TEACHER QUALITY 2.0

AMERICAN ENTERPRISE INSTITUTE

SPECIAL REPORT 4

Anticipating Innovation in Teacher Evaluation Systems

Lessons for Researchers and Policymakers

By Michael Hansen | March 2013

Foreword

There is incredible interest and energy today in addressing issues of human capital in K–12 education, especially in the way we prepare, evaluate, pay, and manage teachers. States have been developing and implementing systems intended to improve these practices, with a considerable push from foundations and the federal government.

As we start to rethink outdated tenure, evaluation, and pay systems, we must take care to respect how uncertain our efforts are and avoid tying our hands in ways that we will regret in the decade ahead. Well-intentioned legislators too readily replace old credential- and paper-based micromanagement with mandates that rely heavily on still-nascent observational evaluations and student outcome measurements that pose as many questions as answers. The flood of new legislative activity is in many respects welcome, but it does pose a risk that premature solutions and imperfect metrics are being cemented into difficult-to-change statutes.

AEI's *Teacher Quality 2.0* series seeks to reinvigorate our now-familiar conversations about teacher quality by looking at today's reform efforts as constituting initial steps on a long path forward. As we conceptualize it, "Teacher Quality 2.0" starts from the premise that while we have made great improvements in the past 10 years in creating systems and tools that allow us to evaluate, compensate, and deploy educators in smarter ways, we must not let today's "reform" conventions around hiring, evaluation, or pay limit school and system leaders' ability to adapt more promising staffing and school models.

Value-added models of measuring teacher effectiveness have grown in prominence in recent years, and we have seen a flurry of state-level legislative activity to establish these metrics as required components of teacher evaluations. But these systems have clear limitations, not only in their application in traditional settings, but also in the way they presuppose a particular design of the teaching profession that may not apply to alternative settings like online or hybrid schools. In this paper, Michael Hansen, senior researcher at the American Institutes for Research and affiliated researcher with the CALDER Center, reflects on the current state of value-added models and anticipates how they might evolve in the near future. According to Hansen, the shape of evaluation systems will shift as states and districts start to take more responsibility for controlling the quality of the workforce. To keep pace with policy change, the research community must adapt to these changing contexts by taking on new research questions, using different metrics, and collecting new data.

I found Hansen's paper to be insightful and engaging and am hopeful that you will do the same. For further information on the paper, Hansen can be reached at mhansen@air.org. For additional information on the activities of AEI's education policy program, please visit www.aei.org/hess or contact Lauren Aronson at lauren.aronson@aei.org.

—FREDERICK M. HESS
Director of Education Policy Studies
American Enterprise Institute

Executive Summary

The growing prominence of value-added models for measuring teacher effectiveness has prompted a recent surge in policies that consider students' classroom performance part of a teacher's evaluation. Yet, in light of the criticism and limitations of the current models, whether and how evaluation systems will adapt over time is unclear. This paper considers how teacher evaluations may likely evolve in the near future, which will have implications for state and district policy adoption.

The future shape of evaluation systems will be determined by who bears the cost of controlling the quality of the teacher workforce. Until now, teachers and students have largely born these costs. But if states and districts are serious about improving workforce quality, they must take on a greater share. Consequently, the current orientation of input- and output-based evaluations will be supplemented with more rigorous process-based evaluation. Heightened cost pressures for school leadership will likely lead to more automated, data-driven evaluation systems.

Improvements in four specific areas will particularly influence teacher evaluations moving forward:

- Small-scale measurement;
- Implementation issues;
- Workforce monitoring;
- Paradigm shifts in education research.

Data analysis plays a key role across all four areas, and will be the necessary precursor to improvements in public-school-teacher evaluation systems.

AEI TEACHER QUALITY 2.0

Anticipating Innovation in Teacher Evaluation Systems

Lessons for Researchers and Policymakers

By Michael Hansen | March 2013

We are well accustomed to the speed of innovation and change in computers and mobile technology. Simply thumb through the tech pages of your Sunday newspaper, where you are sure to find reviews on the latest product releases. Chances are that the products you see reviewed there now—be they smartphones, e-readers, or tablet computers—bear little resemblance to the products that were featured on the same tech pages just five years before. This ever changing technological landscape is part of buying consumer electronics—we know what we buy today will soon be outdated. And, accordingly, we adjust our behavior to reflect this pace of innovation, particularly weighing the advantages of making a purchase today against the anticipated improvements of waiting for the next product release.

Conversely, we are less accustomed to innovation and change in the realm of education policy. Many of the schools we send our kids to today look almost indistinguishable from the ones we attended in past decades. The timing of policy change is hard to predict—it is stuck in neutral most of the time but periodically comes as a watershed. And, anticipating how the substance of future policy choices will vary from those of today is a difficult endeavor. Consequently, the seemingly mundane decision-making process between taking action today versus delaying in anticipation of future improvements becomes infinitely more complex.

This paper aims to inform this decision-making process for states and districts engaged in making policy decisions that affect the way we evaluate teacher performance. I speculate about the trajectory of innovation in teacher evaluation and where value-added models of teacher effectiveness factor into those changes. Many experts have commented and written about the technical aspects of value-added models, the costs and benefits of these decisions, and about how to adapt them into current evaluation frameworks. I, however, want to investigate how America's teacher evaluation "technology" may likely change in the near future, which will have implications for states' and districts' current and future policy adoption.

Michael Hansen (mhansen@air.org) is a senior researcher at the American Institutes for Research and an affiliated researcher with the CALDER Center.

An Evolving Perspective of Teacher Performance

Research on the role of teachers in learning has undergone a substantial shift in recent years. The old model of education production envisioned schools as factories in which various inputs (teachers, funding, and curricula) are combined and transferred to students through the learning process, resulting in outputs in the form of student achievement and proficiencies in a broad sense. Yet the results apparent in a steady stream of research examining longitudinal data sources and spanning many states and years suggest this view does not match reality.[1] Given the variation in observed outputs (student achievement gains) from different classrooms, researchers such as Daniel Aaronson of the Federal Reserve Bank of Chicago and colleagues have come to the conclusion that teachers themselves are not simply uniform inputs.[2]

Rather, teacher effectiveness varies significantly, both *across* schools, where teachers of similar quality have a tendency to teach in the same school, and *within* schools, where teacher quality fluctuates between classrooms. This variation in teacher effectiveness accounts for more of the differences in observed student outcomes than differences in class size or instructional resources.[3] And, contrary to common belief, most of the variation in teacher quality occurs within schools rather than across schools.[4]

As a result of these and similar findings from new research on teacher quality, the consensus among scholars (recently articulated by economists Douglas Staiger and Jonah Rockoff) has shifted to reflect that the classroom is the real factory, and the school is simply a conglomeration of factories of varying effectiveness.[5] In this view, we should focus less on the whole school and more on the teacher. We might attempt to manage the workforce through teacher inputs (a teacher's training, experience, and credentials), but readily observable teacher characteristics, such as licensure and education, are poor predictors of classroom productivity and are hence ineffective at managing workforce quality.[6] Alternatively, because teachers vary so much from classroom to classroom, students stand to benefit considerably if schools and districts simply focus on classroom output resulting from differing levels of teacher effectiveness. Hence, evaluating teachers' classroom performance and actively managing workforce quality have become focal points of recent proposals to systematically improve public education, as Eric A. Hanushek of the Hoover Institution at Stanford University recommends.[7]

This evolution of research that increasingly focuses on teacher quality rather than school performance is illustrated in the policy shift from the No Child Left Behind Act (NCLB) approach to improving schools to that embodied in the recent Race to the Top (RTT) competitive grant competition. At the risk of oversimplifying, NCLB relies primarily on school-based accountability and competitive pressures to effect systemic change. Publicly reported results were intended to feed market pressures, inducing schools to perform at their highest levels.[8] By contrast, RTT favors states with policies that provide students equal access to effective teachers and improve teacher preparation. RTT also promotes policies that explicitly tie teacher evaluation to classroom performance (including student test scores), and encourages states to develop longitudinal data systems linking students with teachers, which clearly invites the use of teacher value-added models to manage workforce quality.[9] Thus, the emergence of new research findings on teacher quality has apparently influenced policy and will likely continue to do so in years to come.

The new generation of school improvement policies puts the issue of managing teacher quality front and center, but the question of how these policies will evolve over time is uncertain. Historically, teacher evaluation systems have generally failed to discriminate between teachers of different quality. Daniel Weisberg and colleagues' 2009 study shows that instead of differentiating between teachers based on student learning outcomes or qualitative differences in classroom practice, district teacher evaluation systems appear to simply treat the large majority of teachers as equally competent.[10] The primary innovation proposed to remedy these fruitless evaluation practices is the value-added measure, which is intended to estimate the effectiveness of a teacher's classroom performance based on student gains on standardized tests. Over the past two years, 33 states have rushed to update their required teacher evaluation systems, and most of these states have adopted value-added estimates to measure teachers' classroom performance.[11] Whether and which of these updated teacher evaluation systems will successfully remedy the ineptitude of past evaluation systems is currently unclear, but it is clear that even today's most state-of-the-art teacher evaluation systems will be outdated in the not-too-distant future. The current versions of these systems are not final by any means, as the heavy reliance on value-added models is vulnerable and leaves many unanswered questions.

Future evaluation systems will continue to use value-added models, but in spite of their current prominence, these models will probably not be the dominant component of teacher evaluation systems moving forward. The form and function of future teacher evaluation systems will depend on who will bear the cost of controlling the quality of the workforce. Teachers and students have largely borne the costs of quality control up to this point, but if states and districts are serious about improving workforce quality, they must take on a greater share of these costs. Heightened cost pressures across the board will likely lead to more automated, data-driven evaluation systems in the future.

The Vulnerability of Value-Added Measures

The recent school-focused to teacher-focused accountability policy changes have taken the research on value-added models to their limits. As I discussed earlier, the use of these models represents the primary innovation in current evaluation systems, and literature on this topic has been integral in the policy shift from school- to teacher-focused accountability. Not surprisingly, these models and the resulting estimates have come under close scrutiny.[12] Such scrutiny has given rise to literature that critically assesses whether value-added estimates successfully capture meaningful differences in teacher contributions to learning. Are these causal inputs or simply a result of bias?[13] Are value-added differences actually

meaningful for long-term student outcomes?[14] Are they reliable enough to accurately discriminate between teachers' effectiveness in practice?[15] Will they be stable enough over time to affect the overall quality of the teacher workforce?[16] Though researchers' collective understanding of value-added measures has evolved over time, the overall picture demonstrates that these are important metrics that can be used in policy settings with a realistic expectation to affect student outcomes. Now, with research having conditionally endorsed the use of value-added measures, states have adopted them surprisingly quickly, and, in doing so, have moved ahead of the existing research on issues of implementation.

This leapfrog of policy ahead of research brings to the forefront many unanswered questions on value-added models. We would be foolish to believe that current measures alone will fundamentally improve the labor market as a whole; this current condition is vulnerable for three reasons. First, value-added models' reliance on standardized testing complicates the process of scaling these measures across the workforce. Second, the underlying assumptions of current value-added measurements linking students to teachers limit the conditions under which those measures can be used. And third, the costs and benefits of implementing a teacher evaluation system in which value-added measures are a primary component are still unproven.

Value-Added Models' Uncomfortable Reliance on Testing. To estimate a teacher's value-added effectiveness, a state or district needs standardized tests aligned with course content, with students measured before and after exposure to a teacher. Though this approach sounds simple, value-added models' reliance on testing will hinder their potential expansion beyond the tested grades and subjects currently required under accountability systems. The availability of test outcomes is the foremost limiting factor. Most states can currently only generate value-added estimates for teachers of reading or math in grades four through eight, which means over half of the teacher workforce cannot have a value-added estimate produced for them. Testing would have to greatly expand into other grades and subjects to make value-added measures a primary factor in most teachers' evaluations. Yet, in the current environment in which both parents and teachers criticize the emphasis on standardized testing, such a proposition seems unlikely to gain much traction.

Moreover, even if schools were to expand testing to enable broader estimation of value-added models, their applicability outside of currently tested grades and subjects is uncertain. An implicit assumption of proposals to expand testing is that teacher variation is present across all dimensions of the teacher workforce. While teacher variation likely exists across grades and subjects, it has not been empirically well documented, and thus may not be as informative for all teachers. For example, does variation in the effectiveness of social-studies teachers make a meaningful difference in the most important student outcomes? Until we know more about value-added estimates across a broad mix of grades and subjects, these measures will be limited to just one segment of the teacher workforce.

Over the past two years, 33 states have rushed to update their required teacher evaluation systems, and most of these states have adopted value-added estimates to measure teachers' classroom performance.

The quality of tests also limits the usefulness of value-added measures. Value-added researchers commonly quip that the estimates are "only as good as the tests." Standardized tests that are poorly aligned with curricula or that fail to discriminate meaningfully along the full distribution of test takers will result in an artificially low amount of variation in teacher effectiveness. For example, teacher value-added estimates in reading generally convey lower variation between teachers and are less stable over time than value-added estimates in math. Yet preliminary findings from the Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) project suggest that this may be an indicator of low-quality reading and language-arts state tests rather than actual low variation in teacher effectiveness in the workforce.[17] Many states would need to improve their tests before value-added estimates could provide much leverage over workforce quality on poorly measured dimensions.

Assumptions about Student-Teacher Links Limit the Use of Value-Added Measures. The link between students and teachers also limits the widespread use of value-added models. The stylized classroom for which a

teacher's value-added estimate has the most straightforward interpretation is one in which classrooms of students are linked to one teacher for a full year, providing the cleanest relationship between teacher contribution and student learning gains. Unfortunately, actual schooling often does not neatly align with this ideal scenario. The following is a list, though by no means exhaustive, of complicating issues:

- Student mobility (across schools and classrooms) occurs during the course of a year, meaning multiple students are exposed to multiple teachers.
- Students frequently receive more intensive instruction in a subject they are doing poorly in, meaning they essentially receive a second dose of instruction that is unrelated to the primary teacher.
- Spillover from other teachers in the school has been documented in value-added estimates.[18]
- Many classrooms have secondary teachers or teachers' aides.
- Some instructional models fundamentally break the one-teacher-per-classroom mold by exposing students to many adults or integrating virtual learning as a key component (for example, the School of One schools in New York City).

For each of these issues, a student's learning over the course of a school year does not map neatly onto a single teacher. This poses a problem when reconciling the resulting estimates (which represent the collective productivity of all adults responsible for a student) with uses in a teacher evaluation system (which attempt to isolate a specific teacher's contribution).

For all of the research done on value-added estimates, surprisingly little has been conducted outside of the most common one-teacher-per-classroom setting. This is not necessarily an indictment of value-added models; a series of simplifying assumptions could deal with each of the previously mentioned special issues, and the value-added estimates for most teachers will likely be very similar regardless. But there has not been a sufficient level of due diligence on value-added estimates in such scenarios to surmise whether results are robust to these modifications, who will be most directly affected by such decisions, or whether these solutions are politically palatable in a policy setting.

Moreover, it is unclear how valid the resulting estimates are in cases that conform to the baseline model with less-than-perfect fidelity. Though value-added measures may be valid and predictive of future performance for a majority of the workforce in tested grades and subjects, the measures may provide little useful information about the minority of teachers for whom special cases may have undue influence on their value-added estimates. These special cases become amplified in nontraditional schooling models. For example, what do value-added estimates tell us when they relate to teachers in language immersion schools? How important is it to be selective about teacher quality when a large share of instruction is delivered through a computer? It is not obvious whether value-added models in their current form will be useful measures in these nontraditional schools.

Most states can currently only generate value-added estimates for teachers of reading or math in grades four through eight, which means over half of the teacher workforce cannot have a value-added estimate produced for them.

Costs and Benefits of Implementation Are Unproven. Few states or districts have actually used value-added estimates as part of a teacher evaluation system for longer than a year or two; therefore, the actual costs and benefits of implementation are unproven. Three areas in which value-added models are unproven are particularly germane to assessing the overall return on investment.

First, it is unclear how teachers currently in the workforce will respond to the use of value-added measures. While output-based measures may be useful for making summative teacher assessments, they provide little actionable feedback to teachers beyond "excellent" or "needs improvement." Hence, how a given teacher's performance improves as a result of this feedback is not obvious, and the teacher's actual response will likely vary

depending on the consequences attached to value-added performance.[19] In that same vein, incentivizing classroom performance is not a popular proposition among the teacher workforce: according to a 2006 teacher compensation survey conducted in Washington State, only 17 percent of teachers favored merit pay.[20] Whether teachers' attitudes toward value-added measures will thaw over time is uncertain.[21]

Second, a key unknown is how the rise of value-added evaluations may influence the pipeline of incoming teachers to the workforce. Proponents of high-stakes evaluations maintain the untested assumption that more high-quality teachers will be attracted to teaching and will stay in the field if value-added models that distinguish based on quality are present. After all, it is a generally held view that high-quality candidates have been wooed away from the profession by the promise of career advancement and high wages in other fields.[22] Whether the teacher pipeline will actually respond in such a way is unknown. This uncertain pipeline is especially risky in hard-to-staff schools, such as those in disadvantaged and rural districts. Using value-added measures to identify and weed out the weakest teachers is not necessarily helpful when there are no better teachers lining up to fill the vacancies.

Finally, the benefits of combining value-added measures with other teacher evaluation measures are unproven. Until this point, the literature on value-added estimation has overlapped with the literature on teacher pedagogy or classroom observation to a limited extent. Few studies have cross-validated value-added estimates with other performance measures.[23] In theory, a multiple-measures approach (like that investigated in the Gates Foundation's MET project) holds the promise of supplementing value-added estimates by reducing measurement error and providing constructive feedback; whether this promise will be realized in practice remains to be seen.

Relying on current value-added measures as the primary means to drive America's teacher evaluation systems is limiting and unproven in implementation. However, we should neither reject the premise of improving teacher evaluation altogether (using value-added measures is simply one approach to evaluating teachers) nor dismiss value-added measurement as a useful tool. What we should reject is the notion that value-added estimates need to take the central role in America's teacher evaluation systems. If we are serious about quality control in the teacher workforce, we need to think more clearly about designing quality-control mechanisms that serve a variety of functions beyond measuring test-score gains in a classroom.

Quality-Control Mechanisms: Two Models of "Cost"

It is obvious that value-added measurement is at odds with some models of education delivery. This raises a key question about the next generation of teacher evaluation: will the structure of schools drive teacher performance measurement, or vice versa? Ultimately, this question is subsumed by a larger issue that needs to be addressed first: which parties will bear the cost of quality-control efforts in the labor market?

Incentivizing classroom performance is not a popular proposition among the teacher workforce: according to a 2006 teacher compensation survey conducted in Washington State, only 17 percent of teachers favored merit pay.

In an ultimate sense, teacher evaluation systems are a means to control the quality of classroom instruction. As previously described, in recent years, we have witnessed an evolution in the way we approach this sort of "quality control" in education. Under NCLB, we have unsuccessfully relied on costs to the organization as a whole to create incentives to improve quality in the classroom. One could argue, however, that a more appropriate way to situate the quality-control challenge in the teaching context is to incorporate an external regulator to monitor the quality of independent teachers that fall under its management. This means shifting the burden from the NCLB-era whole-school approach to putting the onus on school management to monitor the quality of individual teachers. Within this model, there are two ways to administer the costs of this quality-control process: to teachers or to state and local education agencies.

The Producer-Cost Model. One possibility for monitoring workforce quality is requiring teachers to prove they are competent. Consider the purpose of teacher licensure. Teachers bear the upfront cost of taking college course-

work, completing requisite student teaching hours and passing licensure exams, ostensibly to improve their ability to teach and demonstrate their skill. Teacher licensure provides a good example of a producer-cost model of quality control.

Under this model, regulators (state or local education agencies) demand some minimum quality criteria, while producers (teachers) bear costs to demonstrate that those criteria have been met. Typically, based on the presumption that those who meet the criteria are of sufficient quality, regulators do not closely monitor the actions producers undertake to meet those criteria. In theory, the quality criteria should be aligned with actual desired outcomes, which ensures that producer costs are generally beneficial to the public. The desired outcome of this model is that relatively better teachers would remain in the workforce while imposing only minimal costs on the state or district.

Value-added models assume a stereotypical one-teacher-per-classroom model; however, much of the innovation in schooling—ranging from blended learning models to specialized science, technology, engineering, and mathematics schools—appears to be breaking out of this mold.

The producer-cost model does not always play out in practice. Take the teacher-licensing example: college courses, student teaching, and tests are the state's way of keeping out those who are not up to the challenge of teaching; but these barriers to entry may also unintentionally deter high-quality candidates with other options outside of the classroom. The use of standardized tests is likewise a producer-cost approach to quality control, which induces producers (schools under NCLB or teachers under value-added measurement) to adjust their normal practices to accommodate testing. These accommodations from producers have the potential to both help and hinder student learning, and the state is relatively limited in preventing undesirable responses.[24] Viewed through this producer-cost lens, teacher licensing and the use of value-added estimates are closely related and may very well elicit similar responses from the teacher workforce.

Beyond the direct compliance costs for individual teachers, the producer-cost model also imposes indirect costs onto the school system by implicitly harnessing innovation. As described above, value-added models assume a stereotypical one-teacher-per-classroom model; however, much of the innovation in schooling—ranging from blended learning models to specialized science, technology, engineering, and mathematics schools—appears to be breaking out of this mold. Were we to require these experimental schools to separate into clearly delineated classrooms of students under a single teacher, the schools could potentially lose one of the features that makes them distinct. Whether value-added models' implicit adoption of a status quo perspective of schools is an acceptable cost needs to be explicitly addressed in the public debate as we consider adopting these models.

The Regulator-Cost Model. An alternative way to control quality is to let teachers do their jobs, but to have some type of regular inspections. Schools may use principals as embedded quality monitors. Because a school principal (as an embedded regulator) observes teaching in its natural setting, this model requires minimal accommodation from teachers. As such, this approach is amenable to more granular, automated measurement of teachers' classroom performance. Principals can feasibly gather data on teacher performance and provide feedback, which makes the embedded regulator approach well suited for use in formative assessments. This approach to quality control is the regulator-cost model: school districts (regulators) bear the primary cost of directly monitoring teacher performance.

The regulator-cost model brings quality monitoring directly to the site of production. As long as the teacher adheres to stated "best practices" in the content area, the principal can be satisfied with teacher quality without conditioning performance on output. This is a less invasive way to measure teacher performance in an innovative school model; an embedded regulator can recognize quality in practice even if the specifics of delivery deviate from the norm.

Yet this flexibility can breed liabilities. The model's validity hinges on whether we can reliably identify quality

FIGURE 1
PRODUCER-COST MODEL VS. REGULATOR-COST MODEL

	Producer-Cost Model	Regulator-Cost Model
Object(s) of Measurement	Inputs/Outputs	Process
Direct Costs to School Systems	Small to Moderate	Large
Indirect Costs to Teaching	Large	Small
Resulting Assessment Type	Summative	Formative

Source: The author.

across a broad range of settings, which poses the question: is teacher quality something that can simply be broken down into observable and quantifiable actions on an evaluation rubric? Further, how can regulators maintain the reliability of assessment when the evaluation itself is so substitutable? Hence, a tradeoff is implied. Just as value-added models suffer from a lack of validation outside of the stereotypical classroom model, a similar argument could be made for classroom observations that evolve further from their original rubric. New rubrics could be developed and validated across diverse settings, but will necessarily lag behind the pace of innovation.

The regulator-cost model is widely used, as evidenced in the prevalence of principal-led evaluation. Relying on principals as the sole purveyors of teacher quality, however, is problematic for at least three reasons: (1) inter-rater reliability is low; (2) principals' subjective evaluations may not necessarily consider important outcomes of interest; and (3) implementation fidelity is difficult in schools with high principal turnover. In light of these limitations to using principals as evaluators, some districts have begun hiring external evaluators to fill this role who are trained using validated rubrics and coordinate their efforts to ensure high inter-rater reliability. But in an era of dwindling education budgets, the prospect of deploying a team of external evaluators across schools may be prohibitively expensive.

Using the Two Cost Models to Anticipate Changes. These two approaches to quality control are worth comparing directly (see figure 1).

When laid next to each other, it is evident why districts and schools tend to prefer producer-cost-oriented approaches (including value-added estimates) to maintaining quality in the workforce, while teachers tend to prefer regulator-cost approaches. First and foremost, each party prefers the method that incurs the least cost to itself. Second, basing evaluations on test scores or other outputs is risky for teachers (which explains their preference for process- or input-based evaluation policies), while school management prefers these metrics (inputs or processes are more risky to the school system's objectives). And, finally, the formative results from the regulator-cost model are most valuable for teachers, who can use the information to improve their performance, rather than for schools, which presumably place higher value on summative information to manage workforce quality.

Given the apparently mutually exclusive interests of both teachers and schools in this quality-control problem, it is unclear how the system will evolve in the future. I offer three key points that I believe warrant special note in shaping teacher evaluations moving forward. First, direct costs will drive both research and policy adoption. One could argue that the reason why value-added models (see those in the producer-cost column of figure 1) are largely driving the evaluation push is that they are relatively cheap to estimate across many teachers—low costs encourage an active research field, which we know encourages policy. Without a major shift in the cost of process-based measurement to school systems, value-added estimates and other producer-cost approaches to workforce monitoring will increasingly become the norm. The research community would be foolish, however, to dismiss process-based measures (see those in the regulator-cost column of figure 1) of teacher performance because they are more costly and poorly validated.

Rather, given the potential these measures have to provide more information about what quality teaching looks like, we should make cost cutting a priority as this will encourage greater validation and experimentation. Most of these regulator costs are labor related, so we would be well served to seek opportunities to automate process-based data collection. On this point, one may be tempted to resist quality control altogether, given that

quality control is costly and either teachers or education agencies need to bear such costs. This position is unwise. Until recently, because of a combination of both ineffective and poorly implemented quality-control mechanisms, teachers entering the workforce could be of variable quality and both districts and teachers were bearing little cost to ensure workforce quality.[25] Yet, this does not imply that failure to control quality is a costless venture. Students are the residual claimants of quality control (or lack thereof) in public schools; in other words, those students unfortunate enough to get the low-quality teachers implicitly absorb the costs through lower educational outcomes.

Second, useful and unique performance data comes from both models, but we need to know much more about the differences and commonalities of this information. We need to increase the quality and quantity of research that examines the relationships between various performance metrics and quality-control designs that span both of these categories. The MET project seeks to do exactly this, and a small body of research has begun to emerge on the topic—which is encouraging—but there remains much to be learned in this space. As better, more frequent observational evaluation data (hopefully) emerges in the near future, and is merged with data on student outcomes, we may start to better discern teacher quality on a broad range of fine-grained student outcomes.

Finally, we need a larger public discussion about the prioritization of teacher quality-control mechanisms (that is, who should be bearing these costs), and whether and how these may be constructively combined in practice. Virtually every state or school district now revamping its teacher evaluation system is compelled to use some combination of both producer-cost and regulator-cost mechanisms, yet every system is designed ad hoc. With virtually no research evidence to guide how these disparate mechanisms may be combined and used effectively, it is presently uncertain how much the upgraded evaluation systems will improve upon the old ones. In addition, we must be cautious before jumping in and embracing both quality-control mechanisms with open arms. Such an inclusive approach may wind up demanding too much from teachers, unwittingly chasing the best of them from the classroom.

Returning to the initial question: should our drive to measure teacher productivity define how schools are structured, or should schools define how we measure productivity? My view is that for better or worse, measurement is here to stay and will be a major part of educational evaluation in the future. However, we do not have to (and probably should not want to) give control of how we define the teaching profession over to psychometricians and statisticians. On this question, I say we should let our schools define how productivity is measured.

Moving forward, I expect school systems will use value-added measures in whatever cases they can credibly be used to estimate teacher productivity, and we will have to increase the quality and expanse of process-based monitoring to cover the rest of the teacher workforce that value-added measures cannot reach. To the extent that testing can be reliably and cheaply expanded into other grades and subjects to compute value-added measurements (mostly through the expansion of online testing), it will continue to do so. For subject areas, grades, or schooling models that do not lend themselves to teacher-specific value-added measurement, testing will still play a key role, most notably through the use of school-level, value-added measures. In these cases, subjective performance measures will carry double weight for evaluating teachers' individual contributions.

> More frequent, computerized tests during the school year will feasibly allow for both stronger inferences of teacher effectiveness and timely feedback for teachers to improve practice (or for school leadership to intervene).

Do not suppose, however, that either the way we measure teacher performance or the way we educate American children is set in stone. Value-added measurement is sure to evolve over time, as will standardized tests, data on other student outcomes, and data collection on teachers' practices. Expectations of teachers entering the profession and their roles in the classroom will also become more fluid going forward. The challenge is engendering coordination between the two groups such that education may evolve as it will, yet be guided by performance metrics that allow us to identify the teaching and schooling practices most efficient for the next generation of schoolchildren.

Future Innovations in Teacher Evaluation

Future teacher evaluation systems will look very different from today's systems, which draw heavily from value-added measurements. Three things must happen to improve quality-control design in the future: the direct cost of on-site evaluation must decrease considerably, more actionable feedback must be given to teachers to help them improve practice, and better data must be collected on inputs and outputs for both students and teachers. The following section offers a few ways that plausible developments may alter the face of teacher evaluation in the foreseeable future.

Producer-Cost Mechanisms. Given that producer-cost mechanisms make lower direct costs for states and districts, these mechanisms will be the fallback method of quality control until validated regulator-cost mechanisms become cheaper and can be fully capitalized in districts. I offer three predictions for how input- and output-based metrics might be adjusted in future iterations of teacher evaluation systems.

First, value-added measurement will improve. As states begin to adopt the Common Core State Standards and shift away from minimum-competency testing, tests will become a more reliable measurement tool. Also, testing will migrate away from paper-and-pencil tests and toward computer-based tests, providing more accurate measures of student achievement at a lower per-test cost once in place, which may promote more frequent testing (though likely with fewer consequences attached to any one test). More frequent, computer-based testing will bring with it the ability to capture value-added more reliably and the potential to provide useful feedback to teachers in a timely manner to improve their practice; hence, the lines between summative and formative assessment will slowly begin to blur in how standardized tests are used.[26]

Second, input-based teacher measures will improve. Though to date, the empirical research on teacher quality has established little correlation between observable teacher input measures and classroom productivity, it is premature to conclude that these measures do not matter at all. A growing body of literature has begun to point to differences in teacher preparation before entrance into the teacher workforce.[27] Teacher selection on the front end is a key piece of Teach For America's recruitment model, and recent research from Harvard University's Will Dobbie suggests that Teach For America's model discriminates between variables that are differentially effective in the classroom.[28] By improving these measures and getting a better sense of how some of these inputs may be related to educational outcomes of interest, state policies for licensure and district policies for teacher hiring may be adjusted in ways to promote a flow of high-quality incoming teachers.

Third, data on a broad range of student outcomes will improve. Teacher resistance to outcome-based performance measures will decrease when test-score gains are not the only outcome that is monitored. Such measures could be easily calculated if districts began making some already-collected data available for research; for instance, data on student attendance, transcripts, or behavioral discipline could enhance our ability to monitor both students and teachers. Occasional surveys, perhaps administered with testing, might be able to gather information on students' college and career objectives, attitudes toward learning, or engagement in school. As these measures are developed and validated, we can generate outcome-based measures for larger segments of the teacher workforce without having to expand standardized testing. Because these are also outcomes of schooling that we care about, future evaluation systems can make determinations about teacher performance based on a wide range of fine-grained student outcomes.

Regulator-Cost Mechanisms. Because of the sheer number of teachers that typical value-added models cannot take into account, alternative methods to assess teacher quality will be pursued. In the immediate future, states and districts will probably focus on implementing current observation-based and principal-based assessments with more fidelity to assuage the demand for greater control over teacher quality. In the slightly more distant future, these models will adapt in the three following ways.

First, student or parent evaluations will become more frequent and will be used in evaluation systems. These measures attempt to use students' experiences to investigate the quality of the production process. Soliciting "customer feedback" has intuitive appeal in the interest of providing formative information on teachers' classroom practices. Yet, such evaluations may be unpopular among teachers in some settings, and are problematic in the case of philosophical differences between students and teachers or parents and teachers, so they will likely serve a primarily formative purpose. The preliminary evidence to date suggests that these evaluations are correlated with value-added outcomes in teachers.[29] Yet, further research is required—how many surveys are needed for reliable inference? Can interim surveys provide reliable feedback that might allow

teachers to adapt practice? As we learn more about student surveys and fine tune the information collected, the student or parent evaluations will become an important piece of a teacher's performance portfolio.

Second, measures of teachers' professional conduct will improve. Schools have the ability to track useful information on teacher behavior that may inform our assessments of their productivity, but it is unclear whether districts currently use this information in any meaningful way. For instance, absences, tardiness, professional conduct, and peer relationships probably influence student outcomes and likely impose costs on the district; therefore, using them for evaluative purposes is a reasonable proposition.[30] Are specific teachers taking on greater responsibilities to cover for a colleague's chronic absence? What about senior teachers who mentor junior colleagues? Recording peer interactions could not only help inform researchers on effective human resource management within schools, but also improve value-added estimates by more accurately accounting for peer influences. As more of these measures are developed, collected, and validated, one should expect them to be adopted into a future generation of teacher evaluation, either as formative or summative assessments.

Third, classroom observations will become more automated in response to cost pressures, and will become more frequent over time. Currently, classroom observations are subjective measures conducted by a chosen evaluator; however, some elements of observational rubrics could feasibly become quantified and objectively measured using audio or video recordings, reducing the cost of data gathering on some dimensions. Facial recognition software could be used to decipher hints of student engagement or cognition in learning based on videos taken during class time. Automated observation may not supplant in-person observations entirely, but the possible cost efficiencies of automation would likely help it play a major role in future evaluations, particularly in the cases where value-added measures cannot be estimated. Though advances such as these would need to be developed from the ground up, and therefore may be further off on the horizon, such technological innovations will certainly change the way teachers work on the job and how their performance is evaluated.

Innovations in Research Are Necessary

We cannot naively assume the research that brought us value-added models will suffice for teacher evaluation systems of the future. Instead, we must actively modify our research in anticipation of new questions. Researchers can work toward achieving this next generation of evaluation by actively pursuing the following four things.

Ways to Measure on Smaller Scales. Many teaching measures span long time periods or many concepts, and are thus too coarse to help managers staff schools or teachers improve their practice. Consider the length of time used to obtain value-added measurements—the shortest measurements reflect a teacher's productivity over the course of a school year, and many studies suggest several years of data are necessary to make reliable decisions. More frequent, computerized tests during the school year will feasibly allow for both stronger inferences of teacher effectiveness and timely feedback for teachers to improve practice (or for school leadership to intervene). In addition, with the right information from test developers, value-added models could be reduced to the concept level, demonstrating which areas of instruction need to be reviewed for students. If schools were informed about which teachers were most effective at which concepts, schools could adjust their staffing, perhaps through supplemental instruction or teacher rotations, to better meet students' needs.

Classroom observations already measure small-scale behaviors; however, much can be gained from further micro-measurement. To begin, we need to actively determine how to translate established observation rubrics into automated algorithms that can be coded with minimal human involvement; doing so will considerably drive down the cost of providing formative feedback. Moreover, pairing classroom observation data with finer outcome measurements can potentially multiply the efficacy of both data sources by linking learning with specific behaviors. If we knew which classroom practices yielded more learning versus less learning on particular concepts, we would be foolish to not train teachers to adopt their instruction accordingly. For example, we might eventually be able to develop skill profiles for teachers to describe which of their colleagues are most effective with low-achieving students or which ones better promote critical thinking. Schools could then staff accordingly.

Ways to Effectively Implement Evaluation Systems. As described above, teacher evaluation policy has quickly moved ahead of researchers in grappling with issues of implementing value-added measures as a piece of a coherent, effective evaluation system. Researchers need to focus on the sticky points of implementation that are bound to

arise as a result. For instance, the statistical and predictive properties of value-added estimation are reasonably well documented, but those of other performance metrics are not; therefore, how well evaluation systems that combine these measures discriminate teacher quality is not readily clear and needs to be better understood. We also need to better understand the tradeoff between the reliability of observational measures and the frequency of observation. Given the expense involved, districts may decide that two external observations in a year for some teachers is preferable to five observations with slightly higher reliability.

We must further investigate which evaluation measures can best promote positive student outcomes in which settings. For instance, while value-added measures may be informative in high-needs schools, a teacher's contribution to other student outcomes (for example, staying on track in coursework or graduating) may take on greater weight in evaluating teacher effectiveness in these settings. Tailoring evaluation systems also presents the possibility of allowing some teachers or schools to undergo less intensive evaluation if they have positive track records.

We ultimately rely on district and school leadership to use evaluation information in their assessment of teacher quality, regardless of which mechanisms are used to monitor teacher performance in the classroom. Accordingly, developing metrics to monitor the fidelity of principals' implementation (for districts or school boards) or districts' actions (for states) may be used to hold leaders responsible for their actions (or lack thereof) in attempting to control the quality of the labor force. We should probably not hold teachers accountable unless we are willing to hold school leadership accountable as well. We need to think of creative ways to monitor the various entities involved in delivering public education.

Ways to Monitor the Health of the Workforce. Researchers in general know relatively little about the actual productivity of the teacher workforce. If we are serious about manipulating its quality, however, it behooves us to understand what the level of productivity is and track it over time. While state administrative data has been used to investigate teacher productivity across an entire state, we do not know how teachers in, for example, North Carolina compare with their counterparts across the border in Virginia. Differences in the qualifications and productivity of the teacher labor force may certainly be explanatory factors in why some states perform consistently better on the National Assessment of Educational Progress, but we do not currently have the data assembled to say whether this is true. Though the prospect of assembling student-teacher linked data for the whole country is certainly a pipe dream, bringing together a few modest data sources might give us new and credible information. For instance, gathering data from teacher training institutions and licensure testing companies could be a good start in monitoring the teacher pipeline across the country. Integrating administrative state data (when available) would enable researchers to monitor how these teachers move through their careers.

Additionally, most of the research studies on teacher quality have a macro-oriented view of the workforce, looking at overall variation across a state or district. This information is likely not localized enough to be of much use to principals, who are the primary agents capable of affecting the composition of the workforce through hiring and firing decisions. For instance, we cannot tell a principal how his or her school's applicant pool compares against the school's stock of teachers, even though his or her optimal staffing strategy will vary based on that information. Moreover, this information can help districts—looking across multiple schools—recognize where more or less help is needed in staffing. We need to develop ways to keep principals better informed about their own teachers and prospective teachers if we want them to be successful in manipulating the quality of the workforce.

Finally, we need to start tracking workforce quality over time. Teacher value-added estimates are almost universally estimated within a fixed time period relative to teachers in that period; comparisons across periods are not feasible. If we wish to improve workforce quality, we need the means to track progress over time, which means we need to develop credible measures of teacher productivity that are on an absolute, rather than relative, scale. This is certainly easier said than done, but is a necessary part of tracking progress and informing policymakers.

Strategies That Break the Research Mold. The research that will improve teacher evaluation systems and public education overall will almost certainly require that we go against some research norms. First, groundbreaking research in this area will become more interdisciplinary. To cross-validate classroom practice with conceptual learning and value-added models, we need to have some interactions between the various camps where such interactions are discouragingly infrequent. Particularly in the interest of automating many of these evaluation processes, we need to involve researchers from the education technology field. Further, with all of the data collected on classroom practices, concept-level mastery, and teacher

inputs, we will soon find ourselves flooded with data. When that day comes, we should resist standing on principle and only looking for relationships in the data where theory points us; rather, we should embrace data mining and other empirical strategies that can search for patterns in the data that perhaps challenge our assumptions about learning and evaluation.

Finally, with all of this research potentially emerging in the nexus between pedagogy, labor economics, testing, and human resource management, we should also take actions to reduce the time horizons for publishing research findings. Though peer review and requests for revisions do serve as a barrier to low-quality research coming to dominate the field, these barriers also make for slow progress where research findings commonly take years to mature. We should seek ways to reduce timelines for publishing valuable research without compromising quality so that the research community at large may incorporate these findings and adapt accordingly.

Conclusion

Teacher evaluation systems must improve for the current policy interest in teacher accountability to deliver on its promise of fundamentally improving student outcomes. As a result, the teacher evaluation systems of the not-too-distant future will look quite different from today's systems, which rely heavily on value-added models. In the face of this continued innovation, I encourage states and districts to anticipate and prepare for these changes so that as improvements are made, teacher evaluation systems can absorb the newest updates with the least amount of pain. In an earlier paper in this series, Rotherham and colleagues offered some practical guidance on how states and districts may avoid traps that might inhibit the evolution of evaluation systems over time; among other things, they suggest avoiding overly prescriptive policies—particularly in adopting laws, as these inherently stifle the system's ability to adapt to future changes.[31]

In this piece, I explore how teacher evaluation systems may likely evolve in the future, and what research is necessary to enable this process. I view quality control over the teacher labor market as a regulator overseeing industry producers, and consider how the structure of evaluation systems is determined by the party that bears the cost of controlling quality. Following this argument, the prominent teacher value-added models of today, while holding many desirable properties, ultimately provide too little information for too few teachers to use as the central piece of teacher evaluation. However, value-added methods will likely evolve considerably in the coming years, and will play an important role in validating classroom practice, teacher behaviors, and student engagement. Ultimately, I encourage researchers to anticipate the eminent advances in teacher evaluation by focusing on measuring on smaller scales, implementing evaluation systems with fidelity, monitoring the health of the teacher workforce, and breaking the mold of education research.

Furthermore, data will be the *sine qua non* in future iterations of teacher evaluation. A common theme for all of the various ways in which evaluation systems will evolve in the future is data collection. The direct costs of teacher monitoring are too high for it to be broadly implemented while relying on human labor, and lowering these costs will be a major factor in teacher evaluation moving forward. Data collection must become more automated, and data use in decision making must become more widespread if quality control in the workforce is to be pursued in earnest. Investing now in the capacity to accommodate a data-driven evaluation system will be money well spent if the trajectory of interest in teacher evaluation plays out as I envision it.

Finally, it is also feasible that quality-control mechanisms in the future will not be exclusively teacher based, as the current emphasis suggests. Within the context of our current interest in trying to measure and manage teacher quality, we must be cautious that we are not so blinded by ambition as to ignore the role of schools in the productive process. Emerging empirical evidence suggests that principal leadership appears to have large and statistically significant effects on student learning; effective schools hire and develop teachers differently than do less effective schools; the most effective schools are those where a coherent learning and work environment thrives.[32] Indeed, if we knew more about how to improve schools and replicate improvement efforts on a large scale, we might be able to indirectly manipulate and manage the quality of the workforce better than direct efforts to do so.

Notes

1. For a recent review of this literature, see Eric A. Hanushek and Steven G. Rivkin, "Generalizations about Using Value-Added Measures of Teacher Quality," *American Economic Review* 100, no. 2 (2010): 267–71.

2. Daniel Aaronson, Lisa Barrow, and William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25 (2007): 95–135

3. Barbara Nye, Spyros Konstantopoulos, and Larry V. Hedges, "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26, no. 3 (2004): 237–57; Douglas O. Staiger and Jonah E. Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives* 24, no. 3 (2010): 97–118; and Eric A. Hanushek, John F. Kain, and Steven G. Rivkin. "Why Public Schools Lose Teachers," *Journal of Human Resources* 39 (2004): 326–54.

4. Eric A. Hanushek et al., "The Market for Teacher Quality" (working paper, National Bureau of Economic Research, Cambridge, MA, 2005).

5. Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information."

6. Aaronson, Barrow, and Sander, "Teachers and Student Achievement in the Chicago Public High Schools;" and Cory Koedel and Julian R. Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function" (working paper, University of Missouri, 2007).

7. Eric A. Hanushek, "Teacher Deselection," in *Creating a New Teaching Profession*, eds. Dan Goldhaber and Jane Hannaway (Washington, DC: Urban Institute Press, 2009).

8. See, for instance, Robert L. Linn, Eva L. Baker, and Damian W. Betebenner, "Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001," *Educational Researcher* 31, no. 6 (2002): 3–16; and Eric A. Hanushek and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24, no. 2 (2005): 297–327.

9. US Department of Education, *Race to the Top Program Executive Summary* (Washington, DC, 2009).

10. Daniel Weisberg et al., *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (New York, NY: The New Teacher Project, 2009), http://widgeteffect.org/.

11. National Council on Teacher Quality, *State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies* (Washington, DC: National Council on Teacher Quality, 2011).

12. For example, see Richard Rothstein et al., *Problems with the Use of Student Test Scores to Evaluate Teachers* (Washington, DC: The Economic Policy Institute, 2010).

13. Dan Goldhaber and Duncan Chaplin, "Assessing the 'Rothstein Test': Does it Really Show Teacher Value-Added Models are Biased?" (working paper, CALDER, 2012); Koedel and Betts, "Re-Examining the Role . . . "; and Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125 (2010):175–214.

14. Raj Chetty, John Friedman, and Jonah Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood" (working paper, National Bureau of Economic Research, December 2011).

15. Peter Z. Schochet and Hanley S. Chiang, *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains,* (Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, 2010).

16. Dan D. Goldhaber and Michael Hansen, "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance" (forthcoming in *Economica*); and Daniel F. McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4 (2009): 572–606.

17. The Bill & Melinda Gates Foundation, *Initial Findings from the Measures of Effective Teaching Project* (2010). Table 5 of the Gates Foundation's report and its accompanying text discusses the estimated variation of teacher effectiveness from various standardized reading and language-arts tests.

18. C. Kirabo Jackson and Elias Bruegmann, "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics* 1 (2009):85–108.

19. Matthew G. Springer et al., "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching" (working paper, National Center on Performance Incentives, Vanderbilt University, Nashville, TN, 2010).

20. Dan Goldhaber, Michael DeArmond, and Scott DeBurgomaster, "Teacher Attitudes about Compensation Reform: Implications for Reform Implementation," *Industrial & Labor Relations Review* 64, no. 3 (2011).

21. A frequently cited source of teacher resistance stems from viewing value-added estimates through the lens of the socioeconomic status of their students, even though most models explicitly control for these variables. Though researchers tend to scoff at such comments, teachers' concerns must be addressed for them to perceive the system as beneficial. The new IMPACT evaluation system in DC Public Schools, which heavily incorporates value-added estimates in tested grades and subjects, resulted in a disproportionate amount of effective teachers in the most affluent parts of the district and a disproportionate amount of low-performing teachers in disadvantaged schools. Though there is likely positive sorting between teachers and schools (such that higher value-added teachers teach in more affluent school settings), it casts doubt on whether the playing field is tipped against those in disadvantaged schools. See Bill Turque, "Educators Say It Will Take More Than Dollars to Lure Effective Teachers to Struggling D.C. Schools," *Washington Post*, January 23, 2012.

22. Sean P. Corcoran, William N. Evans, and Robert M. Schwab, "Women, the Labor Market, and the Declining Relative Quality of Teachers," *Journal of Policy Analysis and*

Management 23: 449–70; and Caroline M. Hoxby and Andrew Leigh, "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States," *American Economic Review* 94: 236–40.

23.The Bill & Melinda Gates Foundation, *Initial Findings from the Measures of Effective Teaching Project*; Brian A. Jacob and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics* 26 (2008):101–36; and Jonah E. Rockoff and Cecilia Speroni, "Subjective and Objective Evaluations of Teacher Effectiveness," *American Economic Review* 100 (2010): 261–66.

24. Responses to standardized testing could be beneficial for students, so long as the tests are aligned with valued outcomes. See Edward P. Lazear, "Speeding, Terrorism, and Teaching to the Test," *Quarterly Journal of Econ*omics 121 (2006): 1029–61. Yet, many teacher- and school-level responses may be to students' detriment. See David Figlio and Joshua Winicki, "Food for Thought: The Effects of School Accountability on School Nutrition," *Journal of Public Economics* 89, no. 2–3 (2005): 381–94; Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118 (2003): 843–77; and Daniel Koretz, "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," in eds. Eric Hanushek, James Heckman, and Derek Neal, *The Journal of Human Resources* 37, no. 4 (2002): 752–77.

25. Currently, the evidence suggests teacher training and licensure is only weakly correlated with differences in teacher productivity, and public-school districts have ineffective quality-control mechanisms in place. Daniel Weisberg et al., *The Widget Effect . . . ";* Dan Goldhaber, "Everyone's Doing It, But What does Teacher Testing Tell Us About Teacher Effectiveness?" *Journal of Human Resources* 42, no. 4 (2007): 765–94.

26. Pamela Grossman et al., "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores," (working paper, CALDER, 2010).

27. Dan Goldhaber and Stephanie Liddle, "The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement," (working paper, CALDER, 2012); and Kata Mihaly et al., "Where You Come From or Where You Go? Distinguishing between School Quality and the Effectiveness of Teacher Preparation Program Graduates," (working paper, CALDER, 2012).

28. Will Dobbie, "Teacher Characteristics and Student Achievement: Evidence from Teach For America," (working paper, Harvard University, Cambridge, MA, 2011).

29. The Bill & Melinda Gates Foundation, *Initial Findings from the Measures of Effective Teaching Project*.

30. Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "Are Teacher Absences Worth Worrying about in the United States?" *Education Finance and Policy* 4 (2009):115–49; Kirabo Jackson and Bruegmann, "Teaching Students and Teaching Each Other . . . ," 85–108; and Reagan T. Miller, Richard J. Murnane, and John B. Willett, "Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District," *Educational Evaluation and Policy Analysis* 30:181–200.

31. Andrew J. Rotherham, Sara Mead, Rachael Brown, "The Hangover: Thinking about the Unintended Consequences of the Nation's Teacher Evaluation Binge," (Washington, DC: AEI, 2012).

32. Gregory F. Branch, Eric A. Hanushek, and Steven G. Rivkin, "Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals," (working paper, National Bureau of Economic Research, 2012); Anthony S. Bryk et al., *Organizing Schools for Improvement: Lessons from Chicago* (Chicago, IL: University of Chicago Press. 2010); C. Kirabo Jackson, "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers," (working paper, National Bureau of Economic Research, 2010); and Susanna Loeb, Demetra Kalogrides, and Tara Beteille, "Effective Schools: Teacher Hiring, Assignment, Development, and Retention," (working paper, National Bureau of Economic Research, 2011).