Conversations with Arne Duncan
Offering advice on educator evaluations

The national push for new teacher evaluations is real; educators should not miss this opportunity to influence policy makers with solid evidence.

By Carol Corbett Burris and Kevin G. Welner

We can’t say how many high school principals get calls from Secretary of Education Arne Duncan, particularly when he knows he’ll be speaking with a critic of his policies. We do know that he got an earful when he called the principal of South Side High School in New York, Carol Burris (one of the authors of this article).

Burris’ correspondence with Secretary Duncan began on July 5, 2011, when she published an open letter to him on The Washington Post’s “Answer Sheet” blog. She writes in that letter about the test-based evaluation of teachers and principals, a policy that’s received increasing national attention from policy makers. She describes the hard-working teachers at her high school and explains how New York state’s new educator evaluation system would harm the work they do:

[T]he punitive evaluation policies that New York state has adopted (and that many other states have adopted) due to the Race to the Top competition are . . . a dangerous gamble that might score political points but . . . will hinder what you and I and so many others want — better schools for our kids. We already know from research that reforms based on high-stakes testing do not improve long-term learning.

She also writes about the negative climate that New York’s teachers and students now experience due to rating teachers by student test scores. “Both students and teachers feel the brunt of this distrust,” she noted. The open letter described the New York policy as “the legacy of the policies that were rushed into place by states to get the federal Race to the Top money.”
A few weeks later, Secretary Duncan called Burris. They spoke about the evaluation of teachers by test scores, and he listened to her concerns. He recognized the problems inherent in test-score based evaluation systems, especially when year-to-year comparisons are made. They both agreed on the importance of high-quality evaluation systems, and Secretary Duncan asked Burris to send him her ideas regarding effective evaluation policies. The two of us then prepared a response. What follows is the central argument contained in our July 27 letter to Secretary Duncan. (The full text of this letter is available on the web site of the National Education Policy Center: http://nepc.colorado.edu/publication/letter-to-Arne-Duncan.)

Evaluations can be powerful interventions. High-quality, thoughtful evaluation carries the potential to improve schooling. Misguided evaluation approaches, however, have a corresponding potential to harm our schools. It is essential, therefore, that we understand how to design evaluation systems that have the greatest likelihood of improving rather than undermining school performance. In our view, any evaluation system must be judged on the basis of its overall effect on student learning and that determining an evaluation regime’s overall effect implicates at least four categories of information: (a) summative data, (b) formative data, (c) the nature of school working conditions, and (d) incentives for students, teachers, and administrators.

**Summative data.** Summative data have been the dominant focus of recent policies. This has primarily meant test-score results. The key goal of analyses of such summative data is to highlight excellent educators and dismiss ineffective ones.

**Formative data.** Formative data are used to improve teaching and help educators become better at their profession. Formative data and summative data are used together in a sound evaluative system. For educators who are struggling, for instance, formative feedback that is supportive yet frankly discusses the need for and nature of the requested improvements often plays a counseling-out role that can obviate the need for formal dismissal. If the evaluation system has a well-functioning formative component, few educators should require dismissal (because of improvement or voluntary exits), instruction should improve, and student achievement should increase.

**Working conditions.** Research has consistently shown that the character of school leadership and nature of school culture are foremost among the reasons teachers choose to stay or leave a particular school or stay or leave the teaching profession (Boyd, et al., 2011). The best evaluation systems should enhance the working environment of teachers.

**Incentives.** A related and critical question about any evaluation system is how it affects daily incentives and disincentives for students, teachers, and administrators. This is perhaps the most obvious lesson we can take away from the experience with No Child Left Behind. As schools were placed under an incentive system linked to test scores, we saw a narrowing of the curriculum, teaching to the test, and other potentially harmful practices.

Our response to Secretary Duncan expresses other concerns as well, stressing the importance of evidence-based policy making, and explaining why pilot evaluation studies are preferable to the large-scale adoption of unproven state-mandated evaluation systems based on untested assumptions. We note, for example, that there are no formal studies connecting educator evaluation systems that use test-score growth data with learning outcomes, making their effectiveness impossible to judge.

Perhaps most importantly, we point out to Secretary Duncan that teachers with ineffective teaching skills nevertheless might have strong value-added scores, especially when they teach high-achieving students (Hill, Kapitula, & Umland, 2011). As a practical matter, this means that these evaluation systems will reward some teachers not entitled to be rewarded while other teachers might be unfairly dismissed. Further, because higher growth scores are correlated with students who enter the class with

---

CAROL CORBETT BURRIS (cburris@rvcschools.org) is principal of South Side High School, Rockville Centre, New York.
KEVIN G. WELNER (kevin.welner@colorado.edu) is director of the National Education Policy Center, and a professor of education policy, both at the University of Colorado at Boulder.
higher achievement, this system creates a disincentive to teach those with greater disadvantages.

We conclude (as have numerous scholars) that the existing evidence cannot be fairly read to support an educator evaluation system such as New York’s. There’s no reason to believe that such a system will validly identify and remove those who are unable or unwilling to improve; will improve the effectiveness of all others; will identify excellence in teaching or leadership; will provide incentives for good practices (or avoid incentives for poor practices); or will enhance school environment and working conditions.

Just as no pharmaceutical would be brought to market without first being tested for effectiveness and for adverse reactions, neither should a practice with the potential to profoundly impact the lives of the nation’s students and their teachers. Considering both the cost and the high-stakes nature of mandated evaluation systems, our letter offers Secretary Duncan the following recommendations.

1. Put on hold the policy push to use student test scores to evaluate teachers and principals, unless and until data demonstrate the likelihood that such an evaluation approach will positively, not negatively, affect student learning and its accuracy morally justifies its use. Existing systems that use student scores for educator evaluation are already in place. These should be treated as pilots and should be used to understand the systems and their results, including effects on student achievement.

2. More broadly, call upon the National Research Council or the National Academy of Education to document teacher- and principal-evaluation approaches that are proven to successfully meet all four criteria for sound evaluation practices listed above. Such a report might also identify and describe promising additional approaches and evaluate pilot programs and evaluations of those approaches. Based on this report, the U.S. Department of Education could embark on an evidence-based policy that would continue the existing push for high-quality educator evaluation while ensuring that the specific push will benefit the nation’s students.

3. While awaiting evidentiary guidance from the work of the National Research Council or National Academy of Education, focus the federal push on rigor and balance. Educator evaluation systems should pursue the four criteria for sound evaluation practices, recognizing also that multiple measures, pursued diligently and conscientiously, will allow weaknesses in any given measure to be compensated for by others. In lieu of obliging states to impose a nonevidence-based evaluation approach, the federal government should encourage the use of well-designed and well-executed locally appropriate strategies. In this regard, one of the most longstanding and promising teacher evaluation approaches relies on peer assistance and review (PAR) programs, such as those in Toledo, Ohio, and Montgomery County Public Schools in Maryland. We note with alarm the likelihood that current policies are not just failing to promote such programs with apparently successful track records — the new wave of evaluation policies are actually discouraging and terminating these successes (Winerip, 2011).

4. Whatever system is used, require that it be subject to rigorous outcome monitoring; that is, locally designed review and evaluation.

5. Require that all evaluation systems enhance the professionalism of teaching and the principalship. Systems like the one in New York will almost surely undermine that professionalism. Similarly, public dissemination of teacher- and principal-level value-added data will undermine attempts to improve performance. For example, given the different degrees of efficacy among parents, it is likely that demand for highly rated teachers will result in students with the greatest need being assigned to the lowest-rated teachers.

On Aug. 17, three weeks after we sent this letter, Secretary Duncan called Burris again. Among other things, he said he supports peer review of teachers; however, he still believes student test scores are an important evaluative component.

During this second call, Burris described why teachers at her school were now hesitant to take on at-risk students. Secretary Duncan didn’t believe that they should be reluctant, expressing his faith in the capacity of value-added models to compensate for bias in student assignment. The Secretary’s apparent brushing aside of the limitations of value-added modeling illustrates a very important point: If policy makers do not understand the research concerning the technical limitations of this tool, they will support policies that rely on the models to produce valid and reliable numbers for individual educators.

We should note that Burris’ correspondence did not begin with Secretary Duncan. Back on June 3, 2011, she wrote a letter to President Obama, de-
tailing her concerns about the emphasis on student test scores as a prominent part of educator evaluation. Interestingly, she also received a substantive response from the president in a letter dated July 29. He praises her leadership and her school but defends the administration’s policies:

I fully recognize that any system of accountability will not be able to perfectly measure teacher effectiveness, but I respectfully disagree with your suggestion that the closest thing states have to an objective measure of student achievement should not be part of the equation. At the same time, test scores should never be the only factor . . . I am confident that state, district, and school leaders, working in concert with teachers, can properly weigh the many factors that affect student achievement.

Burris’ response, sent to the president on Aug. 22, includes the following:

Allow me to briefly address what I consider to be the primary argument in your letter. You write, “I respectfully disagree with your suggestion that the closest thing states have to an objective measure of student achievement [value-added growth scores based on standardized tests] should not be part of the equation.” As a matter of measurement, I think this contention has some merit . . . .

As you note, the test-based evaluation of schools under NCLB has led to teaching to the test. It has also led to a narrowing of the curriculum. Further, many tests (at least in New York state), have proven to be flawed measures of student learning. In addition, test-based evaluations of schools led, however inexcusably, to fairly widespread cheating. This approach has stifled engaging teaching and learning as well as creativity. And, notwithstanding your compliments to me, a sound policy cannot depend on all schools having principals who will successfully push back against the incentives and disincentives created by such a system.

New York’s new system requires that these scores be used for up to 40% of educator evaluations. If this level were healthy, I would not already be noting that teachers are becoming hesitant to take on a student teacher due to fears about subsequent evaluation by test scores. Similarly, as one of my teachers recently told me, “I always felt flattered, Carol, when you gave me the tough kids to teach. I don’t want to worry if they are assigned to me now — but I will. The kids who are hard to motivate and have weak skills will be the kids that no one wants when their jobs are on the line.” As Professor Welner and I explained in our letter to Secretary Duncan, the research is clear that value-added growth models do not fully compensate for student differences. That is, this teacher’s worries are well-grounded.

I also worry for the students in racially and socioeconomically isolated schools who can least afford poorly supported experimentation. Although poverty can never be an excuse for lack of achievement, neither can its effects on student learning be ignored . . . . As opposed to the test-focused reforms that are currently in vogue, reforms that emphasize classroom supports and equitable structures are well-supported by research evidence and deserve the policy attention they are now being denied.

This correspondence with President Obama and Secretary Duncan captures the divide that has emerged between the Obama administration and various self-designated reformers on the one hand, and educators and researchers wary of the unintended results of policies linking high-stakes consequences to students’ scores on standardized tests. We can find broad agreement on the importance of accountability and evaluation, but the important specifics are in intense dispute.

Although the primary focus of this article has been on the evaluation of educators, we urge readers to keep in mind that teachers and principals work within schools and communities that are themselves crucial to the success or failure of our educational efforts. If we fail to invest in our schools and communities, even the highest-quality educator evaluation will lead to little success. We can’t ignore the impact of poverty, racially and socioeconomically segregated schooling, and inequitable funding. Here is our worry: As we attach reform efforts to evaluation systems, we tend to neglect those factors that matter most. Yes, having a highly effective teacher increases the probability of student learning. But there is no evidence at all that the new evaluation systems will produce more effective teachers. As Secretary Duncan himself has said, “We cannot fire our way to excellence.”

Despite the presumed good intentions of many policy makers, certainly including Secretary Duncan and President Obama, policies that could do irreparable damage to public schools are now in place. New York’s policy has been challenged in court, and that challenge has met with some initial success. But the national push is still very much ongoing. The good news is that if educators and the public speak up, respectfully but loudly, people will listen. If we demand policies grounded in high-quality research evidence, we might eventually see beneficial and substantial changes. ☓

References

