Helen F. Ladd, August 20, 2009

# Comments on Proposed Regulations for Education Stimulus funds

# Docket ID ED-2009-OESE-0006, Race to the Top.

I am writing to object to the heavy emphasis in the regulations on using student test scores for the formal evaluation of teachers and school principals. While student test scores clearly have a role to play in the overall effort of improving schools, they need to be kept in their place. The regulations you are proposing gives them a pride of place that will lead to little good and is likely to do much harm.

As an academic researcher with experience working with longitudinal data on students, teachers and principals, I have estimated value added models examining the effects of teacher credentials, examined teacher and principal labor markets, and evaluated school- based accountability programs.

**Potential for harm**

The main problem with the heavy focus of the proposed test-based approach is that it ratchets up the pernicious narrow test- based approach to education represented by No Child Left Behind (NCLB). The approach is narrow in part because the requirement that all students be tested every year means that students can be tested in only a limited number of subjects. The result is a heavy emphasis on the basic skills of math and reading, to the detriment of other skills and orientations that young people need to become effective participants in the global society. Further, the emphasis on test results for individual teachers will exacerbate the well-documented incentives for teachers to focus on narrow test taking skills and drilling. It is time to move beyond this misplaced emphasis on test scores in a few subjects to return to the broader goals of education that have been such an important part of our history.

**Any positive effects are likely to be limited at best**

Consider the main two arguments underlying the push for test based evaluation of teachers. One theory of action seems to be that holding teachers more accountable for the gains in their students' test scores will induce them to become better teachers. At this point, I am not aware of any credible evidence in support of that proposition. The best direct evidence on that point is likely to emerge next spring from a random-assignment study of performance based pay for teachers by researchers at Vanderbilt financed by the U.S. Department of Education. It seems premature at best to assume that those results will be positive.

Indirect evidence from schools' experiences with the test based pressures of NCLB or its state level precursors suggests that any positive effects are likely to be small at best. While such school-based accountability pressures, which apply to all teachers in each school rather than to individual teachers, have undoubtedly raised scores on the high stakes tests on which they are based, there is far less evidence that they have raised test scores on lower stakes tests such as NAEP. The best recent evidence

on the achievement effects of NCLB is by Tom Dee and Brian Jacob. Using NAEP data, they conclude that the federal legislation appears to have increased student achievement in fourth grade math, but not in fourth grade reading, or 8th grade math or reading. Such findings, which are fully consistent with other research, provide little or no support for the view that test based incentives for individual teachers will lead to significant gains in student achievement, especially at levels beyond the most basic that are the easiest to measure with standardized tests. That outcome is not surprising given that the heavy focus on test scores diverts attention away from the more fundamental task of investing in the capacity of teachers and schools and in the capacity of children to arrive at school ready to learn.

A second possible theory of action is that the linking of students' scores to their teachers will improve student outcomes by providing the information necessary to make it easier for administrators to dismiss low performing teachers. I agree with the underlying presumption that too few teachers are currently dismissed and that current systems for dismissing teachers are cumbersome and not very effective. Moreover, it may well be sensible for administrators to make some use of student test score information in their evaluations of teachers, provided such information is embedded in a more comprehensive approach. The question is how much and how to do it in a fair way. My own recent overview of the research on teacher effects highlights the tremendous difficulties that arise in using student test scores in a fair way to evaluate teachers. Simple approaches that focus on whether a teacher's students in one year perform at higher levels than her students in the previous year are clearly inappropriate because of the changing mix of students from year to year.

The "solution" proposed by researchers such as Thomas Kane ,who is cited in the current draft regulations, is to use longitudinal data for students linked to teachers to measure the gains in achievement of the specific students in a teacher's class , which then can be compared to some standard expected gain. But this "value-added" approach is fraught with difficulties. Even the most sophisticated approaches typically cannot distinguish the contribution of teachers from the classroom context, and they generate estimates of a teacher's quality that jump around from one year to the next, largely because of the small sample sizes for individual teachers.

Although some researchers , such as Thomas Kane and his coauthors, argue that these admittedly flawed value added measures are preferred to existing measures for making high stakes decisions for teachers, one must be careful with that argument. It implies there are only two options for evaluating teachers – the ineffectual current system or the deeply flawed test-based system. In fact many states and districts are currently experimenting with various other options, including peer evaluations. If the Department of Education is seriously interested in improving the evaluation of teachers and making them serve the constructive purpose of improving student achievement, the Department should actively encourage states to experiment with a range of approaches that differ in the extent to which they rely on student test scores. Moreover, those experiments should all be fully evaluated. It is far too early in the development o f new evaluation systems for the Department to impose its preferred solution on the states.

**Appropriate and inappropriate uses of test scores**

Nonetheless, linking longitudinal data on student test scores to teachers can be useful for some purposes. As a researcher I have benefitted greatly from the availability of such data in the state of North Carolina and I trust that state policy makers have benefitted from my research and that of others. The goals of research, however, differ significantly from those of administrators. Researchers are interested in evaluating programs or understanding general patterns and have no interest in the performance of a specific teacher or principal. Thus for research purposes, data linked to teachers can be extremely useful, especially if the outcome data can also be linked to a rich set of social factors that provide information on the social context of the students and the school. But, as I have emphasized above, such data are far less appropriate for making high stake decisions about individual teachers. Nor are they useful for evaluating school principals. With respect to principals, my most recent research using the North Carolina data indicates that the quality of school leadership accounts for at most 10- 15 percent of the predicted variation in the achievement of $4^{th}$ and $5^{th}$ grade students across schools, after controlling for the individual characteristics of the students.

For both teachers and principals, it is neither fair nor constructive to try to hold them accountable for factors over which they have little control, using statistical measures that are based on a narrow range of outcomes, and that are subject to large amounts of random variability.

Helen F. Ladd is the Edgar Thompson Professor of Public Policy, Sanford School of Public Policy, Duke University. She is also a co-chair of the Broader, Bolder Approach to Education. She is writing these comments in her capacity as an academic researcher.