

# Forum on Educational Accountability

<http://www.edaccountability.org>

## What Should Congress Do About Teacher Evaluation?

### *A Public Letter from the Forum on Educational Accountability*

The U.S. Education Department's Race to the Top (RTTT) program requires states to develop teacher and principal evaluation and accountability systems that must include, "in significant part," data on student "achievement." As RTTT-winning states have started to create their systems, student scores on standardized tests are frequently a large component, even half, of a teacher's score. The questions before Congress, as part of the reauthorization of the Elementary and Secondary Education Act (ESEA) or in related legislative or budget language, include: What is the purpose of educator evaluation? What is the proper federal role? Should states be required to create educator evaluation systems? If so, what if any particulars should Congress mandate?

The Forum on Educational Accountability (FEA)<sup>i</sup> believes the most important purpose of educator evaluation is to improve teaching, learning and schools. It is one part of wider efforts to strengthen professional learning and schools. Other potentially important uses of evaluation are secondary. (FEA has outlined its general expectations for student growth in *Empowering Schools and Improving Learning*; these include rich, comprehensive academic learning and the education of the whole child. FEA addresses professional development and systemic reform in *Redefining Accountability* and other reports.<sup>ii</sup>)

FEA urges Congress to proceed with great caution. There is not enough evidence that evaluation systems are ready for large-scale state use or that they are a more important use of school improvement funds relative to other efforts. Thus, Congress should not establish federal mandates, but could provide assistance to promising programs and support research. In addition, substantial research concludes that current student tests cannot reliably, validly and fairly be used to judge educators, as some states and school districts have begun to do. The over-emphasis on standardized testing in No Child Left Behind (NCLB) has had many harmful consequences. These include narrowing the curriculum, teaching to the test, undermining educator morale, and damaging school climate. Those negative consequences predictably will intensify if educators are required to be judged "in significant part" by student test scores.

FEA makes the following specific recommendations to Congress regarding educator evaluation. We base them on evidence about current educator evaluation programs and the consequences of No Child Left Behind, which we summarize following our recommendations.

1) ***Congress should not mandate that states construct statewide evaluation systems as a condition of receiving funds under ESEA.*** Given the history of RTTT, FEA is concerned that a grant program will become de facto a mandate. Congress should allow states and local school districts the flexibility to determine whether and how to conduct educator evaluations. Further, Title II should not become a competitive grant program that requires participating states or districts to develop an educator evaluation system.

2) *If Congress creates a voluntary grant program, it should not mandate any particular, fixed weighting for student learning in general or for test scores in particular.* This includes not mandating the use of “growth” or “value added” measures. To avoid narrow teaching to the test, Congress should require states to use multiple sources of different types of evidence of student learning gathered over time.

3) *Congress should require that educators and local leaders play a major role in the design and implementation of any federally-funded state or local educator evaluation program. Parents, students and various civic and other organizations should be included.*

4) *Congress should require the independent review of federally funded educator evaluation programs for effectiveness and beneficial or harmful consequences.* Require the National Academy of Sciences or a comparable body to review successes, problems and potential improvements. Studies also should determine whether programs are following the FEA recommendations found in this letter. Congress could also study existing programs if it is not ready to support creation of new systems.

## **Discussion of FEA recommendations on educator evaluation**

FEA believes that fair and strong educator evaluations can play a helpful role in improving education. Some evaluation systems have proven their value, such as the one in Montgomery County, Maryland, among other peer evaluation programs (Burriss and Welner, 2011). These were developed locally and implemented carefully, with joint participation by educators and leadership. They have focused on aiding improvement at the individual and school levels.

*Flexibility and caution in a voluntary program.* If Congress chooses to support voluntary programs, it should provide support and flexibility to states and districts seeking to establish or strengthen evaluation programs. Districts could align their programs with their overall approach to teacher and administrator quality and effectiveness. Evaluation would fit with professional learning and the recruitment and retention of educators. However, states and districts may well conclude there are more pressing needs for their resources than evaluation programs. Therefore, Congress should not mandate that states develop educator evaluation systems, much less any particular system. It could provide funding for research and implementation that supports a broad range of locally developed strategies to improve teacher and principal effectiveness.

There are additional reasons for Congress to proceed with caution. While there are strong local evaluation systems, good evaluation systems do not exist on a large scale. Thus, we know too little about the benefits relative to the costs, the positive or negative consequences, of particular approaches. Burriss and Welner (2011) use the Washington, DC, IMPACT evaluation system to show how such systems may not produce expected results. Even assuming that raising standardized test scores was valid as the goal for school reform, student test scores did not improve in the wake of the IMPACT system, though it is clear there has been great emphasis on raising scores in DC. Further, correlations between student scores and observations of teachers were weak, as they have been in other studies.

In such circumstances, it makes sense to proceed slowly. Congress could support some efforts to design or expand systems, test them out, evaluate them very carefully, fine tune them, and then help extend the systems where and how it may be appropriate. It makes far more sense to

conduct independent evaluations of existing programs in order to build a knowledge base than it does to charge hastily down the path of mandating statewide educator evaluation systems.

***Cooperative program development and implementation.*** FEA emphasizes educator participation with local leadership in the development and operation of evaluation systems. Existing local, high-quality systems share this feature. Simply obtaining “buy in” for a plan created by others is not adequate. School boards, administrators, teachers and other educators should work cooperatively to improve educator evaluation and performance. Parents, students and other stakeholders also should play a serious role in guiding the development and proper use of an evaluation system.

***Including evidence of student learning.*** FEA expects that evidence of student learning will be part of an evaluation system, as experts and organizations in the field recognize and support. However, Congress should not mandate how student learning will be included, nor should it specify that student test scores comprise any fixed weight within the “student learning” category. Congress also should not require any additional standardized testing in any subjects or grades.

FEA comes to this conclusion from the hard lessons of NCLB. We have seen great increases in state test scores that are not matched by gains on the National Assessment of Educational Progress. Indeed, under NCLB, NAEP gains have slowed or stopped in both reading and math, for all grades, overall, and for most subgroups in most instances (Neill, 2011). The field is awash in evidence of teaching to the test, narrowed curriculum, and unfortunately increasing stories of cheating. These negative consequences ensued from NCLB’s requirements to test too much and to attach high stakes to the test scores. Mandating use of student test scores as a “significant part” of educator evaluations will only exacerbate the problems of teaching to the test.

That the measurement process can cause distortions is a widely known phenomenon in social science, termed “Campbell’s law.” It says, “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor. . . [W]hen test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways” (Campbell, 1976). NCLB’s focus on school test scores fueled damaging educational consequences. Adding educator evaluations based “in significant part” on student test scores will only compound the problem.

***“Value added” and “growth” measures.*** There are those who say problems with tests can be solved with “growth” or “value added measurement” (VAM). Essentially, these models track student gains on standardized tests across the grades, then use statistical procedures to determine each student’s rate of gain compared with other students or to a statistical projection of how each specific student was expected to perform based on her/his previous score pattern.

However, the models have major problems. One, they rest on the same inadequate, mostly multiple-choice tests used in NCLB, and thus in no way solve the “teach to the test” problem. Two, the models are descriptive, they cannot adequately show causation, that is, attribute student score changes to specific educators (Braun, 2011). Third, they are fraught with errors when used to judge educators. (For an excellent, brief listing of these problems, see Baker 2011a.)

For example, the researchers Peter Schochet and Hanley Chiang (2010) show that, even with three years of student test scores, teachers are rated inaccurately one time out of four. It is worse with only one or two year’s data. Tim Sass (2008) reports that more than two-thirds of the

bottom-ranked teachers one year had moved out of the bottom ranks the next year. One third moved from the bottom 20 percent one year to the top 40 percent the next. Only a third who ranked highest one year kept their top ranking the next, and almost a third of the formerly top-ranked teachers landed in the bottom 40 percent in year two. Other studies have found similar instability in “value-added” rankings. When different VAM models are applied, the results can vary dramatically (Briggs and Domingue, 2011). Wayne Au (2010-11) explained, “Because of these error rates, a teacher’s performance evaluation may pivot on what amounts to a statistical roll of the dice.”

These errors are so significant that inevitably many good teachers and principals will “fail,” while some weak ones will score high. Such results will push a substantial number of high-quality educators to leave the profession. Knowledge of the arbitrary and inaccurate consequences will deter some potentially strong educators from becoming teachers or principals. These problems will particularly affect schools with large proportions of low-income children, as Fend, Florio and Lee show for Florida (2011). Already there is a shortage of qualified teachers of special education and English language learners. Use of student test score-based teacher evaluations would exacerbate this shortage because their students’ conditions unfairly subject these teachers to greater risk of “failing” to adequately boost their students’ scores.

Proponents claim that VAM and “growth” measures can separate the contributions of individual teachers from the effects of the school, the family and the community. However, they cannot do so sufficiently well for making decisions about educators (Baker, 2011a). Baker (2011b) points out that “student growth profiles” are even less accurate in this regard than is VAM. Teachers who work with students in better-resourced schools in more affluent communities will be rewarded, while those working in more difficult circumstances or with more challenging students will be penalized. Teachers will be pit against one another and against students in an inevitably desperate effort to boost test scores, damaging school climate and undermining learning (Massachusetts Working Group, 2011). Use of VAM or “growth models” could exacerbate various forms of segregation. For example, students with disabilities or English language learners could be kept out of general classrooms because they could drag down the scores.

A long list of leading researchers have concluded that “value added” and “growth” measures are not ready for use in judging educators. For example, former Educational Testing Service principal research scientist Howard Wainer concludes, “[T]he more you know about VAM the less faith you have in the validity of the inferences drawn from it” (cited in Braun, 2011).

**Conclusion.** Our children deserve far more than a test-centric curriculum and instruction. Congress should ameliorate, not intensify, that problem. It should not mandate or overly encourage statewide educator evaluation systems or require the use of student test scores in the development of any federally-funded evaluation system.

September 20, 2011

## **Bibliography**

Au, Wayne. 2010-11. “Neither Fair Nor Accurate.” *Rethinking Schools*, Winter, pp. 34-38. Available at [http://www.rethinkingschools.org/archive/25\\_02/25\\_02\\_au.shtml](http://www.rethinkingschools.org/archive/25_02/25_02_au.shtml).

- Baker, Bruce. 2011a, March 13. "7 reasons why teacher evaluations won't work." *The Record*. [http://www.northjersey.com/news/education/evaluation\\_031311.html?page=all](http://www.northjersey.com/news/education/evaluation_031311.html?page=all).
- Baker, Bruce. 2011b, September 3. "Take your SGP and VAMit, Damn it!" <http://schoolfinance101.wordpress.com/2011/09/02/take-your-sgp-and-vamit-damn-it/>.
- Braun, Bob. 2011, Sept. 15. "Braun: Christie misses the mark on grading teachers, author says." [http://blog.nj.com/njv\\_bob\\_braun/2011/09/braun\\_christie\\_misses\\_the\\_mark.html](http://blog.nj.com/njv_bob_braun/2011/09/braun_christie_misses_the_mark.html)
- Burris, Carol Corbett, and Welner, Kevin G. 2011. Letter to Secretary of Education Arne Duncan Concerning Evaluation of Teachers and Principals. National Education Policy Center. <http://nepc.colorado.edu/publication/letter-to-Arne-Duncan>.
- Briggs, Derek, and Ben Domingue. 2011, 2. "Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by The Los Angeles Times." National Education Policy Center. <http://nepc.colorado.edu/publication/due-diligence>.
- Bushaw, William J., and Lopez, Shane J. 2010. A Time for Change: The 42<sup>nd</sup> Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools. *Phi Delta Kappan*, Vol. 92, No. 1 (September 2010): pp. 8-26.
- Campbell, Donald. (1976). Assessing the Impact of Planned Social Change. The Public Affairs Center, Dartmouth College, Hanover, NH. <http://www.eric.ed.gov/PDFS/ED303512.pdf>.
- Feng, Li; Florio, David; and Sass, Tim. 2010. School Accountability and Teacher Mobility. National Center for Analysis of Longitudinal Data in Education Research, Working Paper 47. <http://www.urban.org/UploadedPDF/1001396-school-accountability.pdf>
- Massachusetts Working Group on Teacher Evaluation. 2011. Flawed Massachusetts Teacher Evaluation Proposal Risks Further Damage to Teaching and Learning. FairTest. <http://fairtest.org/flawed-ma-teacher-evaluation-proposal-report-home>
- Neill, Monty. 2011. NAEP Results Show Slowing or Stagnant Results, for Most Demographic Groups, in Reading and Math, at All Ages/Grades, since the Start of NCLB. FairTest. [http://www.fairtest.org/sites/default/files/NAEP\\_results\\_main\\_and\\_long\\_term.pdf](http://www.fairtest.org/sites/default/files/NAEP_results_main_and_long_term.pdf)
- Sass, Tim R. 2008, November. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." National Center for Analysis of Longitudinal Data in Education, Policy Brief 4. [http://www.urban.org/UploadedPDF/1001266\\_stabilityofvalue.pdf](http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf).
- Schochet, Peter Z., and Hanley S. Chiang. 2010, July. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." U.S. Department of Education, Institute for Education Sciences. NCEE 2010-4004. <http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20104004>.

---

<sup>i</sup> FEA is an alliance of education, civil rights, religious, disability, parent and civic organizations. Our work is based on the Joint Organizational Statement on NCLB, now signed by 153 national organizations. Our papers and reports are at [www.edaccountability.org](http://www.edaccountability.org). This letter has not been specifically endorsed by the Joint Statement signers, though it has been circulated to them for discussion.

<sup>ii</sup> See also *All Children Deserve the Opportunity to Learn* and *A Research- and Experience-Based Turnaround Process*, at [www.edaccountability.org](http://www.edaccountability.org).