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

Implications of Current Policy for Educational Measurement: Discussion Comments

Drew Gitomer

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.

Implications of Current Policy for Educational Measurement: Discussion Comments

Drew Gitomer

Educational Testing Service, Princeton, New Jersey

This paper is based on a reaction by Drew Gitomer to a presentation 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.

Assessment reform efforts dating back to the 1980s have been grounded in an argument about the consequential implications of the particular design of tests. Echoing throughout the California Assessment Project, New Standards Project, and Vermont Portfolio Project, among many others, has been the argument that traditional tests skew practice. This is said to occur because educators carry out instruction that attempts to emulate both the surface form of the items and the impoverished view of learning and understanding upon which the items are based. Therefore, if educators are going to teach to the test, it makes sense to alter the test fundamentally so that teaching to the test is a worthwhile activity consistent with the more ambitious goals of education.

Over the years, efforts have been made to develop assessment tasks that ask students to engage in more complex reasoning and produce deeper and more integrated responses. These efforts were discontinued for a range of reasons, and almost always the increased cost of human scoring was a major determinant. Today there is heightened concern about the influence of traditional assessment techniques due to the pervasive, and to many perverse, effects on teaching attributed to No Child Left Behind and its accountability requirements. In response, renewed attention is being given to incorporating more ambitious assessment designs. Further, with the potential of automated scoring tools as well as the promise of substantial federal funding, some of the traditional concerns about cost may be reduced.

Dan Koretz puts forth a cogent argument that says yes, we do need to have assessments that more fully represent the range of educational objectives that are valued by the system. But he departs from conventional wisdom by forcefully noting that simply changing the nature of assessment tasks will not halt the undesirable behaviors brought about by tests that sample a limited number of desired outcomes. In fact, he argues that because of the smaller number of assessment tasks engaging students in complex reasoning that can be placed on a given test, and the likelihood that these sorts of tasks are more memorable, the potential is even greater to focus instruction on particular items rather than the underlying constructs they are intended to measure.

Koretz makes a point that ought to be obvious but is often lost in the rhetoric of test-based accountability. Simply, tests have a rightful place as information that supports an accountability system. But the tests, no matter how cleverly created, can never be *the* accountability system.

Tests are a proxy for a subset of the constructs we intend to measure. Once the proxy becomes a surrogate for the entire set of educational objectives, perversions of desired practices in terms of both assessment and instruction are inevitable. Systems will work to optimize scores on the proxy, and this will result in educational distortions of varying severity. Because current accountability systems are defined by test performance almost uniquely, the consequences for instructional practice may be unintended, but they are not unexpected.

Koretz challenges the very structure of current accountability systems and seeks something that admits a fuller range of data and processes designed to ensure the credibility of information produced by the educational systems. I think there are several key points that his presentation raised for me.

The first is the idea of *mediating variables*. An accountability system that does not give proper attention to how results are produced is ripe for distortion. In the business world we have repeatedly seen companies whose bottom line looked encouraging, only to find something amiss. Sometimes this is due to outright corruption, as in the case of Enron or Madoff. Other times the results are honest but fail to reveal some underlying structural problems. For example, quarterly or annual income may appear strong, but only because short-term results are maximized while compromising the long-term future.

In these cases, the bottom-line data would have been much better understood if attention had been given to mediating variables that clarified how income was generated (or not). Perhaps there were shenanigans, or perhaps assets were being sold for current gain but at the expense of the longer term. On the other hand, one has more reason to trust the bottom line if there is credible evidence that sales increased, if expenses were reduced, and if customer satisfaction was higher.

In the same way it seems that an educational accountability system needs to attend to mediating variables as well. If scores dramatically increase in a school, district, or state, is there any rationale that would explain the growth? Are there different teachers or curricula? Has the student body changed? If nothing changed except the test scores (the bottom line), then there is an increased likelihood that the increased achievement was merely a chimera.

Currently a substantial number of research efforts around the country are taking a serious look at potential mediating variables that could ultimately ensure that the system pays attention to a broader range of indicators. Many of these measures are focused on classroom observations of teacher practice. If, in these low-stakes research studies, it is established that certain characteristics of teacher practice are associated with student learning, then it becomes possible to imagine an accountability system that paid attention to both student assessment data and the mediating teaching quality. Other mediating measures that are being studied include instructional logs, student assignments, measures of teacher knowledge, and teacher and student attitudes. Taken together, mediating variables have the potential to ensure that the accountability system can both make better sense of changes in test scores and also provide information that can support improvements in the system. Only by understanding how the mediating variables affect student outcomes can actions be sensibly taken to improve the practices that will improve student achievement. Without that articulation the student outcome data is little more than a black box. It is not surprising then that, with such a paucity of information, educational systems simply resort to practicing test items as the preferred method of improving outcomes.

A second important implication that I take from Koretz's work is that not only are multiple sources of information required, but they ought to be synthesized through processes of disciplined judgment. Educational goals are complex and multidimensional. This means that measuring attainment of these goals cannot be adequately reduced to a single accountability score. Attempts at this kind of reductionism inevitably will heighten the salience of some goals at the expense of others. Composite measures are also likely to be much less meaningful than if the initial contributors to the composite were preserved in any reported metrics.

To provide an example for the sake of argument, imagine an accountability system based on two variables, weighted equally: mathematics achievement and good work habits and attitudes. An algorithmic model would average the two scores on these measures for each group of students into a composite. Therefore a group of students whose scores were high on mathematics and low on work habits would have the same composite as another group of students who were average on both. Whether from an accountability perspective or an instructional one, the implications for what the results mean and what to do about them are quite different; they also are likely to be much more valid if the conceptually unique indicators are kept separate. Synthesizing the data through judgment rather than an algorithm will be much more illuminating.

This implies, however, that judgment processes are not random or idiosyncratic. Algorithms are often preferred precisely because they do not require judgment, and that is because policymakers often distrust the judgments that have been made and publicly communicated by educational institutions. And certainly there is good reason to distrust the quality of information that is presented by the educational system as some type of public accounting.

In order for a system that employs judgments to have some credibility, judgments must be disciplined and open to inspection. Koretz recommends the use of auditing processes, which also challenges some current assumptions about accountability. In this model, judgments are made at the local rather than state level. However, the role of the state (or other large entity) is to provide checks on the quality of local decision making in order to ensure credibility. Ultimately, if the system is well executed the local decision makers play a much larger role in the accountability process, and hence they are less apt to distance themselves and feel victimized by state-imposed processes.

Taken together, I think that this work provokes us to think about an accountability system that varies, at minimum, in four key ways from most statutory accountability systems. First, the tests ought to be more ambitious and representative of a full range of student outcomes. Second, the system needs to pay attention to mediating variables to provide confirmation of any changes in achievement, but also to illuminate where teachers, schools, and districts might focus their efforts to improve student performance. A third implication is for the inclusion of a broader range of measures. Finally, judgmental processes, including external audits of those judgments, need to be brought to bear to integrate these multiple sources of information.

Koretz is obviously highly doubtful of the efficacy of current accountability practices. Some individuals are more positively disposed, particularly because there is evidence that scores have increased for the lowest achieving students, particularly in the earlier years of education. Even if these data are to be believed, there is scant evidence that accountability practices have increased achievement across the

board, and particularly for higher achieving and older students. Thus at best, current accountability practices may be best at supporting the more routine, procedural thinking that will raise scores for the youngest and least skilled students. If we are, however, going to have an accountability system designed to promote excellence, models more in line with Koretz's thinking must be considered.

It is difficult to think of any other aspects in our lives in which complex systems are accounted for in such a reductionist manner. Whether it is admission to school, a job evaluation, judging the quality of a restaurant, or evaluating the performance of a company's stock, judgments are complex, multifaceted, and attendant to mediating factors. We make complex judgments on a broad set of information and deem these judgments to be fair and credible because, in part, they take into account mediating variables. Students and schools deserve that same deliberation. There is much work to do, however, to make such a system viable. It cannot be done by declaration alone.