Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within the Race to the Top Agenda

December 2009

Test Score Inflation: Comments on *Some Implications of Current Policy for Educational Measurement* by Daniel Koretz

Robert L. Linn

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Test Score Inflation: Comments on *Some Implications of Current Policy for Educational Measurement* by Daniel Koretz

Robert L. Linn

University of Colorado at Boulder

This paper is based on a reaction by Robert L. Linn to presentations by Daniel Koretz at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of the papers presented at the seminar at http://www.k12center.org/publications.html.

As Koretz has noted, over the past 40 years there has been a substantial shift in the ways in which educational achievement test results are used. The stakes attached to test results, for educators and sometimes for students, are much higher today than they were in the 1970s. No Child Left Behind (NCLB) requirements substantially upped the ante for schools over what it had been prior to 2002, and the Race to the Top initiative promises to up the ante for teachers. The current era of high-stakes test-based accountability (TBA) has increased the pressure on tests in many ways.

Koretz referred to statements by Lindquist (1951) and Campbell (1979) that predicted distortions and corruption of results are caused by the high-stakes uses of quantitative indicators, such as test scores. One of the anticipated effects of TBA is score inflation, that is, scores on high-stakes tests that give an exaggerated estimate of the level of student achievement. Score inflation can be evaluated by comparing the gains in test scores on a high-stakes test, such as a state test used for TBA, to the gains made on a low-stakes test, such as NAEP. State-by-state NAEP, however, is only administered every other year and only covers two subjects at two grade levels. Hence there is a need to identify other means of evaluating score inflation.

I know of nobody who has contributed more to the research literature on score inflation than Koretz and his coauthors (see the references in Koretz [2010] for examples). Research by Koretz and his colleagues has provided a great deal of solid evidence that score inflation is one of the unintended consequences of TBA. In the paper of focus for these comments, Koretz has provided a framework for understanding how inflation can take place. He has also provided a number of illustrative items that have characteristics which exacerbate the problem of inflation. When items that have predictable patterns are used or items that are clones of previously used items are used, they cannot be expected to measure the same construct that was measured by counterpart items used earlier when the patterns were fresh and there were no clones that had been used in earlier administrations of the test. It follows logically that one way of reducing score inflation is to stop using items that are clones of previously used items and to avoid the use of items that have predictable patterns.

Even if predictable patterns and item clones are eliminated from tests in the future, there will still be a danger of score inflation due to the way state tests are equated (or linked) from one year to the next. As Koretz noted, the common approach to linking from year to year is to use a nonequivalent group with anchor test (NEAT) design. Although the NEAT design has a number of strengths, it is vulnerable to score inflation pressures because it relies on reusing items on the anchor from one year to the next. From an equating perspective, it is desirable to have relatively long anchor tests embedded in the full test to be equated, that is, to use a substantial number of items as the fixed anchor. If instruction is focused on the specific items on the anchor or on clones of the anchor items, however, then the relationship between the anchor and the construct the test was originally measuring is likely to be changed. The NEAT design depends on the assumption that the relationship of the anchor to the construct remains constant from one administration to the next. Violation of this fundamental assumption is apt to lead to inflated estimation of student achievement. Therefore, the goal of minimizing score inflation pushes in the opposite direction from the goal of having a strong anchor by repeating a substantial number of items.

One approach to equating that might reduce score inflation is to use multiple anchor tests, embedded in the operational test, that link back to different years. With two anchors–one that links back to the most recent year and one that links back to two or three years earlier–the linkage pattern may be less obvious and less predictable than that with a single anchor linking back to the most recent year, thereby reducing the degree of score inflation due to equating.

Both logical analyses of item clones or items with predictable characteristics and research evidence lead to the conclusion that there is score inflation. The degree to which score inflation can account for the increases in test scores that have occurred on most state tests over the past several years, however, is more subject to debate. There have been gains on NAEP mathematics assessments at Grades 4 and 8 over the past few years, for a substantial majority of the states (National Center for Education Statistics, 2009.) The fact that the state gains on NAEP are smaller than the gains on the states' own assessments suggests that there is some inflation in the state results on their own assessments (Center on Education Policy, 2009). However, the gains on NAEP tests suggest that the state-specific test gains represent a combination of real increases in student achievement and some inflation in scores. It would be nice to be able to disentangle the effects.

The gains on NAEP reading assessments, unlike the gains on the mathematics assessments, have been tiny to nonexistent. While this result doesn't prove that all the gains on states' own reading assessments are spurious, it does make one wonder if they are due entirely or primarily to score inflation.

Koretz has proposed the use of self-auditing assessments as a means of routinely checking on score inflation and providing a quantitative estimate of the magnitude of those effects. This might be done by embedding (a) items that are clones of previously used items or (b) items with obvious predictable patterns as nonscored assessment items, in the same way that field test items are currently embedded as nonscored items in operational assessments. Comparisons of performance on the self-auditing items to scored items on the operational assessment would then be used to estimate the magnitude of score inflation. The idea of routinely building self-auditing sections into operational assessments is appealing, but it comes at a cost because it requires testing time and item development resources that could be

used for increasing the number of field test items or the number of operational items, and, therefore, the reliability of the assessment.

Koretz has made a valuable contribution by highlighting some of the factors that contribute to score inflation. His framework for considering the causes of score inflation and his suggestions for the creation of self-auditing assessments are worthy of serious consideration and discussion.

References

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*, 67-90.

Center on Education Policy. (2009). *State test score trends through 2007-08, Part 3. Are achievement gaps closing and is achievement rising for all?* Washington, DC: Center on Education Policy.

Koretz, D. (2010). *Some implications of current policy for educational measurement.* Retrieved from http://www.k12center.org/publications.html

Lindquist, E. F. (1951). *Educational measurement*. Washington, DC: American Council on Education.

National Center for Education Statistics. (2009). *The nation's report card. Mathematics 2009* (NCES 2010-451). Washington, DC: Author.