TEACHER QUALITY 2.0

AMERICAN ENTERPRISE INSTITUTE

SPECIAL REPORT 2

# The Hangover

## Thinking about the Unintended Consequences of the Nation's Teacher Evaluation Binge

Sara Mead, Andrew Rotherham, Rachael Brown | September 2012

## 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As we start to rethink outdated tenure, evaluation, and pay systems, we must take care to respect how uncertain our efforts are and avoid tying our hands in ways that we will regret in the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've made great improvements in the past ten 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 about hiring, evaluation, or pay limit school and system leaders' ability to adapt more promising staffing and school models.

In this second installment of the series, Sara Mead, Andrew Rotherham, and Rachael Brown of Bellwether Education Partners draw out the key tensions and trade-offs associated with our sprint to legislate and build educator evaluation systems. While we can take pride in the progress we have made, the authors call for some humility in thinking about the practical limits of such statutes and processes, and remind us that such rigid structures might unintentionally stifle future innovation. They explain: "the recent evaluation binge is not without risks…headlong rushes inevitably produce unintended consequences—something akin to a policy hangover as ideas move from conception to implementation." Mead, Rotherham, and Brown identify several recommendations as to how to approach the challenges that lie ahead as we move into the next generation of teacher quality reform.

I found the piece to be a useful and engaging contribution to this series, and am hopeful that you will do the same. For further information on the paper, Mead can be reached at sara@bellwethereducation.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

Over the past three years, more than twenty US states have passed legislation establishing new teacher evaluation requirements and systems, and even more have committed to do so in Race to the Top or Elementary and Secondary Education Act Flexibility Waiver applications. These new evaluation systems have real potential to foster a more performance-oriented public education culture that gives teachers meaningful feedback about the quality and impact of their work. But there are pitfalls in states' rush to legislate new systems, and there are real tensions and trade-offs in their design.

Unfortunately, much of the current policy debate has been framed in stark ideological terms that leave little room for adult discussion of these tensions. This paper seeks to move the debate beyond ideology and technical issues by highlighting four key tensions that policymakers, advocates, and educators must consider in the development of new teacher evaluations:

- **Flexibility versus control:** There is a temptation to prescribe and legislate details of evaluations to ensure rigor and prevent evaluations from being watered down in implementation. But overly prescriptive policies may also limit school autonomy and stifle innovation that could lead to the development of better evaluations.

- **Evaluation in an evolving system:** Poorly designed evaluation requirements could pose an obstacle to blended learning and other innovative models in which it is difficult or impossible to attribute student learning gains in a particular subject to a particular teacher.

- **Purposes of evaluations:** New evaluation systems have been sold as a way both to identify and dismiss underperforming teachers and to provide all teachers with useful feedback to help them improve their performance. But there are strong tensions between these purposes that create trade-offs in evaluation system design.

- **Evaluating teachers as professionals:** Advocates argue that holding teachers responsible for their performance will bring teaching more in line with norms in other fields, but most professional fields rely on a combination of data and managerial judgment when making evaluation and personnel decisions, and subsequently hold managers accountable for those decisions, rather than trying to eliminate subjective judgments as some new teacher evaluation systems seek to do.

Recognizing these tensions and trade-offs, this paper offers several policy recommendations:

- Be clear about the problems new evaluation systems are intended to solve.

- Do not mistake processes and systems as substitutes for cultural change.

- Look at the entire education ecosystem, including broader labor-market impacts, pre- and in-service preparation, standards and assessments, charter schools, and growth of early childhood education and innovative school models.

- Focus on improvement, not just deselection.

- Encourage and respect innovation.

- Think carefully about waivers versus umbrellas.

- Do not expect legislation to do regulation's job.

- Create innovation zones for pilots—and fund them.

# The Hangover

## Thinking about the Unintended Consequences of the Nation's Teacher Evaluation Binge

Sara Mead, Andrew Rotherham, Rachael Brown

Teacher evaluation is hot these days. In the past two years, more than twenty US states have passed legislation changing teacher evaluation systems to include evidence of teachers' impact on student achievement. This fevered pace of legislation was sparked by the federal Race to the Top (RTT) program, which called on states to "Design and implement rigorous, transparent, and fair evaluation systems for teachers and principals that differentiate effectiveness." And the US Department of Education's Elementary and Secondary Education Act (ESEA) flexibility waiver process built momentum by demanding such policies as a condition for a waiver from No Child Left Behind (NCLB) requirements. National and local media have added to the clamor; newspapers in New York and Los Angeles even published individual teacher value-added ratings as major stories.

After years of policies that ignored differences in teacher effectiveness, the pendulum is swinging in the other direction. By and large, this is progress—research shows that teachers affect student achievement more than any other within-school factor. Decades of inattention to teacher performance have been detrimental to students, teachers, and the credibility of the teaching profession. Addressing this problem is critical to improving public education outcomes and raising the status of teaching, and neither the issues raised in this paper nor technical concerns about the design and mechanisms of evaluation systems should be viewed as a reason not to move toward a more performance-oriented public education culture that gives teachers meaningful feedback about the quality and impact of their work.

Sara Mead (sara@bellwethereducation.org) is principal at Bellwether Education Partners. Andrew Rotherham (andy@bellwethereducation.org) is cofounder of and partner at Bellwether Education Partners. Rachael Brown (rachael@bellwethereducation.org) is an associate at Bellwether Education Partners.

Yet the recent evaluation binge is not without risks. By nature, education policymaking tends to lurch from inattention to overreach. When a political moment appears, policymakers and advocates rush to take advantage as quickly as they can, knowing that opportunities for real change are fleeting. This is understandable, and arguably necessary, given the nature of America's political system. But headlong rushes inevitably produce unintended consequences—something akin to a policy hangover as ideas move from conception to implementation.

Welcome to teacher evaluation's morning after. As states move from paying no attention to evaluation to starting to design and implement ambitious statewide evaluation systems, it is clear that many are struggling with technical and political challenges. States are figuring out how to systematically evaluate personnel for whom evaluations have been cursory for years. They are incorporating value-added measures of student learning into evaluations for teachers in tested grades and subjects. They are grappling with the even greater challenge of how to measure impact on student learning for the majority of teachers who teach nontested grades and subjects. They are also coping with insufficient managerial capacity. And, in the process, many are running up against the limits of the carefully constructed systems and design features that well-meaning policy wonks told them were critical to effective teacher evaluation systems.

And we are not even beginning to see the greatest of these challenges. The current range of teacher evaluation policies is designed with an eye toward the US education system as it currently exists—even as technological innovation, blended learning, and the growth of charter and portfolio models in many urban areas are fundamentally changing the way the system works. If we are not careful, new teacher evaluations will become another Ice Nine–like element in education, freezing in place what they touch and ultimately becoming just as much of an impediment to progress as the old, inadequate systems they have displaced.

That issue is the focus of this paper. We trace the evolution of teacher evaluation policy, including the very real problems with the previous status quo that created the current situation. We then turn to the question of what costs today's focus on teacher evaluation may carry for innovation, for innovative schools, and for efforts within the public education system to do things differently.

We believe that the current attention to teacher evaluation is long overdue and that as a nation America has yet to wrestle honestly with the issue of teacher quality. Yet one can acknowledge that and still worry about what is happening today. And we do.

So how did we get here?

## Historical Context

Current efforts to establish teacher evaluation systems must be understood in historical context as the most recent in a series of reforms—stretching back to the origins of public schooling— that sought to improve the quality of teaching in public schools. These historical efforts have emphasized different indicators of teacher quality and mechanisms for improvement. But research suggests that many of these mechanisms and indicators in fact have little relationship to improved student learning. Proponents of new teacher evaluation systems have seized on this strategy, not because they "hate" teachers as some suggest, but because the existing evidence suggests that how teachers actually perform in the classroom is a far better indicator of quality than the proxy indicators—certification, years of experience, postgraduate credentials—on which our educational system currently relies.

For over a century, efforts to improve the quality of teaching have largely focused on a "professionalism" agenda, seeking to improve quality through increasing state regulation and formal training requirements for teacher certification and licensure.[1] Beginning in the 1980s, standards-based reformers also called for greater rigor in teacher preparation programs, which reformers argued placed too much emphasis on pedagogical theories and too little on ensuring teachers had deep content knowledge in the subjects they taught. Massachusetts, for example, enacted reforms that required all teachers to hold a major in a subject area other than education and to pass rigorous licensure exams of communication and literacy skills and academic content knowledge.[2]

Building on this concern, NCLB also emphasizes teachers' subject matter content knowledge. The law's "highly qualified teacher" (HQT) provisions require all teachers to hold a bachelor's degree and state licensure and to demonstrate knowledge of the subject they teach, through either a college major, passage of a certification exam, or, for veteran teachers, by meeting a state-defined "highly objective, uniform state standard of evaluation" (HOUSSE). These provisions were designed to ensure that teachers have subject matter knowledge specific to the subjects they teach, reduce the rate of "out of field" teaching in middle and secondary school, and improve equity in the distribution of qualified teachers for poor and minority students. But while the HQT provisions were designed with the best of intentions, they ultimately fell short, creating paperwork hoops for teachers and schools to jump through without necessarily improving the quality of instruction. Because the definition of a "highly qualified teacher" relies almost entirely on a teacher's subject matter knowledge, there is no guarantee that teachers who meet the HQT standard are effective in improving student achievement.[3] NCLB's provisions requiring states to ensure low-income and minority students' equitable access to quality teachers also rely on the HQT standard, so they do not ensure that these students are taught by effective teachers.

Content knowledge alone does not ensure teacher effectiveness.

Unfortunately, research suggests that many of the indicators policymakers have historically relied on as measures of teacher quality are at best very weak predictors of teachers' effectiveness in improving student learning. Indeed, even in studies that account for the range of characteristics commonly perceived as external indicators of teacher quality, these characteristics collectively account for only a small percentage of the observed variation in teacher effectiveness, as measured by impact on student learning.[4] Research shows that holding a master's degree—a proxy for quality that most teacher compensation systems reward—has no positive correlation with improved student learning, with the exception of secondary math and science teachers who hold master's degrees in those subjects.[5] As this finding suggests, indicators of subject knowledge are correlated with improved effectiveness for some teachers, but content knowledge alone does not ensure teacher effectiveness. Research also shows that while experience does matter in teacher quality, the majority of teacher improvement comes in the first few

years of teaching, and returns diminish beyond that point.[6] Similarly, a wide body of empirical research finds little relationship between a teacher's licensure credentials and certification and her impact on student performance; the variation in effectiveness among teachers from the same preparation pathway is much greater than the difference between pathways.[7] Given the lack of evidence that indicates that many commonly viewed indicators of quality actually correlate with improved performance, it is not surprising that the last century of teacher quality efforts have often proved disappointing.

But even as policymakers and advocates for low-income and minority students have grown disillusioned with NCLB's HQT provisions and other policies that rely on teacher credentials, another feature of NCLB—its annual testing requirements in grades three through eight—is generating abundant student achievement data that are transforming the national debate on teacher quality. Data systems in a growing number of states are linking student achievement data to teachers, making it possible to calculate individual teachers' impact on student learning.

**The Growth of Value-Added Measures.** Value-added measures of teacher effectiveness have been used to some extent for many years. In the early 1980s, two statisticians at the University of Tennessee—William Sanders and Robert McLean—began experimenting with statistical methodologies to mitigate some of the challenges of using student achievement data to assess teacher and school effectiveness. Working with data from Tennessee school districts, they found evidence that there are significant measurable differences in schools' and teachers' impacts on student learning and that estimates of school and teacher effectiveness tend to be consistent from year to year. Value-added research also showed that these differences could have significant implications for student learning, potentially large enough to meaningfully narrow or widen achievement gaps.[8]

In 1991, the Tennessee legislature passed the Education Improvement Act, which created a new statewide school accountability system, a major component of which was the Tennessee Value-Added Assessment System, based on Sanders's and McLean's work. Although this value-added model had clear advantages for improving both school and teacher accountability, it remained largely under the radar outside of Tennessee and a small circle of education policy wonks for several years, in part because many states lacked both the annual assessments and robust data-tracking systems needed to replicate the Tennessee model.[9]

This changed in 2002 with the passage of NCLB, which focused national attention on educational accountability, requiring states to annually assess all students in grades three through eight and producing an unprecedented wealth of student achievement data. The law also required states to identify schools and districts not meeting "adequate yearly progress" (AYP)—a measure of a school's student achievement. Discontent with AYP—an admittedly crude measure that calculated annual school performance but not student progress—led some policy analysts to look to Tennessee's value-added model as a potentially better way to measure school performance. At the same time, frustrations with the limitations of NCLB's HQT provisions sparked interest in value-added measures as a means to ensure equitable access to effective teaching for low-income and minority students.

Data systems in a growing number of states are linking student achievement data to teachers, making it possible to calculate individual teachers' impact on student learning.

In a 2004 paper published by the Education Trust, policy analyst Kevin Carey made the case for using value-added data to evaluate teacher effectiveness and offered a set of recommendations for moving toward value-added as a measure of teacher quality. Specifically, Carey recommended that policymakers develop and support systems to collect and analyze value-added data; require evidence of student learning as part of teacher preparation and licensure; and use value-added data to inform teacher recruitment, hiring, compensation, performance evaluation, and professional development. "When used wisely," Carey wrote, value-added data "provides a strong basis for actions that will help states, districts, and schools improve teacher quality, raise overall student achievement, and close the achievement gap."[10]

But the use of value-added data for teacher evaluation really seized attention following a provocative 2006 Brookings Institution report by Robert Gordon, Thomas Kane, and Douglas Staiger. Given the paucity of evidence

that traditional entry credentials—such as certification—are predictive of teachers' effectiveness in the classroom, the authors argued that policymakers would do better to reduce these barriers to entry into the profession. Instead, policymakers should seek to improve quality by focusing on value-added measures of teachers' performance once in the classroom, rewarding teachers who improve student performance and dismissing those who do not. Specifically, the paper's most controversial recommendation called for schools to dismiss the lowest-performing 25 percent of teachers after two years of teaching. By reducing barriers to entry, creating incentives for high performers, and annually dismissing the lowest-performing 25 percent of teachers, the authors argued, policymakers could rapidly increase the level of effectiveness in the teaching profession and generate improved student achievement.[11]

These proposals were particularly provocative when contrasted with prevailing practices in teacher evaluation. In a seminal 2009 report entitled *The Widget Effect*, the New Teacher Project (TNTP) detailed the failure of most existing teacher evaluation systems, in which less than 1 percent of teachers received unsatisfactory ratings. The name of the report, which draws from an extensive analysis of over forty thousand teacher evaluation records, refers to the way in which existing evaluation systems fail to meaningfully differentiate teacher performance, creating a situation in which teachers are treated like interchangeable widgets. "As a result," the authors wrote, "teacher effectiveness is largely ignored. Excellent teachers cannot be recognized or rewarded, chronically low-performing teachers languish, and the wide majority of teachers performing at moderate levels do not get the differentiated support and development they need to improve as professionals."[12] The report called for school districts to replace these broken evaluation systems with comprehensive performance-based evaluation models that differentiate teachers based on their effectiveness in boosting student achievement and that provide targeted professional development to help them improve.[13]

**Teacher Evaluation 2.0.** Even as TNTP detailed the failings of current teacher evaluations, a new kind of teacher evaluation system was emerging—amid considerable controversy—in Washington, DC. In 2009, under the leadership of Chancellor Michelle Rhee, founder of TNTP, the district launched a new evaluation system—IMPACT. Under IMPACT, teachers are evaluated in four areas:

- Impact on student achievement (measured through teacher- and school-level value-added data);
- Instructional expertise (measured through formal classroom observations);
- Collaboration and commitment to school community; and
- Professionalism (which includes attendance, on-time arrival, following policies and procedures, and treating colleagues and students with respect).

For teachers in grades and subjects that take the DC assessment, impact on student achievement comprises 55 percent of a teacher's evaluation (50 percent based on individual value-added data and 5 percent based on school value-added), instructional expertise counts for 35 percent, and commitment to school community constitutes 10 percent. Professionalism is not scored, but failure to meet standards for professionalism can reduce a teacher's rating. Based on their performance in these four areas, teachers receive one of four ratings:

- Highly Effective
- Effective
- Minimally Effective
- Ineffective

These ratings have implications for both teachers' compensation and continued employment. Teachers who are rated highly effective can receive a bonus of up to $25,000, and those who are rated as such for two consecutive years may receive a base pay increase of up to $20,000. Teachers rated as effective simply advance normally on the pay scale. Teachers rated as minimally effective receive targeted professional development to help them improve, and those who do not improve after two years may be dismissed. Finally, teachers rated ineffective are subject to dismissal. With strong encouragement from the federal government, IMPACT soon became the model for a new wave of state and district teacher evaluation systems in states such as Florida and Illinois.

These fledgling efforts to improve evaluations caught the attention of national policymakers. The move toward new systems of teacher evaluation gained steam in July 2009, when the US Department of Education released draft application guidelines for the RTT program. Cre-

ated as part of the American Recovery and Reinvestment Act (ARRA)—the $787 billion stimulus bill passed by the US Congress earlier that year to revive a flagging economy—RTT was a $4.35 billion pot of money from which the US Department of Education was authorized to make competitive grants to states making progress in four "assurance areas" defined in the ARRA legislation:

- Making progress toward rigorous college- and career-ready standards and high-quality assessments that are valid and reliable for all students, including English language learners and students with disabilities;
- Establishing pre–K through college and career data systems that track progress and foster continuous improvement;
- Making improvements in teacher effectiveness and in the equitable distribution of qualified teachers for all students, particularly students who are most in need; and
- Providing intensive support and effective interventions for the lowest-performing schools.

Within those parameters, the department had broad latitude to define specific application criteria and priorities for RTT. And it chose to use the RTT guidance to promote the development of new teacher evaluation systems that reflected *The Widget Effect*'s recommendations. The "Great Teachers and Leaders" component of RTT accounted for more points than any other section of the application—more than one quarter of the total points. The largest component of that section—worth over 10 percent of the total RTT application—required states to articulate their plans for developing teacher and principal evaluation systems that evaluate all teachers and principals at least annually, include student achievement growth as a significant factor in teacher evaluations, differentiate teacher effectiveness in multiple categories, and use teacher evaluation results to inform key personnel decisions, including professional development, compensation, promotion, retention, tenure, certification, and dismissal. The department also created an eligibility requirement for RTT that made any state that established statutory barriers to using student achievement data in teacher evaluation ineligible to win an RTT grant.[14]

RTT instigated a flurry of state activity around teacher evaluation.[15] By the time the first round of applications was submitted in early 2010, eleven states had passed legislation to eliminate statutory barriers to using student achievement data in teacher evaluations, established new standards for school and district teacher evaluations, or created new state teacher evaluation systems.[16] Even more would follow in the next two years, bringing the total to nearly twenty by the end of 2011. Other states accomplished similar results through regulatory action. To support states in establishing teacher evaluation policies that met the RTT criteria, TNTP published *Teacher Evaluation 2.0*, a report drawing on extant research literature to outline six key design features for teacher evaluation systems.[17] These include:

- Annual evaluations (at least), including evaluations of veteran teachers;
- Clear, rigorous expectations that prioritize student learning;
- Multiple measures of performance, primarily the teacher's impact on students' academic growth;
- Multiple rating levels (at least four) with clear descriptions and expectations at each level;
- Frequent observations and ongoing constructive critical feedback; and
- Use in employment decisions including bonuses, tenure, compensation, promotion, and dismissal.

These features have become the grammar of the evaluation conversation. Most of the recent state laws overhauling teacher evaluation requirements also reflect these features, although states vary in the extent to which they do so, as reflected in a variety of ratings of state teacher evaluation laws produced by Democrats for Education Reform, the National Council on Teacher Quality, and Bellwether Education Partners, the latter of which was done by the authors of this paper.[18] While these ratings vary in their focus and issues covered, all reflect the six design principles articulated in TNTP's *Teacher Evaluation 2.0.* Further, some of them go well beyond the six design principals to specify how state teacher evaluation systems should address a number of topics, including the role of teacher input in evaluation development, guidelines for choosing evaluators, and an appeals or mediation process.

Most recently, guidelines for the US Department of Education's ESEA flexibility waiver process—which

allows states to request a waiver of key provisions of NCLB, including the goal of 100 percent student proficiency by 2014—require states applying for a waiver to commit to "develop, adopt, pilot, and implement" teacher and principal evaluation systems that:

- Evaluate educators on a regular basis;
- Support instructional improvement;
- Differentiate performance using at least three levels;
- Incorporate multiple measures of educator performance, including student growth as a significant factor;
- Provide clear, timely, and useful feedback to inform professional development; and
- Inform personnel decisions.

States applying for a waiver must develop and adopt guidelines for these systems and require local educational agencies to develop and implement evaluation systems that comply with those guidelines.[19] Given frustrations with NCLB in most states, these waiver requirements create a powerful incentive for states to adopt new teacher evaluation requirements aligned with federal guidelines, making it likely that *Teacher Evaluation 2.0* will increasingly displace *The Widget Effect* as the norm for teacher evaluation throughout the country over the next few years.

## Tensions and Trade-Offs

Recent state and district policy changes are leading to new evaluation systems that have the potential to provide teachers with more useful feedback about their performance, inform professional development, and ultimately improve teacher practice. The information these systems produce can also help policymakers address persistent inequities in teacher distribution; improve the quality of preservice teacher preparation; better align compensation and resource allocation decisions to what matters most for student learning; identify, reward, and retain high performers through meaningful career ladders; and identify and remediate underperforming teachers—and, when necessary, exit them from the system. In short, improved teacher evaluation policies and systems have much to offer.

That said, much of the public and policy debate around reforming teacher evaluation has been framed in stark either-or terms that obscure—rather than illuminate—real tensions inherent in these efforts. Too often, debates are framed as a choice between adopting policies that dictate the uniform implementation of the six *Teacher Evaluation 2.0* design elements across all schools and districts, and defending a status quo in which teacher evaluations provide little useful feedback to educators because nearly everyone is rated "satisfactory."

Public debate about teacher evaluation tends to vascillate between the technical (implementation challenges, issues related to the validity and stability of value-added measures, and appropriate methods for evaluating teachers in nontested grades and subjects) and the ideological (whether it is fair to hold teachers accountable for student learning and the extent to which test scores truly measure their most important contributions). But the emphasis on technical and ideological questions tends to elide fundamental tensions and trade-offs in the development of teacher evaluation systems, which deserve greater consideration and discussion than our current debate has afforded them. The remainder of this paper addresses four key tensions of the new generation of teacher evaluation systems: flexibility versus control, the role of teacher evaluation in an evolving education system, purposes for which evaluations are designed and used, and what it means to evaluate teachers as professionals. We also discuss two models—developed independently from current state policy reforms—that encompass many elements of *Teacher Evaluation 2.0* but also serve to illustrate what may be lost if states implement evaluation system requirements without caution.

## Flexibility versus Control

Tensions related to centralization, flexibility, and control of key decisions are inherent in the structure of the US education system, which vests multiple levels of government—school district, state, federal—with the power to influence what happens in schools and classrooms. Over time, policy approaches have oscillated between emphasizing centralization as a means of quality control and prioritizing flexibility to place decision making with those "closest to the child." During the Progressive Era, for example, reformers sought to centralize power and decision making at the state and urban school district level; decades later, site-based management and charter school reforms would seek to devolve power to school-level leaders. These shifts reflect the tension between recognition that school- and local-level leaders often have the best insight into those

schools' specific needs, and skepticism or distrust of their ability or will to make the right decisions.

Flexibility and control are not inherently in tension. It is possible for state and federal policies to be, to borrow secretary of education Arne Duncan's formulation, "tight" in defining accountability and outcomes goals for schools, and "loose" on the specifics of how they get there. In practice, though, it can be difficult to delineate exactly where the boundary falls between the "ends" and "means."

Teacher evaluation is a clear demonstration. On the one hand, the recent emphasis on teacher evaluations seems like a shift in a "tight-loose" direction: pay less attention to teachers' training and other characteristics, and instead focus on the results they produce in the classroom. On the other hand, mandates that teacher evaluations include specific design elements could be seen as overly prescriptive. This is particularly true in several states that have gone beyond the broad parameters laid out in *Teacher Evaluation 2.0* and the federal RTT and waiver guidelines, and now require school districts to adopt teacher evaluations that employ state-defined value-added models or specific teacher evaluation rubrics.

While federal waiver criteria require only that states create guidelines for teacher evaluation systems and ensure local school districts to implement systems that meet those guidelines, some states—including Delaware and South Carolina—have chosen to adopt a single statewide teacher evaluation system, in which all districts in the state must participate. It is one thing to say all districts must have teacher evaluation systems that include student achievement data, meaningful observations, and multiple differentiated levels of performance—it is another to mandate the specific tools used for observing teachers and analyzing performance data.

**"People Proof" Systems Are Stupid Systems.** States that mandate statewide teacher evaluation systems, value-added measures, or rubrics are not just being control freaks. Given the poor track record of most school districts in conducting meaningful teacher evaluations, compellingly illustrated in *The Widget Effect*, states have good reason to be skeptical that districts, left to their own devices, will adopt or implement rigorous evaluation systems. Moreover, getting this more rigorous teacher evaluation right is a complicated business—as evidenced by the struggles of front-runner states such as Tennessee and Delaware—and many districts lack the capacity or resources to do so without state support and guidance.

But policies that deny districts the freedom to be bad actors can also restrict their flexibility to adapt or innovate in ways that make these systems work better. This is particularly problematic because the experience of charter schools such as Mastery Charter Schools in Philadelphia and districts that have successfully used teacher evaluations seems to indicate that aligning evaluation to a particular school's or district's culture and philosophy of effective instruction is critical to maximizing the benefit of these systems (see "Mastery Charter Schools" sidebar). If a school or district has already made significant investments or built buy-in around a particular framework of quality instructional practice, requiring it to adopt a statewide rubric that is not fully aligned to that framework may be counterproductive.

It is one thing to say all districts must have teacher evaluation systems—it is another to mandate the specific tools used for observing teachers and analyzing performance data.

Casting this tension in sharp relief is the juxtaposition—in both RTT and ESEA waiver guidance—of teacher evaluation provisions with those intended to expand access to public charter schools and streamline state regulatory burden on schools. Charter schools are independent public schools of choice that, in many states, are granted broad flexibility from regulatory requirements in exchange for accountability. This often includes greater leeway than traditional schools and districts have in personnel matters, including fewer barriers to dismissing underperforming teachers. New state teacher evaluation policies, by mandating teacher evaluations that meet certain parameters, could infringe on charters' historical freedom in personnel matters. Existing exemptions from state regulations may also exempt charter schools from new teacher evaluation requirements—but federal ESEA waiver requirements seem to discourage this.

The ESEA waiver guidance states that "An SEA [state education agency] must develop and adopt guidelines for these systems, and LEAs must develop and implement teacher and principal evaluation and support systems that are consistent with the SEA's guidelines."[20] The term LEA, or "local education agency," is typically

## Mastery Charter Schools

Mastery Charter Schools is a high-performing charter network that operates nine turnaround or merged schools and one new-start charter school in Philadelphia, and it has delivered impressive results for previously low-performing schools and students across its network. Mastery began developing a teacher and principal evaluation system when it became clear that it was going to expand from a single charter school to multiple campuses—creating a need to systematize expectations and evaluation in order to ensure fairness and a consistent standard of teaching. The system, which serves as the basis for a performance-based teacher compensation system, includes three elements:

- **Instructional standards**, drawn from a variety of research- and practice-based sources, as well as educator feedback, establish common definitions and expectations for what high-quality teaching should look like across all ten campuses. Teacher performance against the standards is evaluated through classroom observations.

- **Student performance data** are measured based on student value-added scores from multiple assessments, including Pennsylvania's state test, a nationally normed exam, and regular benchmark evaluations administered as part of Mastery's instructional cycle. Mastery contracted with a statistician to develop its own value-added model using all of these data sources because the value-added data it receives from the state are both too infrequent and too late to inform instructional practice or personnel decisions. Because Mastery uses regular benchmarks as part of the instructional cycle in nearly all grades and subjects, it does not face the same nontested grades and subjects problem that most state teacher evaluation systems do.

- **Mastery responsibilities, values, and contribution** is a more subjective measure of the extent to which teachers demonstrate Mastery values and contribute to their school community, as evaluated by their principal. The first two factors comprise a greater share of the evaluation than this more subjective measure, but Mastery feels it is important that the evaluation system reflect its particular values.

More important than the specifics of Mastery's evaluation system is its integration into a broader culture and commitment to developing effective teaching. Mastery makes significant investments in professional development and individual coaching to improve teachers' performance against the instructional standards, and coaching and expectations for professional growth are not limited to underperformers. In fact, high-performing teachers actively seek out such opportunities. Coaches are evaluated based on their success in improving teachers' performance in targeted areas of the standards.

As a result, evaluations are but one piece in an ongoing conversation about instruction, performance, and improvement that is happening throughout the year among teachers, coaches, and principals. Because of this ongoing conversation, teacher evaluations are not just a score at a particular point in time, nor are the results a surprise for teachers who are engaged in ongoing conversations about their performance.

Mastery founder Scott Gordon views this larger ecosystem focus on instructional quality as critical to the evaluation system's success. "Without the clear standards, focus on professional development, and linear link between all the pieces," Gordon said, "I don't think we'd get that big a bang out of it." Gordon worries that state and district policies that create teacher evaluations without putting in place this larger ecosystem and culture will wind up disappointing their supporters. Gordon also notes that creating an evaluation system makes conversations with staff about their performance even more important. Building principals' capacity to engage teachers in these honest conversations and give them productive feedback has been one of the most difficult—and critical—tasks in implementing Mastery's teacher evaluation model.

Mastery's experience and current high performance demonstrate the benefits of an effective and well-designed teacher evaluation system: rewarding and retaining high performers, improving weaker ones, exiting people who do not improve, and building a performance culture at the school level. But obtaining these benefits takes cumulative work in implementing and refining the system and shaping a culture that values performance.

used in federal law to refer to school districts, but charter schools in many states are also their own LEAs, so this language appears to suggest that they must also adopt evaluation systems consistent with state guidelines. In the District of Columbia, for example, the Office of the State Superintendent of Education (OSSE) prepared an ESEA waiver request stating that all LEAs in the district would adopt teacher evaluations meeting OSSE-defined guidelines—even though 95 percent of the LEAs in DC are charter schools, and DC's charter school law gives OSSE no authority to mandate that they adopt such systems. In Florida, teachers' unions decried a new evaluation law as an assault on due process. But, because the law applied to charter as well as traditional district schools, it actually subjected charters to more complicated evaluation and due process requirements than before.

It is impossible to effectively evaluate teacher performance without a significant role for human judgment.

The problem with these requirements is not simply that they interfere with charter autonomy. They could also jeopardize sophisticated human capital systems that some high-performing charter schools (such as the Achievement First and Mastery charter networks profiled in each of the sidebars) have developed using their autonomy over personnel. These systems, which predate the current push for new systems of teacher evaluation, integrate teacher evaluation, development, and performance-based compensation with the school's instructional philosophy and culture, in ways that may not fit into state requirements for more uniform evaluation systems.

Much of the current rhetoric and policy surrounding teacher evaluation has, as its subtext, a desire on the part of reformers and policymakers to "people proof" or "politics proof" teacher evaluation. Value-added measures and specific observation rubrics are cast as a way to reduce the danger of subjective administrator judgments in evaluations and base them instead on objective, "scientific" criteria. The impulse to "people proof" schooling, to establish systems, protocols, and tools that eliminate room for human judgment and error, is a longstanding impulse in education—and one that often creates more problems than it solves.

Indeed, many of the systems that proponents of new teacher evaluation systems are seeking to replace—such as the "steps and lanes" teacher salary schedule—were created in part to "people proof" decisions or protect teachers from administrator bias. Our culture tends to yearn for scientific solutions that allow us to make decisions based on purely objective information without demanding human judgment. But, in fact, "people proof" decisions are often stupid decisions, and it is impossible to effectively evaluate teacher performance without a significant role for human judgment.

New evaluation systems, particularly those including student growth data, are being treated as a way to get around a specific management problem: because the culture in K–12 education traditionally discourages candid conversation about performance, school leaders have failed to take responsibility for making difficult personnel decisions. Well-designed teacher evaluation systems can help change this by giving principals a tool to enable these conversations, making them more accountable for having them, and reducing some current barriers to dismissing low performers.

But evaluations alone cannot transform a culture that is so resistant to frank discussions of performance. "Educators are not trained and have no context for supervisory conversation of that nature," says Mastery Charter Schools founder Scott Gordon, noting that his schools' implementation of evaluation systems has required intensive training of school-level leaders, not only to use the system and tools consistently—the focus of much of the training associated with current state teacher evaluation efforts—but also to provide constructive feedback and have meaningful discussions with teachers about their performance. Changing the culture and building capacity for such feedback and conversations become particularly important when we consider that the purpose of new evaluation systems is not just to dismiss low-performing teachers, but to help all teachers improve their professional practice and performance.

**Political Dynamics Create Incentives for Greater Control.** Policymakers and advocates who favor new systems of teacher evaluation recognize that they are operating in a unique political moment, in which both a Democratic presidential administration and Republican governors and legislative majorities in many states favor major changes to current teacher evaluation policies. Furthermore, unusual circumstances such as the RTT contest and state demands for relief from ESEA requirements have provided this Democratic administration with unprecedented leverage to drive state policy changes, even in states with reluctant leadership.

## Achievement First Charter Schools

Achievement First operates twenty high-performing charter schools in New York and Connecticut. Unlike the current wave of state activity on teacher evaluation, which has emphasized the need for evaluations to support dismissal of underperforming teachers, Achievement First's teacher evaluation system grew out of a desire to develop and recognize the most effective teachers.

The Achievement First network has been around for fourteen years and has grown rapidly in the past decade. To support this growth, Achievement First made significant investments in developing existing staff members to lead new schools. But effective teachers who had been in Achievement First classrooms for several years also wanted pathways to grow and develop as professionals without becoming principals or school leaders. Achievement First developed its evaluation system as part of an effort to provide these teachers with such opportunities. Achievement First's teacher evaluation system is based on four elements:

- **Student achievement**, which Achievement First believes should be a critical component of evaluations for all teachers. For teachers in grades and subjects that take the state test, it uses its own value-added model. For teachers in nontested grades and subjects, it uses end-of-course assessments to measure student learning. Student achievement accounts for a larger percentage of teachers' evaluations in tested grades and subjects, reflecting differences in the reliability and validity of the different types of assessments.

- **Student character development**, a key part of Achievement First's mission, is also a component of every teacher's evaluation. It is measured using parent and student surveys and accounts for 15 percent of a teacher's evaluation.

- **Planning and instruction** are measured through classroom observations conducted at multiple points throughout the year using a specific rubric developed by Achievement First. Most observations are conducted by building-level administrators, in some cases supplemented by other Achievement First network staff from outside the school. A comprehensive rating from the teacher's instructional coach is also incorporated into the overall planning and instruction rating. Planning and instruction counts for a higher percentage of the evaluation for teachers in nontested grades and subjects.

- **Core values and contribution** to the team are measured through principal assessment and a peer survey and count for 15 percent of a teacher's evaluation.

Achievement First's experience offers three potential lessons for broader efforts to implement teacher evaluation systems. First, because Achievement First designed its evaluation system as part of a broader effort to develop and recognize effective teachers, every aspect of the system is designed to support this goal. Second, Achievement First was honest with teachers about the challenges of developing an evaluation system and engaged them throughout its development. Finally, because Achievement First's emphasis has been on developing master teachers, it has designed its evaluation rubric to define a high standard of instructional excellence, rather than a minimum threshold. As a result, average teacher scores on the rubric have been only middling—even though Achievement First is an extremely high-performing network. This ensures that the system can support ongoing development and improvement even for high-performing teachers.

This unique—and clearly limited—political opportunity creates an incentive for champions of teacher certification overhauls to lean toward control over flexibility in designing new policies in order to lock in as much of their agenda as possible before the window closes. This means seeking to mandate now many key design features of teacher evaluation systems—see, for example, the detailed provisions of Florida's Senate Bill 736—rather than legislating broad parameters that leave space for district variation. It also means a tendency to put key provisions and requirements in legislation rather than regulation—even when regulation may be a better vehicle for addressing many complex issues involved in teacher evaluation—because legislation is more difficult to change down the road.

While the political incentives are obvious, there

is a real danger that, in seeking to lock in the most "rigorous" standards for teacher evaluation systems, policymakers and advocates may simply be locking in the current state-of-the-art knowledge regarding teacher evaluation. Policymakers need to take action, but they also need to be willing to evolve—to learn from the results of the policies they put in place and to make adjustments as results demand. Legislating key design elements in an effort to protect them can also restrict the ability of future policymakers to adapt policies in response to the new knowledge that will inevitably emerge as teacher evaluation systems are implemented and studied, or as our education system itself evolves in other key respects.

## Evaluation in an Evolving System

Teacher evaluation policies do not exist in a vacuum. Other forces are currently at work transforming our education system in ways that have real implications for the design and use of teacher evaluation—forces that could create real challenges if policymakers do not take those implications into account.

The challenge is broader than just charter schools and a few alternative schools. Technological innovation is already beginning to reshape our understanding of what schooling looks like—and with it, the work teachers do. Blended learning models, such as Rocketship, Carpe Diem, and School of One, are leveraging technology to deliver education in new ways that have the potential to increase educational productivity and personalization to meet individual student needs. Blended learning models vary in design, approach, costs, and the extent to which technology supplants in-person teachers or enables teachers to be deployed in new ways. These models fundamentally change the way teachers do their jobs; teachers spend less time on traditional lecture and practice, and more time working with students one-on-one or in small groups, analyzing data to diagnose student needs, and crafting instructional experiences to meet them.

Student groupings in these models are more flexible and fluid, and students receive instruction and tutoring from a variety of teachers, as well as from technology-based modalities. As a result, it can be difficult or impossible in some of these models to attribute student learning gains in a particular subject to a particular teacher. This creates complications for teacher evaluation systems that rely on linking teachers to their students' academic results. Existing observational rubrics—such as those included in the Bill & Melinda Gates Foundation-funded Measures of Effective Teaching project and being adopted by states and districts—were not designed for the more personalized modalities in which blended learning educators often deliver instruction.

Nor is blended learning the only force increasing the number of teachers whose students do not have state test scores. Over the past decade, states and school districts have dramatically expanded the number of children they serve in early childhood and pre–K programs—and the long-term trend of expansion in publicly funded preschool is likely to continue once states and districts rebound from the recent recession. Early childhood students also do not take state assessments, and because of the low adult-to-child ratios necessary in early childhood classrooms, preschool teachers often work in coteaching settings that make it difficult to attribute students' progress to an individual teacher.

In other words, the expansion of blended learning and early childhood programs is likely to dramatically increase the number of teachers who pose the most difficulty for new teacher evaluation systems—those whose work does not directly map onto the state assessment performance of a specific, identifiable group. The greatest challenge facing most states currently seeking to implement new teacher evaluation systems is determining how to evaluate student achievement gains for teachers in nontested grades and subjects—those not covered by NCLB's mandate to annually assess every student in grades three through eight in math and English language arts. The experience of Tennessee, winner of $500 million in RTT's first round and the birthplace of value-added measures, is a good example.

Tennessee has struggled to identify appropriate measures of student growth for teachers in nontested grades and subjects—over half of Tennessee's teachers. In the initial stages of implementation of their new system, these teachers' impact on student achievement will be scored using schoolwide value-added data in math and reading—a move Tennessee teachers and outside observers have criticized. Further, some educators whose responsibilities do not typically include classroom instruction—such as media specialists—may be required to teach mock classes to receive ratings on required observational measures.[21] Delaware, which received $120 million during the first round of RTT, has also faced challenges implementing key components of its planned statewide teacher evaluation system.[22] Like Tennessee, Delaware has struggled to identify "student growth measures" for grades and subjects not subject to state testing, and as a result delayed

teacher and leader evaluation systems based on those measures for a year.

Just because blended learning models decouple the link between an individual student's progress and an individual teacher, this does not mean that blended learning teachers cannot be held accountable for their impact on student learning. Teachers at the Florida Virtual School, for example, do not get paid if their online students do not complete their courses. Indeed, because blended learning teachers work collaboratively in teams with shifting groups of students and constantly collect and assess data on students' progress, there is more transparency for their work than for more traditional teachers. School of One uses open classes, so there is much more ongoing observation and engagement between classrooms and educators. The types of data analysis and collaborative teaching that occur in blended learning situations actually open the door for entirely new forms of teacher performance evaluation that are still tightly linked to impact on students but look very different from the 2.0 models that federal and state policymakers are currently mandating.

Unless state policies provide for additional flexibility around teacher evaluation in blended learning and other innovative approaches to schooling, we will miss out on the opportunity to develop these new forms of evaluation. More troubling, teacher evaluation requirements could actually become a barrier to the expansion of blended learning models that delink student learning from individual teachers, or to the development of new models combining in-person teaching with technology and other delivery mechanisms to personalize student learning experiences. Charter school authorizers may be unwilling to approve schools using new blended models if those schools cannot explain how they will comply with state teacher evaluation laws.

## Trade-Offs in the Use of Evaluations and Value-Added Data

Proponents of new teacher evaluation systems have emphasized a set of specific design principles—using student achievement data as a major factor, including classroom observations as a measure, and using evaluation data to inform key personnel decisions—to define what quality evaluations must look like. There is a reasonable basis for each of these design features, but proponents of *Teacher Evaluation 2.0* have too often succumbed to the temptation to oversell these systems, rather than acknowledging the complexity involved and the broader reality that progress means moving forward with imperfect models and refining them over time. Several key trade-offs deserve greater attention than they have received in the current debate. These include trade-offs between "deselection" and professional improvement as strategies to improve teaching quality, trade-offs among potential evaluators, trade-offs between individual and organizational accountability, and trade-offs between teacher evaluation and other potentially potent uses of value-added data to improve student learning.

> Evaluations alone cannot transform a culture that is so resistant to frank discussions of performance.

**Deselection or Professional Improvement?** The move toward new teacher evaluation systems is rooted in the belief that better evaluations are needed to identify underperforming teachers and remove them from classrooms. As *Teacher Evaluation 2.0* has gained political traction, however, proponents—including Secretary Duncan and President Obama—have begun to emphasize that new teacher evaluation systems are not just about firing teachers but should also provide teachers with useful feedback to help them improve their performance. This politically necessary rhetoric has the added benefit of being true. Because current value-added measures are most reliable in identifying only those teachers at the high and low end of the spectrum, with the vast majority falling in the middle, dramatically improving teacher quality through evaluation will require using those evaluations to help the majority of teachers who do not fall at those extremes to improve.

But the design of states' and districts' 2.0 evaluation systems do not necessarily back up this rhetoric. For better or worse, the design focus in most 2.0 evaluation systems has been on building systems that are sufficiently fair, valid, and reliable to hold up in court as a basis for dismissing low-performing teachers. State efforts to design new teacher evaluation systems have focused much more attention on what happens to teachers at the bottom of the spectrum than those in the middle or at the top. For example, despite the rhetoric of teacher development, several states' new teacher evaluation laws require the creation of a professional development plan only for

low-performing teachers, and primarily as a way of giving these teachers an opportunity to improve prior to dismissing them. Similarly, current design efforts have not emphasized features that are critical to ensuring that evaluations really help teachers improve. Useful feedback comes not just in the number generated by a particular evaluation measure (whether student achievement data or an observation rubric), but from the conversations that occur between a teacher and a principal, observer, or expert coach based on that data. Effective feedback requires evaluators to engage in an analytic conversation with the teacher about his or her performance that identifies strengths to build on, areas for improvement, and critical action steps to accomplish both. Evaluation systems need to be designed to facilitate and provide time for such conversations. While state evaluation systems are devoting attention to training evaluators in specific observation instruments, most are not devoting the same time to building their capacity to give meaningful feedback or engage in productive conversations about performance.

**Who Evaluates?** There are also trade-offs in deciding who should be responsible for teacher observations and evaluations. One school of thought holds that evaluating, developing, and ensuring the quality of teaching in a school is the principal's job, and therefore principals should have the primary responsibility for teacher observation and evaluations. But this raises real issues of principal capacity and time, particularly to link evaluation with ongoing professional development, which is usually delivered by coaches, master teachers, and others in the school—not by the principal. Others argue that to eliminate potential principal biases and ensure independence and reliability, independent evaluators from outside the school should conduct the evaluations, either alone or in tandem with principal reviews—and some states are including language in teacher evaluation legislation to require this. Some also argue that evaluations will be most useful for teacher development if those conducting the evaluation are also responsible for supporting and developing teachers on an ongoing basis, such as instructional coaches and master teachers. But some critics fear that evaluators who are also engaged in providing professional development may be biased in their evaluations or may be more likely to go easy on underperforming teachers. (In fact, research suggests the opposite—that coaches and mentor teachers appear to be harder in their evaluations of teachers than principals are.)[23] Each of these approaches represents differing philosophies about the ultimate aims of teacher evaluation, the incentives and behavior of different players in the system, and how to weigh trade-offs between independence, reliability, and usefulness for professional development—but these trade-offs are not always explicitly discussed in public debates about teacher evaluation systems.

Experience from high-quality evaluation models (such as those described in the appendices) suggests that effective teacher evaluation systems need to be integrated into a broader system of teacher development, support, and advancement that includes a clear definition or framework for what effective instruction looks like; intentional supports for developing teaching practices against that framework; ongoing conversations among teachers, peers, supervisors, and coaches or master teachers about professional performance, data, and improvement; and performance-based compensation. In these models, all individuals in the organization share a common organizational vision about effective instruction and are accountable for both individual and organizational progress toward aligned goals. Most current state and district efforts do not include all these components.

State efforts to design new teacher evaluation systems have focused much more attention on what happens to teachers at the bottom of the spectrum than those in the middle or at the top.

**Individual versus Organizational Accountability.** Measures for which teachers are held accountable must be aligned to those for which the larger school or organization is accountable. If teacher and organizational measures are not clearly aligned, evaluation systems may not drive teacher performance toward organizational goals. The current state ESEA flexibility waiver process raises concerns in this vein. Some states that have applied for or received ESEA waivers do not currently include value-added or growth measures of student performance in their accountability system for schools, but they are proposing to create such value-added measures to meet the

waiver requirements related to teacher evaluation. Georgia, for example, which recently received a waiver, proposes to measure school performance based on student proficiency and graduation rates, rather than growth. Some states are also proposing to integrate new value-added or growth data into revised school accountability systems—but others are not. This is troubling because it creates the potential for evaluation metrics and systems for teachers that are disconnected from those for schools and school districts. Moreover, it is a completely backwards way of adopting the use of growth or value-added measures for accountability purposes. Because value-added or growth measures for individual teachers are much less robust and more subject to error than those for schools—and also, in the proposed evaluation systems, often carry higher stakes—it makes little sense to implement them before, or even at the same time as, value-added or growth accountability for schools. In fact, it could undermine teacher and public trust in these systems.

Colorado's experience is illustrative here. When the state first developed its well-regarded Colorado growth model, it originally used the data only for public information purposes. After familiarizing citizens and educators with the measure and the information it produced, and identifying and working out kinks in the tool, Colorado began to use growth data as part of its school accountability system in 2009. Only after the growth data had been in place as part of the accountability system for schools was it extended to individual teacher evaluation and accountability. Because the need to include value-added or growth data in teacher evaluations is driving efforts to develop these measures in many states, they are not taking the time to roll out these measures in a deliberate way that builds public and educator understanding, trust, and acceptance of them.

**Evaluation Is Not the Only Use for Value-Added Data.** A focus on value-added measures for teacher evaluation has also distracted attention from other potentially valuable uses of value-added data to improve student performance—some of which may interact with teacher evaluation systems in ways that complicate their results. For example, schools and teachers can use value-added data to analyze individual students' learning trajectories and target aid or interventions accordingly. Value-added data can also help schools and districts strategically assign students to teachers in ways that match teacher strengths with student needs.[24]

Such strategies have the potential to improve student learning outcomes, but they could also impact the validity of value-added measures of teacher performance, if students are assigned to teachers on a nonrandom basis that takes into account teachers' and students' past value-added data. Other strategies to benefit students could also impact evaluations. For example, several states have passed or are considering policies that prohibit districts from assigning the same child to an ineffective teacher for two or more consecutive years. These policies make a great deal of sense, given the data on the cumulative impacts of ineffective teachers on student performance, but they would also affect assignment of students to teachers in ways that could impact evaluation results—particularly if more effective teachers received an influx of students who had been taught by ineffective teachers in the previous year.

Obviously, student and teacher assignments never have been, and never will be, random—an issue that also has implications for evaluation systems. But we should not allow concerns about the impact on evaluation systems to prevent us from doing things that might benefit students. A recent Center for American Progress report, for example, argued that one reason states should not provide raw value-added data to journalists is that parents might use this data to advocate for more effective teachers for their students, and the impact on student and teacher assignments might undermine the validity of the teacher evaluation system.[25] This seems like a backwards argument.

Ultimately, value-added and evaluation data can be used in a variety of ways to improve both teacher effectiveness and student achievement, beyond the uses that tend to feature prominently in most policy debates about teacher evaluations. Failure to consider those uses and the trade-offs involved in evaluation design could result in missed opportunities to improve student learning.

## Evaluating Teachers as Professionals?

Efforts to establish new teacher evaluation systems are often accompanied by heated disputes about their impact on the standing of the teaching profession. Advocates argue that holding teachers responsible for their performance will bring teaching more in line with norms in other fields. Furthermore, they maintain that new evaluations will help raise the status of the profession by encouraging dismissal of poor teachers who give teaching a bad name and by facilitating the implementation of performance-based compensation programs that increase salaries for the most effective teachers. Critics of new evaluation sys-

tems argue that mechanistic teacher evaluations based on test scores and rubrics are demeaning and demoralizing and neglect the nuanced art that goes into teachers' work. Both sides have a point.

New teacher evaluation systems should be understood as part of a larger effort to move attitudes toward human capital in education away from an industrial-era model that treats all workers as interchangeable parts with little differentiation of pay, status, or responsibility, and toward a more performance- and talent-sensitive orientation along the lines of law, medicine, and other professions. The title of *The Widget Effect* alludes to this. Gordon, Kane, and Staiger also argue that increased accountability for student achievement will ultimately improve the status of teaching and attract more skilled people to the field.[26]

But some features of 2.0 teacher evaluation systems and policies are in fact very different from evaluation norms in other professions. Most professional fields, including business, medicine, and law, rely on a combination of data and managerial judgment when making evaluation and personnel decisions, and subsequently hold managers accountable for those decisions. Methods that lack human judgment and discretion are rare. Indeed, far from eliminating subjective feedback from personnel evaluation, many firms have moved to adopt "360 degree" feedback mechanisms in which employees receive feedback from managers, direct reports, peers, and clients. Others combine objective quantitative measures—such as dollars or clients brought into the firm—with more subjective or soft performance indicators. 2.0 policies, in contrast, have sought to minimize the role of managerial judgment through the use of "objective" value-added data, common rubrics, and third-party evaluators. Rather than building the capacity of managers and educators to provide meaningful development feedback to their direct reports and peers, these systems seek to minimize the extent to which they are expected to do so.

There are clear tensions associated with using human judgment in teacher evaluations. As is often the case in education, politics and mistrust exacerbate the challenge of designing evaluation systems. The current emphasis on "objective" data and rubrics, as well as third-party evaluators, is in large part a response to the opposition of teachers' unions and teachers' associations to past evaluation and performance pay initiatives that were viewed as giving principals too much managerial discretion. Obviously no one wants a system in which employees are at the whim of arbitrary or biased managerial judgments. But some labor concerns reflect a pre-union, pre–civil rights era when employees lacked many of the protections against discrimination and arbitrary dismissal that exist today—and that will continue to exist regardless of changes in evaluation systems or policies.

> Teaching is ultimately a people business, and getting improvements in teacher effectiveness will require human judgment and interaction—processes cannot do it alone!

Moreover, the best protection against arbitrary or biased managerial judgment is not to eliminate that judgment altogether, but to ensure that the managers themselves are also held accountable for performance.

In designing value-added systems, policymakers should consider whether the design elements they are putting in place move education away from or toward professional norms in other similar fields. For example, policy and design decisions are currently driven in large part by concerns about whether or not evaluation systems can withstand legal challenges to dismissal of the worst teachers—a legitimate focus in systems that convey a property interest in continued employment on tenured teachers. But the result is that design elements that support professional improvement can become secondary.

Human judgment is critical both to making smart assessments of teacher performance and to using those assessments in ways that improve instruction for students. Given the limitations of our current value-added and observational rubric tools, "people proof" evaluation systems will almost certainly result in nonnegligible numbers of both Type I (false positive) and Type II (false negative) errors. That is not an argument against improving evaluation systems through the use of these tools—our current system, in which more than 99 percent of teachers are rated satisfactory, yields few false positives but likely many false negatives—but it is an argument for providing space for human judgment to intervene when such errors seem obvious, along with the right incentives and capacity building to encourage sound managerial judgment. By the same token, building evaluation and professional

growth systems that truly develop teachers as professionals is impossible without providing space for human judgment and feedback apart from purely "objective" and impersonal mechanisms.

## Policy Implications

Because there are real trade-offs and tensions in the design and use of teacher evaluation systems, we need systems that are flexible and able to adapt to particular circumstances and changes over time. It is also critical to publicly and transparently engage policymakers, educators, and the broader public in conversations about trade-offs, rather than glossing over them. Such conversations can provide an opportunity for input about the priorities policymakers should take into account in making trade-offs related to teacher evaluation systems, and in the process can build public and educator understanding and trust in those systems. There are several key implications for policymakers to consider:

**Be Clear about the Pain Points.** Policymakers need to be clear about the problems teacher evaluation systems are intended to solve. Right now, teacher evaluations are too often marketed as an educational wonder drug, without a clear theory of action about how evaluation results will translate into improved teaching or the other system elements necessary to foster effective teaching. Policymakers must be clear about the problems they are trying to address, their goals, and their theory of action—and they must make design choices and trade-offs that reflect that theory of action.

**Do Not Treat Processes and Systems as a Substitute for Cultural Change.** Policymakers are relying heavily on new teacher evaluation systems and the processes they mandate to improve teacher effectiveness. But teaching is ultimately a people business, and getting improvements in teacher effectiveness, whether through professional improvement or deselection, will require human judgment and interaction—processes cannot do it alone! That, in turn, requires deep cultural change within the US educational system. Teacher evaluation systems and processes can help facilitate this cultural shift, by setting clear expectations, creating language and venues within a school to talk about performance, and empowering leaders to confront or dismiss low performers. But these systems cannot carry all the work on their own. Other elements of the education system, such as meaningful outcomes accountability, competitive pressure, and new approaches to training are also needed to foster a culture that takes performance seriously.

**Look at the Entire Ecosystem.** Some advocates and policymakers talk about teacher evaluation systems as if they are being implemented in a vacuum, without considering how they relate to other elements of the education ecosystem. How will reforms that tie personnel decisions to evaluations based on student achievement affect the labor market for teachers? What changes in principals' pre- and in-service training are needed to develop their capacity as effective evaluators? How should teacher preparation requirements change to align better with new evaluations or standards? Policymakers must pay attention to the entire teacher labor market ecosystem, not only to the point at which high-stakes evaluations take place or have professional consequences for teachers.

Similarly, policy changes are increasing high-stakes individual accountability for teachers even as they in some cases reduce school-level accountability, and the direction of new state accountability systems under ESEA waiver requests is not always well aligned with states' educator accountability proposals. This is unwise. Strong school-level accountability is essential to effective teacher evaluations and must be aligned with them. Because well-designed and aligned school- and system-level accountability create the right incentives and pressures for school-level leaders, they can reduce the need for excessive state control on teacher evaluations and create space for more judgment.

**Focus on Improvement, Not Just Deselection.** Research shows that there is a subset of teachers who adversely affect student learning and are less effective than the average first-year teacher—which means that deselecting these teachers and replacing them with a random new teacher would more likely than not improve student achievement. But these teachers are only a fraction of the total workforce, and deselecting them is not enough to generate the level of improvement needed. Evaluation systems based solely on creating legally robust methods for removing low performers are insufficient. Excessive efforts to decouple evaluation from human judgment and create standards that can withstand legal scrutiny move education away from, rather than toward, the professional norms that guide similar fields.

Several key policy implications arise here: First, policymakers must be honest about the need for teacher evaluation to include a component of professional judgment and should design systems accordingly. Second, principals

and other evaluators need training not only in the technical aspects of evaluation systems, but also in how to provide effective feedback and engage in honest conversations about performance to help teachers improve. They also need adequate time to do so. Related to this, policymakers should carefully consider trade-offs in selecting evaluators and should not discount the impact of the choice on the ability to support professional growth and development. Finally, evaluation components must be aligned with one another so teachers do not receive mixed messages about what they need to do to improve.

**Encourage and Respect Innovation.** Education has a long history of "one best system" thinking that leads to policies that freeze in amber the current state of the art, which curtails innovation and renders the system unable to adapt to changing needs and challenges. It would be a mistake to bring this one best system thinking to teacher evaluations. A better model to emulate might be the approach taken by law or other professional services firms where there are some clear commonalities in operations, norms, and performance expectations, but evaluation and accountability metrics reflect diversity in what is valued most. Similarly, policymakers must ensure that evaluation requirements do not curtail existing autonomies. Charter schools, for instance, should be held accountable to their authorizers for their outcomes, rather than bound by new evaluation requirements that curtail their autonomy in personnel matters.

**Think Carefully about Waivers versus Umbrellas.** One potential strategy to ensure that new teacher evaluation systems do not become a barrier to innovation is to provide broad waiver authority enabling innovative providers to gain an exemption to teacher evaluation requirements if they can demonstrate strong performance or well-designed alternative teacher evaluation and development mechanisms. Relatively few states have built such well-designed waivers into their teacher effectiveness legislation, and those that have not would be well advised to do so. At the same time, policymakers need to think carefully about whether it is possible, and perhaps smarter in the long run, to design teacher evaluation policies with a broad umbrella that can cover both traditional and innovative models. The answer to this question likely depends on one's assumptions about the potential scale of blended learning and other innovative models and the speed at which they will grow, as well as how teaching arrangements in these models are likely to differ from traditional schools. But it is worth taking some time to explicitly consider these questions, and policymakers' hypotheses about them, when making decisions about teacher evaluation policies.

**Do Not Send Legislation to Do a Regulation's Job.** In a rush to institutionalize, reformers have turned to laws to protect reforms. Laws are obviously more durable than regulations, and the legislative process can—although not always—build greater buy-in from stakeholders. But legislation can also lock in policies that should be tweaked or even overhauled. Because the old model of evaluation was so ubiquitous and we are only beginning to experiment with alternative models, there are many things we do not know, and implementation of different models is likely to yield considerable learning about what does and does not work in different contexts. What can and should be handled in legislation versus regulation varies with state context, but in general, policymakers should try to avoid locking in legislation components of evaluation systems for which implementation is likely to provide important lessons about how to do things better.

**Create Innovation Zones for Pilots—and Fund Them.** Within the overall context of federal and state policy, there must be room for school districts, consortia of school districts, or even entire states to try new approaches. Federal dollars from Title II of the ESEA and the Teacher Incentive Fund can support innovative projects to try alternative teacher evaluation methods that combine quantitative and qualitative methods or are designed specifically for new education delivery models. In keeping with its role in fostering research and innovation, the federal government should not simply require states to establish new teacher evaluation policies, but should simultaneously provide waivers and investments to support state, local, and charter school innovations that meet certain standards for rigor and protection of civil rights. As part of its fiscal year 2013 budget request, the Obama administration proposed a $5 billion competition for states to work with colleges of education, teachers' unions, and other stakeholders to reform the teaching profession. An evaluation innovation pilot fits squarely with the goals of that initiative.

## Conclusion

Public education in the United States has for too long lacked a performance mindset or a strategic orientation toward developing human capital. The current move

toward new teacher evaluation systems represents significant progress to correct these shortcomings. That said, if advocates of 2.0 teacher evaluation rush too quickly to create new systems or do so without appropriate humility about what we do and do not know, there is a risk that they will end up replacing old broken systems with new ones that, while better, are equally inflexible or create barriers to innovation and reform. In other words, the nation's teacher evaluation spree could turn into a big headache. The best way to mitigate this risk is not to ignore it or brush it under the rug, but to be honest and transparent about the trade-offs and tensions, the reality that new systems will not be perfect, and the need to learn as we move forward.

## Notes

1. Andrew Rotherham and Sara Mead, "Back to the Future: The History and Politics of State Teacher Licensure and Certification," in *A Qualified Teacher in Every Classroom?*, ed. Frederick M. Hess, Andrew Rotherham, and Kate Walsh (Cambridge, MA: Harvard Education Press, 2004), 11–48.

2. Sandra L. Stotsky, "Revising Teacher Licensing Regulations to Advance Education Reform in Massachusetts," in *Education Success Stories* (Amherst: National Evaluation Systems, Inc., 2001), www.pearsonassessments.com/hai/images/NES_Publications/2001_14Stotsky_466_1.pdf (accessed August 1, 2012).

3. Emma Smith and Stephen Gorard, "Improving Teacher Quality: Lessons from America's No Child Left Behind," *Cambridge Journal of Education* 37, no. 2 (2007): 191–206, www.tandfonline.com/doi/abs/10.1080/03057640701372426 (accessed August 1, 2012); and Robert Rothman and Patte Barth, "Does Highly Qualified Mean Highly Effective?" Center for Public Education, 2009, www.centerforpubliceducation.org/Main-Menu/Staffingstudents/How-good-are-your-teachers-Trying-to-define-teacher-quality/Does-highly-qualified-mean-highly-effective.html (accessed August 1, 2012).

4. Dan D. Goldhaber, "The Mystery of Good Teaching," *Education Next* 2, no. 1 (2002): 50–55.

5. Charles Clotfelter, Helen Ladd, and Jacob Vigdor, "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects," *Economics of Education Review* 26, no. 6 (2007): 673–82; and Goldhaber, "The Mystery of Good Teaching."

6. Jennifer Rice King, "The Impact of Teacher Experience: Examining the Evidence and Policy Implications," *CALDER Policy Brief*, no. 11 (August 2010), www.urban.org/uploadedpdf/1001455-impact-teacher-experience.pdf (accessed August 1, 2012).

7. Dan D. Goldhaber and Dominic J. Brewer, "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," *Journal of Human Resources* 32, no. 3 (1997): 505–23.

8. June C. Rivers and William L. Sanders, *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement* (Knoxville: University of Tennessee Value-Added Research and Assessment Center, 1996), 9, http://heartland.org/sites/all/modules/custom/heartland_migration/files/pdfs/3048.pdf (accessed August 1, 2012); and Kati Haycock, "Good Teaching Matters: How Well-Qualified Teachers Can Close the Achievement Gap," Education Trust, 1998, www.take2theweb.com/pub/sso/eastlinton/images/Good_teaching_matters.pdf (accessed August 1, 2012).

9. William L. Sanders and Sandra P. Horn, "The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment," *Journal of Personnel Evaluation in Education* 8 (1994): 299–311, www.redmond.k12.or.us/145410515152938173/lib/145410515152938173/The_Tennessee_Value-Added_Assessment_System-_Mixed-Model_Methodology_in_Ed_Assessment.pdf (accessed August 1, 2012).

10. Kevin Carey, "The Real Value of Teachers: Using New Information about Teacher Effectiveness to Close the Achievement Gap," *Thinking K–16* 8, no. 1 (2004): 3–42, www.edtrust.org/dc/publication/the-real-value-of-teachers-using-new-information-about-teacher-effectiveness-to-close (accessed August 1, 2012).

11. Robert Gordon, Thomas Kane, and Douglas Staiger, "Identifying Effective Teachers Using Performance on the Job," Brookings Institution, 2006, www.brookings.edu/papers/2006/04education_gordon.aspx (accessed August 1, 2012).

12. Daniel Weisberg et al., *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (Chicago: The New Teacher Project, 2009), http://widgeteffect.org/ (accessed August 1, 2012), 6.

13. Ibid.

14. US Department of Education, "Race to the Top Program Executive Summary" (Washington, DC, November 2009), www2.ed.gov/programs/racetothetop/executive-summary.pdf (accessed August 1, 2012).

15. Bellwether Education Partners staff, including two coauthors of this paper, were involved in writing or advising state RTT applications.

16. Learning Point Associates, "State Legislation: Emerging Trends Reflected in State Phase 1 Race to the Top Applications," June 2011, www.learningpt.org/pdfs/RttT_State_Legislation.pdf (accessed August 1, 2012).

17. The New Teacher Project, *Teacher Evaluation 2.0* (Brooklyn, NY, 2010), tntp.org/files/Teacher-Evaluation-Oct10F.pdf (accessed August 1, 2012).

18. Ron Tupa, Jocelyn Huber, and Barbara Martinez, "Built to Succeed? Ranking New Statewide Teacher Evaluation Practices," Democrats for Education Reform, October 17, 2011, http://www.dfer.org/Report%20-%20Evaluation%20Ratings%20DRAFT9.pdf (accessed August 1, 2012); National Council on Teacher Quality, "State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies" (Washington, DC, 2011), www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf (accessed August 1, 2012); and Sara Mead, "Recent State Action on Teacher Effectiveness: What's in State Laws and Regulations?" Bellwether Education Partners, August 2012, http://bellwethereducation.org/ideas/publications (accessed August 1, 2012).

19. US Department of Education, "ESEA Flexibility" (Washington, DC, September 23, 2011), www.ed.gov/esea/flexibility (accessed August 1, 2012).

20. Ibid., 7.

21. Michael Winerip, "In Tennessee, Following the Rules for Evaluation Off a Cliff," *New York Times*, November 6, 2011.

22. US Department of Education, "Race to the Top Annual Performance Report" (Washington, DC, 2012), www2.ed.gov/programs/racetothetop/annual-report.pdf (accessed August 1, 2012).

23. Craig D. Jerald and Kristan Van Hook, *More than Measurement: The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems* (Santa Monica, CA: National Institute for Excellence in Teaching, 2011), www.tapsystem.org/publications/eval_lessons.pdf (accessed August 1, 2012).

24. Craig D. Jerald, "The Value of Value-Added Data," Education Trust,