



HARVARD
GRADUATE SCHOOL OF EDUCATION

DANIEL KORETZ
HENRY LEE SHATTUCK PROFESSOR OF EDUCATION

August 28, 2009

The Honorable Arne Duncan
Secretary
U.S. Department of Education
400 Maryland Avenue, S.W.
Washington, D.C. 20202

Re: “Race to the Top” Docket ID: ED-2009-OESE-0006

Dear Secretary Duncan:

As someone who has studied education reform and educational testing for more than 25 years, I have serious concerns about the Notice of Proposed Priorities for the Race to the Top Fund and State Fiscal Stabilization Fund published in the *Federal Register* on July 29, 2009. While the proposed priorities include a number of laudable features, some of the core elements are unlikely to succeed and may cause substantial unintended harm.

The proposal places excessive focus on test scores

Most important, the draft priorities would exacerbate the already excessive focus on a narrow range of test scores as measures of educational success. Evaluations of test-based accountability systems have shown that this narrow focus often leads educators to de-emphasize other critically important goals of education and induces undesirable instructional strategies and time-wasting test preparation. And as a result of the latter, it often substantially inflates scores, giving the public an illusion of success and making it impossible to distinguish with confidence between high-performing and low-performing schools. These disturbing findings are consistent with decades of research on performance accountability systems on other sectors of the economy. Moreover, there is not even much indication that the nation’s recent experience with this approach under NCLB has even substantially improved trends in tested subjects.

Evaluating schools accurately and fairly requires more than a set of test scores. It requires a variety of indicators, not only indicators of student achievement. The priorities allow the use of other measures in addition to scores, but they do not require them, and they place vastly greater emphasis on test scores than anything else. Moreover, accurate evaluation requires human judgment, to understand the additional factors influencing student performance and the appropriateness of specific interventions.

I urge you to require explicitly that states make substantial use of additional indicators. States should be required to specify the range of measures they would use to evaluate schools and the methods they will employ to make use of expert judgment in evaluating schools, either overall or on a targeted basis.

The proposal places unwarranted reliance on value-added measures

The proposed priorities place great emphasis on using value-added measures to evaluate teachers and principals. This is unjustified by the scientific literature and will result in frequent and serious errors in evaluating practitioners. Among the most important issues are the following:

1. Value added estimates for individual teachers are extremely imprecise. A great deal of the variation among educators—particularly teachers—will be simple noise, and many teachers will therefore be misclassified as effective or ineffective. The only practical way to lessen this problem is to accumulate estimates for each educator over a period of years, to combine value-added estimates with other data, or both.
2. Value added methods provide estimates of student growth, but they cannot be trusted to indicate teachers' contributions to that growth. There is considerable disagreement in the scientific community at present about the extent to which—and the conditions under which—value-added estimates even approximate measures of teachers' effectiveness. The effects of many other variables, such as school organization, characteristics of the student population, etc., will often masquerade as “teacher effects.”
3. Apart from the problem of unwarranted inferences about teacher effectiveness, value-added methods will often provide misleading comparisons of student growth. If two classes are making identical rates of progress but one has a curriculum that better matches the test, that class will erroneously be shown as having greater growth. Particularly where there is substantial curricular differentiation, e.g., in middle-school mathematics, these errors may be very large.
4. A number of studies have already shown that value-added estimates can be markedly inconsistent from test to test or even across parts of a single test. This inconsistency might reflect differences in alignment (as in the previous comment), variations in score inflation, or both. Regardless of the reason, the consequence is that ranking of teachers will vary, sometimes markedly, depending on the choice of test or even the choice of weights given to different parts of a subject within a test. Given that these choices are to some extent arbitrary, rankings of teachers will be also.
5. There is as yet no convincing evidence that switching the accountability system to a value-added approach will have an appreciable positive effect on educators' practices or on student learning.

In sum, there is no scientific justification for making value-added estimates the primary basis for evaluating teachers. However, tracking growth has some clear advantages over the current

system. If the Administration has decided to explore the utility of growth models in the ESEA accountability system, a more useful approach would be to require states to:

1. Take appropriate steps to implement growth models for describing the progress of schools;
2. Devise a system for combining the results with other data about school quality; and
3. Facilitate or sponsor independent evaluations of the effects of these innovations on educational practice and student learning.

The proposed use of NAEP puts NAEP at risk

The proposed priorities call for using the National Assessment of Educational Progress (NAEP) to monitor the impact of RTT programs:

We propose using the NAEP to monitor overall increases in student achievement and decreases in the achievement gap over the course of this grant because the NAEP provides a way to report consistently across Race to the Top grantees as well as within a State over time as the State transitions from its current assessments to the high-quality assessments (as defined in this notice).

While using NAEP to audit gains on state tests has been common, using it as a direct measure of program impact will very likely degrade the quality of NAEP. Research suggests that one of the most serious problems with high-stakes testing is inflation of scores as the accountability pressure distorts instructional practice. NAEP must be protected from this bias.

Instead, the NPRM could call for using state tests, but with the following conditions:

1. Scale scores must be reported as well as performance standards, and all changes in achievement gaps must be reported in terms of scale scores and effect sizes. Trends reported in terms of performance standards are very difficult to evaluate, and using performance standards will always distort conclusions about trends in achievement gaps.
2. Descriptive data must be provided for all subgroups to ensure that any apparent shrinkage of achievement gaps is not merely an artifact of ceiling effects, etc.
3. To the extent not precluded by (a) legitimate needs for item security, such as linking, and (b) legitimate needs to protect individual confidentiality, both actual test items and detailed data (at the level of individual items) must be released to qualified researchers for evaluation.
4. Priority should be given to states that will implement methods for verifying score gains, such as administration of state-sponsored sample-based audit tests or linking scores to postsecondary performance.

NAEP could then be used separately by USED to audit state gains.

The priorities for improvements of assessments should be reconsidered

The proposal focuses on the importance of improving assessments used in our accountability systems. I agree that tests used for accountability need improvement, and I have recently focused my own work on this problem. However, neither the rationale for changes to current testing programs nor the principles for improving them are clear in the proposed priorities.

First, we should not embark on a wholesale replacement of current state tests until it is clear which aspects of current testing programs warrant replacement or modification. Current state tests vary in terms of content, cognitive demand, mix of formats, breadth of coverage, and many other aspects of design. In the terms used in the proposed priorities, they vary in terms of their ability to “measure a student’s understanding of, and ability to apply, critical concepts through the use of a variety of item types, formats, and administration conditions.” Moreover, because of the demands for additional testing imposed by NCLB, current state tests represent a very large commitment of development resources from the testing field, which is small and already stretched very thin. Simply discarding all of the current tests would waste resources and run the risk of low quality in the development of replacements.

Second, new test development should be based on a clear, research-based understanding of the design principles that are most important for this particular use of tests. The proposed priorities do not demonstrate this. For example:

1. The priorities ignore what may be the single most important failing of tests currently used for accountability, which is the predictability of their content and format. This predictability provides opportunities for narrowed instruction, inappropriate test preparation, and score inflation.
2. There is no rationale given for the focus on technology. There are ways to employ technology to improve assessments used for accountability, but the use of technology in and of itself would be of no benefit, and it would divert resources that could better be put to other uses. Technology should be encouraged only as a means to a specific end.
3. Extensive use of performance tasks, while desirable in for some purposes, is difficult in an accountability context for technical reasons, entails substantial compromises (e.g., lessening the coverage of the test), does not always result in measurement of higher-order cognitive skills, and will not address the core problems of inappropriate test preparation and score inflation. It too should be encouraged only as a means to a specific end, and states should that propose heavy reliance on complex performance assessment should address its disadvantages in their designs and evaluation plans.
4. A critically important issue in current accountability programs is the use of a single survey test to evaluate all schools and classrooms, regardless of their curricula. This is particularly problematic at the high-school level, where curricular differentiation is often great. When variations in alignment with a single test are substantial, that test will provide both biased comparisons and inappropriate incentives to educators. A few states have responded to this issue by implementing end-of-course exams, but there has

been little effort to evaluate this alternative. The proposed priorities do not address this issue.

A competition for funds for assessment development should not be held until there is an opportunity for careful discussion about the most important design principles for tests used for this specific purpose and about the tradeoffs they entail. These questions entail complex technical issues, and the Administration should convene groups of professionals in the measurement field to help craft appropriate guidance.

Thank you for your consideration of these comments. I would be happy to elaborate on any of these issues if you or your staff would find it helpful.

Sincerely,

A handwritten signature in black ink, appearing to be 'D. Koretz'.