FairTest

National Center for Fair & Open Testing

Race to the Top Assessment Program
Public & Expert Input Meeting
Boston – General Assessment Meeting
November 12, 2009

Public Comment
By Monty Neill, Ed.D.
Interim Executive Director

General Remarks

Before the nation can successfully implement better assessment practices, it must first reject the incorrect assumptions and flawed logic of No Child Left Behind (NCLB). To ensure effective education reform, including high-quality assessment, the Administration must overhaul NCLB, its draft requirements for Race to the Top (RTTT), and the "Assessment Program Design" to which we are responding today.

NCLB has failed to improve educational quality and equity. U.S. children have made less academic progress since NCLB came into effect than in the preceding period, and the achievement gap has not narrowed as significantly. Secretary Duncan's proposals to date would reinforce the errors of NCLB.

The problem is not only that tests used under NCLB are inadequate, but that the fundamental assumption behind the law has proven wrong: America cannot test and punish its way to better schools, no matter how good its standardized tests might become. That said, the nation does need high quality assessments that are properly used.

A revised RTTT could provide a great stimulus for states to overhaul their assessments. This would require developing new systems of local and state formative and summative assessments that can assist student learning, help gauge students' academic progress, and provide an important source of evidence for evaluating teachers, principals and schools. These new systems should be built within a framework that provides flexibility and diversity while ensuring high quality opportunity and expectations for all students.

15 Court Square, Suite 820, Boston, MA 02108 (857) 350-8207 FAX (857) 350-8209
Web Site: www.fairtest.org Email: fairtest@fairtest.org

Unfortunately, the framework before us today appears designed to ensure the continuation of highly centralized, top-down state assessment systems. To a great extent, it perpetuates the flawed conceptions of NCLB. It is far too limited and will inhibit the most necessary and valuable improvements in assessment. Its structure reduces teachers to administering and perhaps scoring tests. It does not even suggest that teachers should be part of the process of creating new, high-quality assessments. It completely misconstrues formative assessments, as if the issue were teachers selecting a test off a shelf instead of responding to the emerging needs of highly diverse learners engaged with a specific curriculum.

Therefore, FairTest's first specific recommendation is that the "Assessment Program Design" itself be overhauled. Fortunately, there exist well-thought-out approaches that can provide a new framework. Among your expert presenters are people who have helped develop such approaches. I urge the Department to listen carefully to their recommendations and concerns.

The Forum on Educational Accountability (FEA), an alliance of dozens of education, civil rights, religious, disability, parent and civic organizations that I chair, commissioned an Expert Panel on Assessment, which included some of your expert presenters, to develop recommendations on what a comprehensive, educationally beneficial assessment system could look like.

The report explains how to use multiple sources of evidence -- teacher evaluations of student work over time, locally developed assessments, performance assessments of various kinds, and statewide standardized exams -- to determine both achievement levels and student growth. It recommends external monitoring to assure the quality, accuracy and fairness of the various assessments. A system built from these elements would provide solid data for evaluating schools, districts and states. A growing body of evidence from the U.S. and other nations supports these recommendations.

Assessment is both a quantitative and qualitative endeavor. Thus, states should be able to use these federal funds to engage in qualitative evaluation, such as an inspection system, as recommended by the Broader, Bolder Agenda. Inspectors are trained experts who visit schools to observe, review data, and hold discussions, then evaluate the school and issue a report. This process is central to accountability in England and New Zealand.

Legislation introduced in Massachusetts and supported by FairTest would build a system that includes state standardized test results, incorporates an inspectorate, and relies most heavily on assessment of student classroom work. This legislation provides the three "legs" on which new assessment systems should stand.

In my written comments, I propose concrete steps the Department should support. These are based in large part on three attached documents - FEA, BBA and the Massachusetts bill. Some of these ideas can be incorporated into the "Assessment Program Design" before us, but many require or could be done far better and more easily if the Department significantly modifies the program design.

Responses to specific questions

In hopes that the U.S. Department of Education both revises the "Design" structure and agrees to positive steps that can provide major improvements to assessment, I provide detailed answers to questions posed in the Federal Register notice.

I highly recommend that Department staff first read the attached documents from the Expert Panel on Assessment of the Forum on Educational Accountability; the Broader, Bolder Agenda; and the Massachusetts legislation. Much of what I discuss below assumes content detailed in those documents.

1. Propose an assessment system (Questions 1 and 2)

Putting the FEA, BBA and the Massachusetts proposals together, states and districts, with federal assistance, can build educationally sound assessment systems. They would provide series of assessments that use different formats, rely on classroom-based evidence (summed up in grades and scores, portfolios or learning records), extensively utilize performance assessments, and include both quantitative and qualitative evidence, each contributing to fair evaluations of students, educators, schools, districts and states. They can provide data on status, growth and improvement (see FEA, 2007, Principle IV, for discussion of these three dimensions). They provide a framework for professional development through which teachers can greatly strengthen their formative and summative assessment knowledge.

States can collaborate in multiple ways to build their new assessment systems, including:

- develop new large-scale common assessments;
- build multi-state banks of performance tasks and projects teachers can access as appropriate;
- work together to solve complex problems of assessing highly diverse populations of students, in and across demographic categories of race, class, disability and language;
- exchange ideas and experiences on the best ways to use multiple sources of evidence, including verification of the accuracy of teacher evaluations of student learning;
- share evidence of what works, with whom, in what circumstances, with what mix of local control and state guidance, in assessment, professional development and school improvement.

Using a mix of state and local assessments, including classroom-based evidence, is permissible under NCLB, provided that the evidence is reliable, can support valid inferences, and ensures comparable definitions of basic, proficient and advanced achievement across schools and districts. What is needed is not authority, but funds and federal support to enable states to build such systems.

What follows is a discussion of the use of performance tasks, other sources of classroom-based evidence of learning, and local assessments, that with large-scale exams (census or sampling) and inspectorates can create a strong yet flexible assessment and evaluation system based on state or common standards. The proposals from the Department should reflect the goal of helping

states develop comprehensive assessment and evaluation systems, not simply a new set of large-scale exams.

The National Research Council Board on Testing and Assessment (2009) recently reiterated the wisdom of the measurement profession that standardized tests alone cannot provide an adequate basis for making high-stakes decisions about students, educators, schools or systems. This point has been a basic premise of the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999). Thus, proper assessment requires multiple sources of evidence. Further, the extensive use of performance tasks is necessary for evaluation of complex learning (Sheperd, Hannaway and Baker, 2009).

To obtain multiple sources of evidence and use a wide range of performance tasks, states could mandate more and more standardized assessments, using varied formats and types of items. That would be a bad idea, costing far too much and pushing the system to the point of overload and explosion. For example, a performance assessment system for a biology course would require about 10 performance tasks of an hour each to be able to make a defensible judgment about a student (Shavelson, Gao and Baxter, 1993). That is not feasible for a centralized system, but it is feasible for a good teacher to use 10 such tasks a year.

While the teacher should be capable of designing some of those good tasks – in part to be able to teach well – it is infeasible for her to develop all 10 of them. Thus, there is a need to develop assessment banks, to which teachers would contribute as well as draw on existing tasks to use when appropriate to her particular curriculum and students, including ELLs and SWDs. The results of those tasks would provide a significant component of the composite evidence of each student's learning, feeding into the overall data system.

To keep costs low or avoid use of locally-controlled (though state-guided) assessment evidence, a state might choose to use only one or two tasks in each of its large-scale exams. It would then have to continue to rely primarily on multiple-choice and short-answer items. That would effectively prevent the state from gathering evidence of higher-level student learning, for which those item types are inadequate. Worse, especially if the tests remain high-stakes, it would perpetuate the current problem in which too-limited tests dominate curriculum and instruction. We conclude that the far preferable solution would be to encourage extensive use of high-quality performance tasks, with teachers doing the grading.

Our hypothetical biology teacher would teach more than the performance tasks, requiring further assessment evidence, such as quizzes, tests, reports, experiments, and more. Some of them would involve work on computers as new technologies develop.

Thus, the core assessment data for evaluation purposes should come from ongoing student work (classroom-based evidence). Students produce great amounts of work every year, which their teachers evaluate. Taken together, those teacher evaluations are better predictors of college success than are even very technically sound single tests, such as the SAT and ACT (College Board, 2009. Still, the quality of current teacher assessment knowledge and practice needs improvement. The best solution to this problem is professional development, which is generally essential to the development and success of the new system (more on that below).

In addition to using performance tasks and conducting other teacher made or selected assessments, schools and districts can use varied types of common assessments to gather additional information, including for use as a check on teacher evaluations. These assessments can be part of each student's composite score.

The result is in effect a grade, a summative judgment reduced to a number. The grade or score would be based on the applicable content and performance standards. It would be a grade rooted in richer evidence than most teachers' grades now are because of the use of approved performance tasks, a grade rooted in greater teacher skill due to extensive professional development. Thus, the classroom-based evidence, which would incorporate assessments generated from beyond the classroom, would provide a high-quality basis for evaluating student progress. It can provide one important piece of evidence for evaluating teachers and principals, schools and districts, as well as states.

Such classroom-based and local assessment data requires verification to ensure the assessments and the scoring meet state and federal standards. There are several ways this can be done.

"Moderation," re-scoring samples of student work, is one option (Wood, Darling-Hammond, Neill and Roschewski, 2007). This assumes either an organized compilation of student work (a portfolio or "Learning Record"), or at a minimum some of the performance tasks drawn from the "bank." This process works best if the design of the portfolio is very strong, and if teachers have a few years of practice (c.f., Learning Record, n.d.).

Moderation enables independent readers to rescore work from samples of students in each classroom, thereby checking on the originating teacher. This can provide a means to ensure that performance standards are applied across all schools and students are evaluated based on common standards. In essence, if reviewers conclude the five randomly selected work samples from Ms. Jones' classroom have been scored accurately, it is a reasonable conclusion that her other students have also been accurately scored. Other nations use moderation successfully (Wood, et al., 2007).

The toughest question may well be what to do when reviewers disagree with the originating teacher. One option is to produce new scores that would replace the original scores for use in final judgments on student work, as is done in Queensland, Australia. Another is to simply provide feedback, as the Learning Record did. The latter would keep the stakes low as the system develops over time until such discrepancies are very rare. It likely will take three to five years to build a system with the needed accuracy, due to the need for rigorous development and extensive professional development.

Another tool is to carefully design requirements local assessments used for accountability would have to meet, and then review the assessments for quality (Wood, et al.). This was the approach Nebraska used in developing a system of local assessments. Over just a few years, the quality of local assessments improved significantly (Gallagher, 2007). A third option is to triangulate with other forms of data, such as the large-scale state exams (which may include some performance

tasks). In this, discrepancies between test scores and local results can be investigated and resolved.

All three can be used together to verify each of the components of information that contribute toward the evaluation. By employing all three, strong design can be built in from the start, while moderation and triangulation ensure comparability and maintenance of standards.

In such a system, large-scale assessments need not be given to every student every year, because locally-based evidence would be available to be gathered, analyzed and evaluated every year, in all subjects (Joint Statement, 2004). In addition, this approach would allow for evaluation of all subject areas without the burden of large-scale exams in every subject area.

States could use sampling systems such as employed by the National Assessment of Educational Progress (NAEP) and by the now-ended Maryland State Performance Assessment Program, a set of performance tasks that was ended when NCLB required every student in each grade to obtain comparable scores. Local assessment evidence could combine with sampling tests to produce the needed annual information for each student in each grade. That is, students would receive grades that are validated through the moderation process. This does leave some leeway, but with proper moderation and triangulation for the years tests are administered, the leeway is quite small. That small leeway is a price well worth paying for lower state test burden. However, a state could continue to administer large-scale exams that produce individual scores to all students in selected grades. Many details regarding these possibilities are provided in the FEA Expert Panel (2007) report.

Thus, "innovative and effective approaches to assessment" rely on multiple sources of evidence which, to be used effectively and with reasonable cost and administrative burden, must employ significant local and classroom-based evidence accumulated over the course of the year. Even with assessment tasks provided by computer (as some can be, but others cannot), the choice of which ones to use at what time must be locally determined unless a state is to insist on a state-mandated curriculum in which all students will proceed through identical curriculum at the same pace. I assume that is not a goal of this Department.

The rich evidence flowing from such a system can be used to strengthen teaching, learning and program improvement. It can provide an important contribution to the determination of school effectiveness (which also requires a great deal of other evidence, from within and outside of schools; see Forum on Educational Accountability, 2009). It can contribute to teacher and principal evaluations, though student achievement, even if determined using multiple sources of evidence, must only be one part of those evaluations.

I noted in my introduction (oral testimony) the use of inspectorates. These are major components of accountability in England and New Zealand (Rothstein, 2009). They have been used occasionally in the U.S. Essentially, an inspection system has teams of trained experts conduct multi-day visits to each school to observe, review data (including school self-evaluations), and hold meetings and discussions. They then issue a report on the schools. They may be involved in providing assistance, as needed, or that work may be done separately. Rothstein proposes inspections once every three years; they are less frequent in other nations. They resemble current

accreditation processes, but as Rothstein notes, those process would have to be improved. The Broader, Bolder Agenda in Education Campaign (2007) has proposed that inspectorates be part of a reauthorized ESEA.

Legislation proposed in Massachusetts (Sciortino, 2009) would create an inspectorate, but it would also continue state large-scale exams, and it would rely on locally-based evidence of student learning, including classroom work, as described above. Education, civil rights, parent and other groups helped develop this plan.

In response to a few additional system design requirements:

- Standardized tests do not provide much evidence of college readiness. For example, though there is a correlation between ACT scores and college success, many students who score below ACT's cut-off in fact succeed in college, while many above the cut do not. SAT and ACT scores predict only about 16% of the variance in college grades, less than what teacher grades provide, particularly grades in college preparatory classes (College Board). From the other end, the knowledge and skills sought by college professors and higher-level employers cannot be measured by current or improved tests one-shot standardized exams (Achieve, 2005). It is the accumulation of a rich array of evidence, primarily classroom-based, that will enable fair and accurate decisions; and it is a rich educational program supported, not undermined, by assessments, that will ensure student success.

- High quality performance tasks also provide multiple entry points so students of different abilities and knowledge can access them. Their ability to provide multiple entry points should be one means by which they are evaluated for inclusion in an assessment bank. The Expert Panel on Assessment (2007) also called for using universal design principles to ensure assessments are accessible and valid for students with disabilities and English language learners (see section III).

- Local teachers would do most of the scoring, of their own students and as part of teams working at the school or district levels to score local and state assessments. As discussed above, moderation to evaluate teacher accuracy should be part of the system. This can be done fairly quickly, at the end of a school year for example.

- The costs should be feasible, but we lack precise evidence, in part because there are many design options. I did some rough calculations of the per-student costs of the Queensland, Australia "Rich Tasks," sets of performance tasks at grades 3, 6 and 9 that were scored locally then subject to moderation. (I could not obtain direct cost statements.) My estimate is $25-35 (Australian) per student annually, including design, moderation and reporting, but not including teacher classroom time. When part of teachers' regular work, group scoring of complex tasks also serves as valuable professional development.

- The Department expressed particular interest in assessing individual student growth. The FEA Expert Panel (2007) concluded that multiple sources of evidence can and should be used in determining individual student status, improvement by groups of students, and growth of individual students (See esp Section IV). Using multiple sources of evidence is the means to ensure accurate, valid assessments of the very diverse populations in our schools, while

obtaining a sufficient variety of kinds of evidence will require reliance on classroom-based evidence.

- The Department's design framework on professional development reads as being done to teachers (e.g., "delivering high-quality professional development"). Accumulating evidence shows professional development is best done with, not to, teachers, and is a continuing practice, not a deliverable. Similarly, it treats teachers as consumers of assessments, not as designers of assessments. Participating in the design of assessment provides excellent professional development opportunities for teachers, strengthening their knowledge of curriculum content and assessment and providing opportunities for teachers to share knowledge and skills.

- Assessment for learning (formative assessment) has been recognized as a very valuable component of teaching and learning. In the opening portion of my testimony, I criticized the conception of "formative" assessment deployed in this program design. Formative assessment is a process, using a variety of tools, employed by teachers and students. Its use should be rooted in a teacher's particular curriculum and instruction (see Brookhart, 2009, including the appendix "Position Paper"; and Shepard, 2009). Unless the Department is encouraging states to implement a standardized curriculum, not just common standards, it should not focus efforts on constructing pre-fabricated assessments to be used "formatively." The far more important task is to provide funding for professional development in formative assessment. That said, properly designed and useful tasks and projects can provide opportunities for formative assessment. For example, an extended performance task should provide opportunities for feedback by the teacher and reflection by the student.

- Large-scale assessment items should be released, as is now done in Massachusetts and several other states.

Question 3: Local Education Authority expenditures. Unfortunately, the Department's Design views school districts (to say nothing of schools and teachers) merely as implementers of the statewide assessments. A better framework sees teachers, schools, and districts as partners with the states in developing a new assessment system, as outlined above and in the attached documents. They would also contribute to building data systems to use the information from the multiple sources of evidence of student learning. In this approach, LEAs would spend a significant portion of their share of the funds on professional development so that teachers can greatly improve their assessment capacity.

Question 4, I addressed above: Teachers, as part of their paid work, would score anything that is not machine-scorable, either in their schools or in multi-school settings. This can be done quickly, facilitating fast turnaround and reasonable costs. More time-consuming would be a moderation process. Assuming the primary purposes of moderation are to provide feedback for teacher learning and to improve data quality for public reporting and accountability, moderation can be organized for the close of the school year, providing sufficient time to use the information in planning improvements to instruction and curriculum.

In conclusion, the Department must overhaul its proposed Design. The Design is too limited and will inhibit the most necessary and valuable improvements in assessment. It perpetuates a top-

down and centralized approach to assessment and improvement that continues to marginalize educators. The proposal here, including in the attachments, not only provides a different, positive framework, it begins the process of filling in many of the details states will need to individually and collaboratively create and implement new assessment systems. What is essential is to show reasonable options for ways states can design systems, not to proscribe one method all states must follow. The technical complexities will vary across different options. They can be addressed and solved with political will and adequate funding.

Thank you for your consideration.

References

Achieve. 2005. Rising to the Challenge: Are High School Graduates Prepared for College and Work? Available at http://www.achieve.org/node/548.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Broader, Bolder Approach to Education. 2009. *School Accountability*. Available at http://www.boldapproach.org/report_20090625.html.

Brookhart, S. 2009. Editorial. *Educational Measurement: Issues and Practice*. Fall, V. 28, N. 3, pp 1-4. The Appendix is on pp. 3-4.

College Board. 2009. SAT Program Handbook 2009-2010. New York, Author.

Expert Panel on Assessment. 2007. *Assessment and Accountability for Improving Schools and Learning*. Forum on Educational Accountability. Available at http://www.edaccountability.org/reports.html.

Forum on Educational Accountability. 2009. *Empowering Schools and Improving Learning.* Author. Available at http://www.edaccountability.org.

Gallagher. C. 2007. *Reclaiming Assessment: A Better Alternative to the Accountability Agenda*. Portsmouth, NH: Heinemann.

Joint Statement on No Child Left Behind Act. 2004. Forum on Educational Accountability. Available at http://www.edaccountability.org.

Learning Record. (n.d.). Materials about the Learning Record can be found on the web at http://www.fairtest.org/learning-record.

National Research Council Board on Testing and Assessment. 2009. Letter Report to the U.S. Department of Education on the Race to the Top Fund. Available at http://www.nap.edu/catalog/12780.html.

Rothstein, R. 2009. *Grading Education: Getting Accountablity Right*. Washington, D.C.: Economic Policy Institute, and New York: Teachers College Press.

Sciortino, C. 2009. An Act to improve assessment and accountability to ensure students acquire 21st century skills. Massachusetts House of Representatives, H. 3660. Available at http://www.mass.gov/legis/bills/house/186/ht03/ht03660.htm.

Shavelson, R.J., Gao, X, and Baxter, G.P. (1993, March.) Sampling variability of performance assessments. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: CSE Technical Report 361.

Shepard, L. 2009. Commentary: Evaluating the Validity of Formative and Interim Assessments. *Educational Measurement: Issues and Practice*. Fall, V. 28, N. 3, pp 32-37.

Shepard, L., Hannaway , J.., and Baker, E. 1999. Standards, Assessment and Accountability. An Education Policy White Paper. Washington, DC: National Academy of Education.

Wood, G., Darling-Hammond, L., Neill, M., and Roschewski, P., 2007. Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills. Available at http://www.fairtest.org/refocusing-accountability.