7-30-2011

Letter to Secretary of Education Arne Duncan Concerning Evaluation of Teachers and Principals

Kevin G. Welner
University of Colorado Boulder, Kevin.Welner@colorado.edu

Carol C. Burris
burriscarol@gmail.com

School of Education, University of Colorado at Boulder
Boulder, CO 80309-0249
Telephone: 802-383-0058

NEPC@colorado.edu
http://nepc.colorado.edu

NEPC POLICY MEMO

LETTER TO SECRETARY OF EDUCATION ARNE DUNCAN CONCERNING EVALUATION OF TEACHERS AND PRINCIPALS

Carol Corbett Burris, Rockville Centre School District

Kevin G. Welner, University of Colorado at Boulder

This NEPC Policy Memo presents the text of a letter from Drs. Burris and Welner to Secretary of Education Arne Duncan. The letter was invited by Secretary Duncan during a phone conversation with Dr. Burris. It offers concrete guiding principles for evaluation of educators and suggestions for a way forward.

Dear Secretary Duncan:

Thank you for calling me on July 14th and listening to my concerns regarding the current policy push for teacher and principal evaluations linked to student test scores. Because you and I agreed on the importance of evaluation in assuring high-quality educators for our students, you invited me to email you additional thoughts about evaluation and alternative evaluation ideas. I have taken that invitation quite seriously, and I appreciate the opportunity to enter the dialogue on this most important topic.

In order to ensure that I'm accurately representing the research on this topic, I enlisted the help of my friend Kevin Welner, who is a professor of education policy and evaluation at the University of Colorado at Boulder and also directs the National Education Policy Center.[1] Kevin and I have collaborated on several research studies focused on what we call "universal

acceleration" – combining high expectations and strong supports for teachers, principals and students.

What follows is our best thinking on how to best evaluate public educators in order to better serve our students.

High-Quality Evaluation

Although our focus is on evaluation of educators, it is imperative 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 cannot ignore factors such as racially and socio-economically segregated schooling and its associated impact on student achievement, nor can we ignore factors such as inequitable funding. These have proven to be intractable problems, but they must be addressed.

None of this, however, detracts from our agreement that high-quality evaluation is fundamental to good policy and practice. We also agree that many teachers and principals have not been evaluated in a systemic and useful way in the past. That is, we agree that there is a great deal of room for improvement.

Evaluations can be powerful interventions. Although high-quality, thoughtful evaluation carries the potential to improve schooling, misguided evaluation approaches have a corresponding potential to harm our schools. Like most policy tools, evaluation can be used soundly and beneficially, or it can be abused.

How then do we evaluate an evaluation system? We propose that it be evaluated by its overall effect on student learning. Such an overall effect implicates at least four overlapping areas: (a) summative, (b) formative, (c) working conditions, and (d) incentives. Each is these is described briefly below.

The summative role. This has been the dominant focus of recent policies, including SB 191 in Colorado and APPR in New York. The key goal is to highlight excellent educators and dismiss ineffective ones. We agree that this is an important evaluative function. Accordingly, it is imperative that evaluation systems be both valid and reliable, to ensure that the decisions that serve the best interests of students are made.

The formative role. The key formative goal is to improve teaching and help educators become better at their profession. The formative role can and should co-exist with the summative role in a sound evaluative system. For educators who are struggling, formative feedback that is supportive yet frankly discusses the need for extensive 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 voluntarily exit), instruction should improve, and student achievement should increase.

Working conditions. As we discussed during our July 14th conversation, an evaluative system can change the way a school functions, for better or for worse. In the latter case, the system can waste valuable time on tasks that do not support the essential mission of schooling: teaching and learning (see examples in my open letter).[2] Proponents of rigorous evaluative systems argue that evaluation's effects reach beyond the information derived; good systems can encourage better teaching. We agree that this can be a positive outcome. We also, however, think it foolish to ignore the downsides—specifically, negative impacts on school culture, collegiality, and teachers' relationships with administration and with students. Research has consistently shown that the school leadership and school culture are foremost among the reasons teachers remain at a school and even stay in the teaching profession.[3] The best evaluation systems should enhance these key working conditions; an evaluation system that harms these conditions should only be used as a last resort.

Incentives. Relatedly, a crucial question about any evaluation system is how it affects daily incentives and disincentives for 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. This alone does not mean that test scores shouldn't be used for high-stakes evaluative systems; however, decision-making is improved if it includes a very careful look at unintended consequences. The unintended consequences of test based systems were well expressed in the recent report from the National Research Council.[4]

In sum, good evaluative systems should have a positive effect on student learning and achievement by identifying and removing those who are unable or unwilling to improve, by improving the effectiveness of all others, by identifying excellence in teaching and leadership, by providing incentives for good practices (and avoiding incentives for poor practices), and by enhancing school environment and working conditions.

What We Know and Don't Know

As the recent RAND Corporation research brief on the New York City pay-for-performance plan stressed, "[p]ilot testing and evaluation are essential" when embarking on an unproven program.[5] Indeed, because of such pilot testing and to the credit of the district's leadership, an ineffective and expensive New York City practice (additional pay for higher test scores) was summarily abandoned.

There is no question that educator evaluation systems based in large part on student test scores are unchartered waters. Yet the statewide systems are not pilots; they are full-blown mandates imposed on all public schools. What we are engaging in is a national experiment that is costly in public dollars attached to high-stakes consequences for educators and students alike.

There are three possible results, then, that can result from the state-mandated systems (systems that unfortunately were in most cases prompted by the guidelines of Race to the Top):

1) student outcomes might increase;

2) student outcomes might remain the same; or

3) student outcomes might decline.

Even if we were to measure just short-term outcomes, pilot studies linked to careful evaluations could tell us a great deal about which of these three results is most likely.

Result 1 would be welcome. The second and third results would not. With either of these latter two results, millions of tax dollars dedicated to training, implementation, testing, and data analysis would be wasted. More effective solutions would not be pursued. Our students would pay the price.

Although there are no formal studies connecting educator evaluation systems that use test score growth data with learning outcomes, there are two recently published reviews of the Washington DC teacher evaluation system IMPACT, which is in some ways the prototype of test-score-based evaluation of educators. It has been in existence for two years. When student scores are used for IMPACT teacher evaluations,[6] its design looks similar to many of the state-adopted models that rely heavily on student scores. For this reason, we thought a closer look might provide insight into educator evaluation systems and in particular the relationship between the two main types of evaluation used in IMPACT: observations and test scores.

What we found out gave us cause for concern.

Ideally, there would be a strong correlation between a teacher's value-added score and the score derived from careful observations. A correlation of 0.60 and above is generally accepted to be a strong relationship. This would mean that the district is measuring, with each part of the evaluation, something akin to true teacher quality. Yet the DC IMPACT program showed a relationship of only 0.34 between teacher value-added scores and the scores from evaluations (primarily observational) linked the district's Teaching and Learning Framework observation scores. This "modest correlation" concern was raised in an evaluative report of IMPACT published by the Aspen Institute.[7]

At one level, this relatively weak relationship between the two components of the IMPACT evaluation is testimony to the district's wisdom in incorporating both elements in the evaluative system. But at another level, it raises red flags about the reliability and validity of one or both.

Indeed, this is not the first time a lack of a strong relationship was found. A prominent peer-reviewed article published a few months ago found that teachers with ineffective teaching skills nevertheless might have strong VAM scores, especially if they taught high-achieving students.[8] As a practical matter, this means that some teachers will receive bonuses when they should not, others will not receive bonuses when they should, and still others might be unfairly dismissed—to the detriment of students as well as the teachers themselves. Further, because higher growth scores are correlated to students who enter the class with higher achievement, this system creates a disincentive to teach those with greater disadvantages. That is, even models like DC's that attempt to control for prior achievement fail to capture the full effect of ongoing advantages and disadvantages.

In light of these concerns, we next looked at the associated student scores since IMPACT was enacted. One would expect that if the system were effective we would see an accelerated increase

in student scores as teaching improved due to training, coaching, evaluation and the pressure on teachers to increase test scores. This was not the case.

In 2007, only 37.5% of all DC elementary students (through grade 6) were proficient in reading and 29.3% were proficient in mathematics. By 2009, the final year prior to IMPACT, the percentage of elementary students who were proficient was 49% in both reading and mathematics. There was a dramatic increase between 2007 and 2008, followed by an additional year of growth. However, two years later (during the IMPACT years) half of those gains were lost. The percentage of students proficient in reading and mathematics fell to 43% and 42.3%.

Between 2007 and 2009 the percentage of students proficient in secondary reading increased by over 10 percentage points; the increase in math was nearly 13 percentage points. However, during the IMPACT years there was only a 4% increase in students proficient in reading and a 7% increase in math. Although secondary students did not lose ground during the IMPACT years, progress decelerated.

We note in particular that evaluations based on student test score growth would be more common in elementary schools, covering classroom teachers in grades 4 through 6, as opposed to secondary schools, where that component is only applicable in grades 7 and 8 and even then only for reading and math teachers. And we note that the post-IMPACT results are worse at the elementary level.

These are correlational results, and we cannot make any causal inferences or claims. In fact, a good argument could be made that IMPACT's effects – good or ill – would not likely be felt so quickly.[9] However, a sound research design attached to pilot programs could carefully address all these issues. Certainly the data we found suggest nothing to be enthusiastic about, even though IMPACT is only one of many factors that may affect scores. Put another way, wouldn't the children in New York, Colorado, and other states moving toward such systems benefit from solid research evidence from DC, particularly if IMPACT is indeed having a negative effect? And put yet another way, if teachers are being evaluated and dismissed based on the IMPACT data, shouldn't the program itself also be subject to a summative evaluation? Shouldn't such evaluations take place before any scaling up of this experimental policy?

The existing evidence cannot be fairly read to support an educator evaluation system such as APPR in New York. There is 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. In short, while an evaluation system should fulfill summative, formative, working condition, and incentive roles, the current push appears to fail in each area.

Recommendations

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, we offer the following interrelated 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. Existing systems, such as IMPACT, 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.[10] Such a report might also identify and describe promising additional approaches and recommend 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 be beneficial for 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 non-evidence-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 long-standing 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 having the effect of discouraging and terminating these success.[11]

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

5. Insist that all evaluation systems enhance the professionalism of teaching and the principalship. As I explained during our conversation, the New York APPR policy 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.

Thank you for the invitation to contribute our thoughts about this most important issue. We hope that the ideas set forth above are received in the spirit with which they are offered: a desire to contribute to a healthy dialog, grounded in the assumption that we are all pursuing the same

goal of reform that benefits the nation's students. Please do not hesitate to contact us with any questions, concerns, or requests for further input.

Notes and References

[1] Please note that our ideas presented here do not necessarily reflect the views of our respective employers.

[2] http://www.washingtonpost.com/blogs/answer-sheet/post/an-open-letter-to-ed-secretary-arne-duncan/2011/07/04/gHQABpIJyH_blog.html

[3] See, e.g., Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, *48*(2), 303-333; Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *The Journal of Human Resources*, *XXXIX*(2), 326-354; Lankford, M., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools. A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62.

[4] Committee on Incentives and Test-Based Accountability in Public Education (Michael Hout and Stuart W. Elliott, Editors) (2011). *Incentives and Test-Based Accountability in Education*. Washington DC: National Research Council.

[5] Li, J. (2011). *What New York City's Experiment with Schoolwide Performance Bonuses Tells Us About Pay for Performance*. Washington DC: RAND Corporation. (Page 3.)

[6] For reading and math teachers in grades 4 through 8, the student growth component of the evaluation is set at 50 percent. For other educators, various other evaluation components take on greater importance.

[7] Curtis, R. (2011). *District of Columbia Public Schools: Defining Instructional Expectations and Aligning Accountability and Support*. Washington DC: The Aspen Institute. (Page 22.)

[8] Hill, H. C., Kapitula, L., and Umland, K.A (2011). Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal*, *48*(3), 794-831.

[9] An exception might be the immediate effects on school environment and working conditions.

[10] Approaches have also been described by experts in this area and should be carefully considered as part of the National Research Council or the National Academy of Education examination. See Goe, L., Bell, C., & Little, O.

(2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. See also Hinchey, P. H. (2010). *Getting Teacher Assessment Right: What Policymakers Can Learn from Research*. Boulder, CO: National Education Policy Center. Retrieved July 23, 2011 from http://nepc.colorado.edu/publication/getting-teacher-assessment-right.

[11] See Winerip, M. (2011, June 5). Helping Teachers Help Themselves. *New York Times*. Retrieved July 23, 2011 from http://www.nytimes.com/2011/06/06/education/06oneducation.html