Testimony
Senate Standing Committee on Education Public Hearing
The Regents Reform Agenda: "Assessing" Our Progress
October 29, 2013
Rosalie Friend, Ph.D.
Information Coordinator, NYC
Save Our Schools

The education reform policies adopted by the New York State Regents derive primarily from business interests and foundations created by businessmen. They know little about child development or how children learn. Instead they base their suggestions on their experiences in business. The differences between businesses and public schools are vast and deep. Public schools were established as a public service. Children and their families are not customers. School attendance is compulsory; our society requires children be educated to prepare them to be active citizens and productive members of an increasingly complex society.

Businessmen should realize that children are not products. Educated citizens cannot be manufactured. Children have their own minds and their own goals. They are all different in temperament, motivation, background, and how quickly they learn, so teaching groups of children is very complex. The teacher must work with a variety of children at the same time and motivate them to engage in the activities decreed by the school system. Teaching is a profession, not a business.

Furthermore, there has not been much consensus about what type of preparation our society wants children to have. Traditional schools have emphasized obedience to authority, while progressive schools have emphasized teaching children to take initiative. We have seen decades of acrimonious debates over the most desirable kind of education. Now we see an even more complex iteration of this tension. The Common Core State Standards (CCSS) decree an emphasis on critical thinking, analysis, and student writing. The CCSS specifies the importance of using performance assessment in which children create projects or reports with depth and individuality. Despite this stated goal, the two consortia developing tests for the Common Core are both using multiple choice questions, which can have only one answer. Of course the big issues facing our society do not have one right answer. Real problems need the deep thinking which was proposed as the goal of the CCSS. Making multiple choice questions difficult and calling them "performance assessments" is misleading. These questions still have only one right answer; they are not calling on children to think through problems and create solutions.

The consortia do intend to include essay questions in their assessments, but to use computers to grade the essays. I doubt that computers will be able to analyze coherence or rhetoric, let alone originality. Some of the rubrics (guidelines) now issued to teachers scoring essays have been almost mechanized, with instructions to use intact phrases from the texts that students are responding to. Of course this is plagiarism, which is at odds with our society's standards for writing. My own research ties in with others who find that if students are to learn information in such a way as to be able to use it, they must restate the information in their own words.

In light of these observations, the regents should examine the premises of the reforms that are being introduced and determine whether there is any research to support their use before imposing them on New York State.

I am especially concerned with the weight that school reforms have given to testing. I am an educational psychologist, so I know what it takes to develop valid reliable assessments of school achievement. I taught assessment at the Hunter College School of Education. I agree with the reformers' interest in

measuring outcomes, but I fear that the reformers know very little about how we measure learning. We cannot look directly into children's minds to find out what they have learned. We can observe their responses in class, evaluate their work on assignments, and see what kind of contributions they make to classroom discussions. We can also develop tests, compilations of tasks that require skills and knowledge, and see how well children carry them out. Examination of children's work on class tests gives teachers the opportunity to determine which aspects of lessons children have mastered and which aspects need more work. To test school achievement, the items must be based on information and skills taught in school.

Standardized tests were developed to compare children across schools. They must be administered under the same conditions whenever they are used. They are limited by the fact that they are not closely related to the lessons in any particular school. The use of standardized tests for accountability is very appealing and is now established in the United States. The scores are quantitative and seem objective. However these tests lack an essential requirement. There is no standard unit of measurement. When we measure height or weight, the inches or pounds meet a preset standard. There is no standard unit of measurement for learning. Professionals can use standardized tests to adjust instruction, "It seems that Johnny needs more work on trigonometry." However, the scores cannot be used to set policy for school systems. Without a standard unit of measurements the differences in scores are not real.

No standard unit of measurement! This is one reason that testing programs can be easily manipulated. Mayor Bloomberg was able to run for reelection bragging about rising test scores in the schools. The rising scores were achieved by commissioning tests that got easier every year. In 2013 NY State achievement tests were designed to be so difficult that only 30 percent of the children passed. If these tests are allowed to stand as a baseline, there will be claims that students are learning more in future years, as the tests are made easier, because there is no standard unit of measurement.

The first requirement for a test is validity; that means we must be sure that our interpretation of test results is appropriate and meaningful for our purpose. A valid test must identify the children with real knowledge and skills in the domain being tested, not just children who are good at taking tests. The classic example of low validity is the child whose low score on a math test is due to poor reading comprehension – the inability to determine what the questions were asking not the inability to do the math. On essay tests, knowledge of the content being tested is conflated with the ability to write. Of course if the children are tested on things they were not taught, the tests won't tell us anything about what the children learned in school.

The second requirement for a test is reliability. Essentially this means consistency. Does the test give the same results over time? Do students whose classwork is comparable get comparable scores?

Lets look at the reliability of student test scores for teacher evaluation. It is extremely weak. Haertel (2013) cites a typical study in which over half the variation in teacher scores was random or unstable. He mentions teacher distress at the way scores are improbably different from year to year or even from class to class Indeed an analysis of New York City's teachers' scores by Gary Rubenstein found no correlation at all between different "measures" of the same teachers from one year to the next. For teachers in junior high or high school, there was no correlation between a teacher's scores from one class period to another. Value added modeling, the method used to compare student scores, is a sophisticated statistical procedure. Statisticians say it cannot be accurately used for a small sample like a single class. (Murnane & Steele, 2007).

Even more serious, when we try to use student test scores to evaluate the effectiveness of teachers or schools, we have lost validity. The standardized tests are designed to tell us about the learning of an individual child in an individual subject; the technical qualities of the test are measuring those taking the

test, not their teachers. In a statement issued in 2010 ten leading education researchers declared that student scores cannot be aggregated to draw a conclusion about a teacher or a school. (Baker et al, 2010)

Another problem with using student test scores to evaluate teachers is that students differ greatly in their response to instruction. Those with better backgrounds or more of what is termed intelligence , will more readily grasp the lessons. Those whose temperaments or cultures lead them to be more cooperative or hard working will learn more than their peers who lack these qualities. Children are not assigned randomly to classes. We cannot assume that differences in teachers' aggregated test scores are due to the teacher's ability unless the students who earned the scores are comparable in every way.

A third factor affecting test scores is that regardless of the teacher's skill, students influence one another. A disruptive student will reduce the learning of the entire class. M.M. Kennedy (2010) found that even a gregarious student reduces student achievement by distracting classmates from their lessons. On the other hand, in a cooperative learning group, a child who can mediate or lead can help the entire group meet learning goals.

Furthermore, W. Edwards Deming (1993), the business management expert says, "It is the structure of the organization rather than the employees, alone, which holds the key to improving the quality of output." In the case of schools, a teacher's effectiveness is dependent on the resources provided and support from the administration. Is the building harmonious with an atmosphere encouraging learning? Is there a reasonable curriculum with appropriate supplies and materials for delivering instruction? Are interruptions minimized? Are teachers given time to plan instruction, especially when assigned to teach new grades or new subjects? Is there staff to support teachers dealing with troubled students and their families? Do children have access to interesting books for independent reading?

When we look across New York State we must ask about a fifth factor, income inequality. What opportunity to learn is available in the community? Families in poverty cannot provide the support and enrichment that are available in middle class communities. This week the New York Times (Rich, 2013) reported on a study by Anne Fernald of Stanford University. She found that by age two, middle class children's vocabularies were thirty percent higher than those of children raised in poverty. Neighborhoods with good libraries, book stores, art stores, puppet shows, etc. have a different impact from neighborhoods with liquor stores and payday lenders. Of course attitudes affect learning and scores. Students and therefore teachers in communities in which there is social pressure to do well in school get better scores than those from communities in which urge children to resist authority and defy teachers. Thus, Haertel (2013) suggests that aggregating student scores tells us who the teacher taught, not how the teacher taught.

Unless we can correct for all these factors we cannot know what portion of aggregated test scores is influenced by the teacher's work. We do know that since the Coleman Report in 1966, all subsequent studies have shown that the major factor in children's school achievement is their home environments (Viadero, 2006). Most studies estimate that around 60% of children's school achievement is accounted for by the family or primary caregivers. Thus attributing the variation in test scores to an individual teacher is an appalling distortion of professional assessment standards.

A different problem with the high stakes placed on testing is that the more the tests count, the more they change the behavior of teachers, students and their families, and school administrations. This phenomenon has been labeled the Campbell Effect, named after Donald Campbell who found this influence in many different situations. Campbell, one of the founders of the field of program evaluation, found that attaching high stakes to evaluation led to the distortion of the processes that were being evaluated (Campbell, 1976). Teachers and principals feel pressured to "teach to the test" in order to earn high scores. School achievement tests usually focus concrete thinking so answers can be scored easily. If

we want children to learn to deal with complex problems that do not have a single answer, we should not be diverting time to preparing to give stylized answers to simplistic questions. We should be teaching skills and strategies for problem solving, critical thinking, and the analysis of texts and situations. Children must learn to respect one another's thinking and work in teams to analyze complex problems.

The absence of a standard unit of measurement and lack of validity and reliability make the use of student scores to judge teachers and schools a tragic waste of time and money. This policy is required by the federal Race to the Top, but the federal money is not used for instruction, and time and money are taken from instruction to give "data" that is not accurate. New York State should withdraw from Race to the Top and give more attention to instruction and less to testing.

The Common Core State Standards are appealing in seeking higher level thinking and more complex lessons, but they have never been field tested. As mentioned above, simplistic tests, labeled as aligned with the Common Core, cannot determine whether children can use complex cognitive processes. The idea of setting higher standards because children could not meet the standards set by No Child Left Behind, is counter intuitive, and there are many problems with the way the standards are being implemented and assessed. The standards for the primary grades are not developmentally appropriate with vocabulary and concepts far beyond the capacity of small children. In addition, testing authorities over 50 years ago said that children could not cope with test booklets with separate answer sheets until they were in the third or fourth grade (Lindquist, 1963). Teachers are reporting distress among children faced with tasks that are incomprehensible.

Achievement gaps due to income inequality and differences in opportunity to learn cannot be overcome by decree. Years of research have shown that small classes, good libraries, highly trained and experienced teachers, and the Reading Recovery program can improve the achievement of children from low income families. All these require devoting more money and care to educating needy children. New York State could begin by complying with the court's dictates in the Campaign for Fiscal Equity. Educational policies should be based on educational research by those who understand learning and instruction, not on the business practices, false analogies, and invalid "data."

References

Baker, E. L., Barton, P. E., Darling –Hammond, L. D., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L.A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers: Briefing Paper 278. Washington, DC: Economic Policy Institute.

Campbell, D.T. (1976). Assessing the Impact of Planned Social Change. Dartmouth College, Occasional Paper Series, #8.

Deming, W. E. (1993). The New Economics for Industry, Government, Education (2nd ed). Cambridge, MA: MIT Press.

Haertel, E.H. (2013) Reliability and validity of inferences about teachers based on student test scores. William H. Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Service.

Kennedy, M. M. (2010). Attribution Error and the Quest for Teacher Quality. Educational Researcher, 39, 591-598

Lindquist, E.F. (1963) Educational Measurement. Washington, DC: American Council on Education.

Murnane, R. J. & Steele, J. L.(2007). Measuring teachers' effectiveness through value-added modeling. Future of Children 17. 26-27.

Rich, M. (2013, October 22). Language-Gap Study Bolsters a Push for Pre-K. The New York Times. CLXIII, 56297, 1.

Viadero, D. (2006). Race Report's Influence Felt 40 Years Later: Legacy of Coleman study was new view of equity. EdWeek [Online] Available http://www.edweek.org/ew/articles/2006/06/21/41coleman.h25.html

Winerip, M. (2011, December 18). On Education: 10 Years of Assessing Students With Scientific Exactitude. The New York Times. retrieved from http://www.nytimes.com.

Submitted October 24, 2013

Rosalie Friend, Ph.D.

Rosalie Friend is a retired adjunct associate professor, Hunter College/CUNY School of Education
Dr Friend can be reached at rfrien@hunter.cuny.edu or 718-965-4074.

Among the materials on the reference list, Haertel's 2013 paper gives an expert in-depth analysis of the reliability and validity of the use of student test scores to evaluate teachers. I would strongly recommend that the committee study this report. I would be happy to discuss this information with the committee's staff.

Isabel Nunez of Concordia College and Bruce Baker of Rutgers University have also done valuable research in this area.