drawbacks of using estimates of teachers' impact on student achievement to inform workforce policies aligned with the strategic goals of increasing student achievement overall and providing for a more equitable distribution of teaching talent.

Teachers may have the political power to ignore this moving bar, but doing so only reinforces the popular belief that their profession is hostile to accountability, which in turn constrains public funding for education, the prestige of teaching, and the caliber of college graduates willing to enter the profession. There is reason to believe that teachers understand this. In the words of Randi Weingarten, president of the American Federation of Teachers, "With exception of vouchers … no issue should be off the table."[31] It is up to policymakers and school officials, then, to invite teachers to the table.

Inviting teachers to the table is one thing, but having constructive discussions there is another. This paper has offered three tools to facilitate constructive discussions. First, because the term value-added may strike teachers as alienating and deceptive, the phrase context-adjusted achievement test effects, or CAATs, should be used to denote indicators of teacher effectiveness formerly known as value-added. Second, the conceptual framework relating the seriousness of decisions to the number and trustworthiness of indictors of effectiveness can act as an icebreaker for difficult discussions about potential uses of CAATs. By focusing attention on the nature of a decision in question, the framework may temper expectations about the trustworthiness of CAATs, which will never be 100 percent. Third, a set of three guiding principles establish a baseline of due diligence for those crafting proposals to use CAATs to inform decisions about teachers. Subscribing to these principles bolsters trustworthiness, a lack of which is probably the greatest threat to the legitimacy of proposals to use CAATs to inform decisions about teachers.

Endnotes

1 Daniel Aaronson, Lisa Barrow, and William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25 (1) (2007): 95-135; Steven Rivkin, Eric Hanushek, and John Kain, "Teachers, Schools and Academic Achievement," *Econometrica* 73 (2) (2005): 417-58; Jonah E. Rockoff, "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94 (2) (May 2004): 247-252; Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, "Identifying Effective Teachers using Performance on the Job" (Washington: The Brookings Institution, 2006).

2 McKinsey & Co., "The Economic Impact of the Achievement Gap in America's Schools," April, 2009, available at http://www.mckinsey.com/App_Media/Images/Page_Images/Offices/SocialSector/PDF/achievement_gap_report.pdf.

3 The following evidence enlists proxies of effectiveness (qualifications). These proxies are weak predictors of individual effectiveness, but as aggregate measures of workforce characteristics, they provide reasonable support for the claim of disproportional assignment: Charles T. Clotfelter and others, "High-Poverty Schools and the Distribution of Teachers and Principals" (Washington: Urban Institute, 2007); Hamilton Lankford, Susanna Loeb, and James Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation & Policy Analysis* 24 (1) (2002): 37-62; Heather Peske and Kati Haycock, "Teaching Inequality: How Poor and Minority Students are Shortchanged on Teacher Quality" (Washington: Education Trust, 2006).

4 Nick Anderson, "Education Chief Urges Union to Aid Reform Push," *The Washington Post*, July 14, 2009, available at http://www.washingtonpost.com/wp-dyn/content/article/2009/07/13/AR2009071303058.html.

5 Book 2, Chapter 3 of *The Wealth of Nations, Adam Smith*, 1776.

6 Steven Brill. "The Rubber Room: The Battle over New York City's Worst Teachers." *The New Yorker*, August 31, 2009, available at http://www.newyorker.com/reporting/2009/08/31/.

7 Lesli A. Maxwell, "Human Capital Key Worry for Reformers," *Education Week*, December 3, 2008, available at http://www.edweek.org/ew/articles/2008/12/03/14human_ep.h28.html?qs=value+added+communication.

8 James S. Coleman, "Equality of Educational Opportunity Study" (Washington: U.S. Department of Health, Education, and Welfare, 1966).

9 Daniel Fallon, "Case Study of a Paradigm Shift: The Value of Focusing on Instruction", (Education Commission of the States, November, 2003) available at http://www.ecs.org/clearinghouse/49/00/4900.htm.

10 Bruce G. Knecht. *Hooked: Pirates, Poaching, and the Perfect Fish*. (New York: Rodale Books, 2006).

11 Jean Johnson, Andrew Yarrow, Jonathan Rochkind and Amber Ott, "Teaching for a Living: How Teachers See the Profession Today," Public Agenda, October 2009, available at http://www.publicagenda.org/pages/teaching-for-a-living.

12 Estimates of the percentage of teachers whose impact on student achievement is not susceptible to measurement vary. Any estimate is based on premises as to what kinds of assessments furnish acceptable measures of student achievement before and after experiencing a teacher's instruction. On the one hand, limiting such assessments to the annual state-sponsored ones satisfying the requirements of No Child Left Behind (Mathematics and English language arts in every grade from 3 to 8 and one high school grade; science in at least one of grades 3-5, one of grades 6-9, and one of grades 10-12 by 2007–08) results in an elevated percentage. On the other hand, a vision including district-wide common assessments and the more flexible computational approach (less reliance on computation of achievement gains and more reliance on statistical controls for prior achievement) yields a lower percentage.

13 Sam Dillon, "Schools Cut Back Subjects to Push Reading and Math." *The New York Times,* March 26, 2006, available at http://www.nytimes.com/2006/03/26/education/26child.html.

14 The National Education Association's comments on the Department of Education's proposed guidance for the $4.35 billion Race to the Top program highlights a broad antipathy towards the use of test scores, and value-added estimates of teacher effectiveness by extension, as the basis for decisions, available at http://www.nea.org/home/35447.htm.

15 Daniel Koretz, "A Measured Approach." *American Educator*, Fall 2008.

16 Tim R. Sass, "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." (National Center for Analysis of Longitudinal Data in Education Research, Brief 4, November, 2008) available at http://www.caldercenter.org/upload/teacher_compensation_policy.pdf.

17 Teachers, of course, also lobby for particular grade or course assignments. Such lobbying introduces additional unobserved relationships that may potentially bias estimates of teacher effectiveness.

18 Additional examples involving job-sharing and team-teaching would only punctuate the case. In general, whether students are exposed to a single teacher for the duration of a grade or course, or whether the "dosage" of teacher is moderated by variable attendance, is a matter that researchers leave alone.

19 Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics,* forthcoming, available at http://www.princeton.edu/~jrothst/published/rothstein_vam_may152009.pdf.

20 Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, "Identifying Effective Teachers using Performance on the Job" (Washington: The Brookings Institution, 2006).

21 Risk-adjusted mortality rates are more commonly estimated at the hospital level, but as this paper focuses on teachers, it makes sense to call attention to a setting where doctors are the focus of value-added estimates of effectiveness, available at http://www.health.state.ny.us/diseases/cardiovascular/heart_disease/docs/2004-2006_adult_cardiac_surgery.pdf.

22 Any causal relationship between a specific medical intervention and a patient's eventual death becomes less clear over time. Accordingly, risk-adjusted mortality rates focus on a postoperative window of 30 days during which a surgeon's efficacy is most suspect.

23 Jackie Bennett, "Value-Added Accountability Requires Context," EdWize Blog, December 18, 2008, available at http://www.edwize.org/value-added-accountability-requires-context.

24 The word "trustworthy" is used consciously here as an accessible stand-in for the more formal terms "validity" and "reliability." Validity speaks to a measure's fidelity to the construct that it aims to measure; reliability speaks to consistency of a measure across instances of measurement.

25 The maximum number of indicators of effectiveness available to inform a decision is somewhat arbitrary. Three indicators serves the purpose of illustration, however, and a much larger number does not seem indicated in any case. Moreover, there may be aspects of job performance other than the ability to induce student achievement that would be appropriate to consider for a given class of decision.

26 Morgaen L. Donaldson, "So Long Lake Wobegon: Using Teacher Evaluation to Raise Teacher Quality" (Washington: Center for American Progress, 2009); The New Teacher Project, "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness," (2009).

27 Daniel Goldhaber and Michael Hansen, "Assessing the Potential of Using Value Added-Estimates of Teacher Job Performance for Making Tenure Decisions" (National Center for Analysis of Longitudinal Data in Education Research, November, 2008), available at http://www.urban.org/UploadedPDF/1001265_Teacher_Job_Performance.pdf; Sass, "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy"; Aaronson, Barrow, and Sander, "Teachers and Student Achievement in the Chicago Public High Schools."

28 See note 1.

29 See Elena Silva, "The Benwood Plan: A Lesson in Comprehensive Teacher Reform." (Washington: Education Sector, 2008) for a description about one plan in Hamilton County; for more examples, see http://www.tennesseescore.org/index.cfm?Page=PromisingPractices.

30 Groups do not necessarily need to be equal in size. A 20-60-20 breakdown formed by the top quintile, the bottom quintile, and the three middle quintiles has a certain appeal.

31 Randi Weingarten, "Making the Right Choices for Education and the Economy," a speech at the National Press Club, November 17, 2008, available at http://www.aft.org/presscenter/speeches-columns/speeches/downloads/npc_171108/NPCSpeech_Written.pdf.

The Center for American Progress is a nonpartisan research and educational institute dedicated to promoting a strong, just and free America that ensures opportunity for all. We believe that Americans are bound together by a common commitment to these values and we aspire to ensure that our national policies reflect these values. We work to find progressive and pragmatic solutions to significant domestic and international problems and develop policy proposals that foster a government that is "of the people, by the people, and for the people."

Center for American Progress

1333 H STREET, NW, 10TH FLOOR, WASHINGTON, DC 20005 • TEL: 202-682-1611 • FAX: 202-682-1867 • WWW.AMERICANPROGRESS.ORG