Assessment and Data Quality Issues

Written Testimony Submitted to the Commission on No Child Left Behind in Preparation for a Public Hearing in West Hartford, Connecticut on May 9, 2006

Submitted by:

Stuart R. Kahl, Ph.D.
President and CEO
Measured Progress, Inc.
100 Education Way
Dover, NH 03820

Assessment and Data Quality Issues

This testimony is offered from the perspective of a testing company specializing in customized statewide assessment, holding general or alternate assessment contracts with approximately twenty states for all or some components of the states' high stakes, accountability assessment programs. The commentary has implications for the reauthorization of Title 1 legislation and for implementation (enforcement and guidance).

Industry Capacity

With the high stakes associated with the results of statewide testing, errors in the computation or reporting of results have received a great deal of attention in the media and have become the cause of concern. At a recent (April 25, 2006) meeting between Secretary of Education Margaret Spellings and executives from many testing companies, it was suggested that the error rate in the testing industry may be considerably smaller than that in other industries. However, while it was acknowledged that errors can probably never be totally avoided, the incidence of errors has been at a level that cannot be accepted given the consequences of the errors for individual students and schools.

The logical question being raised in many circles is, "Does the testing industry have the capacity to handle the testing required by NCLB?" There are now ten or twelve companies capable of and currently operating statewide educational assessment programs. With fifty states, it would seem that capacity should not be an issue. There

may be a few small states who are concerned about getting companies to respond to their requests for proposals (RFPs); however, RFPs generally do elicit several proposals. In the rare situation in which one does not, there are usually reasons beyond contract scope, that make the project unappealing, and it is up to the state to deal with them.

At Secretary Spellings's meeting, one major testing company's executive stated that his company has the capacity to handle several times the volume of testing they are currently handling. Representatives of other companies felt the same. If this is the case and capacity is not an issue, then it is important to better understand the actual nature of the problems the testing companies are encountering. The remainder of this section is intended to facilitate that understanding and to suggest possible ways to address various problems, whether through changes in legislation or implementation.

Systems. A major reason testing companies believe they have the capacity to handle the volume of testing required by NCLB is that most, if not all, of them have developed efficient, high-tech systems to apply during the various stages of an assessment program. These systems have not only increased the level of automation applied to various processes such as shipping, log-in, materials tracking, analysis and reporting, they also perform these processes far better than in the past. For example, bar-code technology allows some companies to not only know that a secure test booklet has not been returned, but also to know which student's booklet it was. Image scoring of constructed responses results in better quality control with respect to the work of human "readers" and faster scoring of written student responses because it enables ready access

to scorers anywhere in the country. Psychometric processes of scaling and test equating now take hours, rather than days and weeks. Thus, because of these systems, testing companies have indeed increased capacity dramatically. The large numbers of students and different tests involved in testing programs, in and of themselves, are not the cause of the difficulties companies encounter. One company official has suggested that if his company had only one state assessment contract, rather than many, yet still had the same size staff, the company could still not produce results any faster and the timelines for reporting would still be challenging – the latter a major factor increasing the risk of errors.

Special demands of statewide accountability assessment. The yardstick many have mistakenly used for turnaround times for test results has been the turnaround times historically associated with local district testing using off-the-shelf instruments. Those instruments, years in the making and reused for many years, had pre-established "look-up" tables associated with them so that reporting could be accomplished very quickly once points earned were determined for students. More importantly, school and district results were based on whatever students' answer sheets were returned to the testing company. The company had no idea whether all or just some of the students in a particular grade were tested, and had no responsibility to find out. In fact no one was accountable for including all students.

For today's high stakes, accountability testing programs, all students in tested grades must be accounted for and their reported assignment to not only schools, but also to

subgroups within schools must be accurate. Furthermore, because new tests are often used every year, in part for security reasons, significant psychometric work (scaling and equating) must be accomplished each year, and final analysis and reporting cannot be completed until the data files are “clean” and complete.

Data quality and business reality. The challenges of statewide accountability testing identified above were raised at a meeting of testing company executives with a former Secretary of Education in February of 2003. The obvious solution too many of the challenges suggested at that time was that the states and their contractors had to get a “head start” on the data files before the student response data became available. There is no question this is necessary, and many states have implemented student information systems for that reason. However, these systems are not “mature.” Typically the data files provided to the contractors are incomplete and contain inaccurate information. As any experienced researcher or testing person knows, far more time is spent cleaning data files than running final analyses. Reconciling returned materials with enrollment information in existing data files can take weeks, even months. Adding to this challenge is the fact that even with established due dates for the return of testing materials to the contractors for processing, many materials “straggle in” for a variety of reasons. Additionally, they are not always returned in the manner requested in direction manuals. Some student response documents are inevitably left inserted in test booklets, rather than included in the separate envelopes or cartons designated for the return of response documents. Accountability for this sort of occurrence is generally unclear. Of course, schedules are

created under the assumption that there are no problems with data files from the student information systems or irregularities in the return of materials.

Guidance and regulation. The U.S. Department of Education recently issued "Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting" (April 2006), as guidelines for states, LEAs, and schools. While many of the operational procedures identified in this document are already being followed by most testing companies (and in fact, many have "gone beyond"), there is information in the guidance on the needs of states and their districts and schools regarding data management. This is clearly an area in which additional funding is needed for staff, staff training, and guidance. Many states and LEAs currently do not have the capacity to create and maintain student information systems in such a way that data clean-up required of testing contractors can be minimized.

Because of the high visibility of testing errors of late, regulation of the testing industry is sometimes proposed as a solution to the problem. Regulation would not in any way reduce the risk of errors. Regulations would lead to the review of companies' operational procedures, technical approaches, quality control procedures, etc., which generally are already very good. The systems described earlier are impressive. Psychometric techniques are head and shoulders above what they may have been years ago, and most states have technical advisory committees (TACs) to which some of the leading psychometric experts in the country belong. Furthermore, the testing community takes very seriously the "Standards for Educational and Psychological Testing" produced

jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education in 1999.

Choice and Timing. One of the guidelines listed in the April 2006 non-regulatory guidance from the U.S. Department of Education states, "Reserve extensive time in the schedule exclusively for assessment data collection and validation, as soon as possible after the assessments are administered." NCLB's requirement that parents of students in schools identified as "in need of improvement" (for not meeting percent proficient targets two years in a row), be given the option of sending their children to other schools is what drives reporting schedules and does not allow time for a significant results/data validation effort. States tend to prefer testing in the late spring so that their tests can address their grade level expectations (GLEs) completely. While they are free to establish schedules calling for testing earlier in the year, this could entail the determination and possibly mandating of when specific topics are covered in every school.

So that parents can make their decisions about school choice when they must be made, AYP results must be reported in mid summer. A more reasonable approach to reporting, allowing a validation period, would give schools individual student results fairly quickly, but allow two months from the time individual student results are provided until school level AYP results are produced. The individual student results could be used by the school for various purposes, but that two month period could be used to verify that each school has results for all its students, that the various subgroups to which students are assigned are accurate, and that there are no irregularities in each student's response data.

Also during that period, corrective action can be taken where discrepancies and irregularities exist. In short, this change would allow everyone to have more confidence in the accuracy of the reported results.

To allow such a data validation period would require a very different approach to the school choice provision of NCLB. Interestingly, statistics from the Center on Education Policy indicate that few parents take advantage of the choice option. Truthfully, they do not have time to make informed decisions, nor are changes in schools real possibilities in many cases. Many educators are calling for "services before choice." Delaying the choice option until the next year would allow the schools time to address issues the parents might have regarding the education of their children, to provide appropriate services to students in need, and to convince the parents that their concerns have been addressed effectively. The risk, of course, is that the results of the next round of testing could take the school out of the "in need of improvement" category not very long after the option of choice was extended to parents. Nevertheless, this change would allow time for both the critical data validation step and "services before choice," as well as much more time for thoughtful deliberation and planning following the release of results that, by law, might create the choice option.

Performance Targets and Growth Models

There are those who question how realistic NCLB annual targets for percent of proficient students are. With those targets on a track for 100 percent proficient students by 2014,

many expect that ultimately the vast majority of schools will be considered “in need of improvement.” Assuming a low likelihood of any change in that ultimate requirement of NCLB, the acceptance of the use of growth models in AYP determinations seems a viable approach to rewarding schools that are successful in moving students who are initially quite far from proficient to points closer and closer to proficient over time. The simplest growth model involves establishing, for students scoring far below proficient on the state tests, interim target scores leading them to the proficient level in two, three, or four years, depending on how far they were from proficiency initially. The percentage of students meeting their interim targets would be added to the percentage actually achieving at the proficient level to determine the critical percentage that is to be compared to the annual target percentage for AYP.

From a testing company perspective, there are no reasons to avoid applying a growth model, and in fact, depending on the model chosen, the growth targets and progress toward them can be easily determined provided a good student information system exists to allow the tracking of students over time. Thus, an additional recommendation for the Title I reauthorization is to allow the use of growth models in AYP determinations and to provide guidance in selecting and implementing them.

Test Characteristics

NCLB called for the use of “multiple measures” in the assessment programs, yielding results used for AYP determinations. Multiple measures historically referred to different

tests or different types of tests. An argument for multiple measures in the form of different tests is that decisions (e.g., proficiency level of a student) based on multiple tests administered at different times would be more sound than decisions based on a single test. Some states have interpreted the term "multiple measures" to mean tests with multiple item formats (e.g., multiple-choice and constructed-response). Some states have chosen to use only multiple-choice items in their high-stakes tests to reduce costs and turnaround time for results. Research in the early 1990s showed that sole reliance on multiple-choice items in a high stakes environment can have a negative impact on instructional programs, and therefore, many educators are concerned that the demands of NCLB have led states to implement assessment programs with negative instructional impact. The Department's April 2006 guidelines state, "Data quality is more easily assessed for multiple-choice, directed response instruments than for constructed-response or essay instruments." This is not true. For example, common statistical analysis packages generate the same reliability coefficients for tests whether they include dichotomously scored multiple-choice items or constructed-response items scored on a zero-to-four continuum.

The guidelines further indicate that states should "ensure that assessment instruments clearly associate each item with one state academic content standard." While this guidance may make life easier for those doing alignment studies relating items to content standards, following it would also make it more difficult to measure higher order thinking skills or higher levels of what is called "depth of knowledge." Higher order skills require students to make connections across content strands and disciplines, to make

comparisons, and to synthesize and evaluate diverse information. Not surprisingly, alignment studies are finding too few higher level questions in state tests.

It is suggested that in the reauthorization process, attention be given to the clarification of the intent of the language around multiple measures. Furthermore, it is important that guidelines not inhibit the use of constructed-response, higher order, or otherwise complex items so that the higher expectations that are at the heart of NCLB can be encouraged, assessed, and ultimately realized.

Summary of Recommendations

- provide additional funding for staffing, staff training, and guidance related to data management to assure the effective development and maintenance of accurate, complete, and up-to-date student information systems;
- delay school choice decisions by parents to allow for a much longer period for data verification and for parents to have time to make more informed decisions;
- allow growth models to be applied in the determination of AYP results so that annual growth targets are more realistic;
- clarify the intent of language in NCLB regarding multiple measures and avoid guidance that inhibits the use of higher order questions in the assessment instruments.