Educational Reform Through High Stakes Testing—Don’t Go There

Richard A. Huber

Watson School Of Education

Christopher J. Moore

Science and Math Education Center

University of North Carolina at Wilmington

601 S. College Rd.

Wilmington, NC 28403-3297

phone: (910) 962-3561

fax: (910) 962-3988

e-mail: huberr@uncwil.edu

Educational Reform Through High Stakes Testing

Educational Reform Through High Stakes Testing—Don't Go There

Given the current state of knowledge about the negative impacts of standardized testing, it may seem reasonable to conclude that science education supervisors need not be overly concerned about how such tests are employed within educational policies and initiatives promulgated by state and federal agencies. After all, a rich literature base documents the risks and potential pitfalls of standardized testing, and a number of well-researched national standards, such as the National Science Education Standards (National Research Council, 1996) provide unambiguous guidance based upon that literature. Before we grow complacent, however, we should take a careful look at what is going on in at least some of the 20 states that have begun implementing "high-stakes" accountability testing programs. In this paper, we outline the impacts of one such state program, North Carolina's New ABCs of Public Education, on K-8 science instruction and reform efforts. The evidence presented indicates that the North Carolina ABCs Program, which was held up by President Clinton in his 1999 State of the Union Address as a model for the nation, has derailed efforts to implement Standards-based reforms in many of North Carolina's classrooms. We believe that science education supervisors across the nation should view North Carolina's ABC Program like a warning alarm from a coal miner's canary.

In a 1997 article, Time Magazine referred to "high-stakes testing" as, "the latest silver bullet designed to cure all that ails public education" (Kunen, 1997). Unfortunately, as pointed out by Robert Shaffer of FairTests, an advocacy organization concerned with equity issues in standardized testing, this trend demonstrates a disturbing disregard for the scientific literature on testing. Shaffer states, "every professional guideline says, 'don't use these test scores as a sole criteria to make decisions.' But this is guidance that is widely abused and ignored" (CNN.com, 1999).

The admonition to use test results cautiously is highly consistent with the National Science Education Standards. While the Standards does recognize accountability testing as an acceptable practice (National Research Council, 1996, p. 89-90), the Standards also stresses that such testing should be conducted in accordance with rigorous quality assurances safeguards--which are consistent with Standards-based reform practices and goals. For example, the testing should utilize (1) authentic assessment tasks, (2) unbiased assessment tools, and (3) sound sampling and analysis strategies. Additionally, the programs or initiatives that give rise to the assessment should be sound and supportive of broader educational reform

efforts and goals, including those within the Standards-based reform movement.

Importantly, the Standards specifically addresses the need to be wary of policies set by elected or politically appointed leaders. Within the discussion of System Standard C, which calls for coordination among reform initiatives, the Standards explain that,

> New administrations often make radical changes in policy and initiatives and this practice is detrimental to education change, which takes longer than the typical 2- or 4- year term of elected office. Changes that will bring contemporary science education practices to the level of quality specified in the Standards will require a sustained effort (National Research Council, 1996, 231-232).

These concerns are echoed in the literature on testing-based reform initiatives. For example, Corbett and Wilson (1991) discuss how "reform by comparison" initiatives, such as the New ABCs, lack "the inherent patience needed to nurture better educational results over the long run" (p.3).

System Standard F of the National Science Education Standards directs science education supervisors to address concerns such as those noted above. System Standard F calls upon the science education community to be vigilant in reviewing new educational initiatives and policy instruments that may have negative "unintended effects on the classroom practice of science instruction" (National Research Council, 1996; p. 233). As explained within the discussion of System Standard F, "Even when as many implications as possible have been carefully considered, well-intentioned policies can have unintended effects. . . . unless care is taken, policies intended to improve science education might actually have detrimental effects on learning" (National Research Council, 1996). This standard is perhaps nowhere more important than in the case of "silver bullets," where program implementation tends to progress more rapidly than does program evaluation and validation.

Overview of the New ABCs as an Example of High-Stakes Testing

High stakes testing programs are a product of the growing movement to improve public education by ensuring that schools more fully reflect and conform to the needs of business and industry and the criteria used to judge success and effectiveness in business and industry. Towards this end, the New ABCs and other high-stakes testing programs emphasize the importance of holding teachers and schools accountable for student learning as measured by students' performances on standardized tests. In order to

ensure that these personnel are adequately motivated, substantial rewards and hefty sanctions are linked to testing outcomes. The central role of accountability testing in the New ABCs is demonstrated by the fact that the letter "A" in the "ABC" acronym stands for "accountability." The other two core objectives of the program are to increase the emphasis placed upon the "Basic" subjects and to provide an increased level of "Control" of program implementation and pedagogy at the local (school) level (North Carolina Department of Public Instruction, 1996).

The "control" component of the program appears to have been intended to help mitigate known or predicted problems associated with standardized and/or accountability testing. In theory, the increased local control would ensure that personnel working close to the students--teachers and school-level administrators--would keep the students' best interests in mind when making decisions about how to implement the program. Unfortunately, this component of the program does not appear to have been successful (Jones et al, in press).

As suggested above, the poor performance of the New ABCs with respect to its "control" objectives could have been anticipated had attention been paid to warnings in the Standards and in the professional literature regarding the need for "long-haul" (rather than 2- or 4- year) perspectives. The Standards strongly emphasizes that the process of increasing the authority and control afforded to teachers will be very slow because the changes in teachers' roles will only occur only as substantive reforms are made throughout the educational system. Thus, it is unreasonable to hold teachers accountable for executing such authority effectively prior to having successfully completed the pre-requisite substantive systematic reforms necessary to support teachers in their new roles.

Additionally, some researchers are concerned that the types of high-stakes utilized within the accountability component of the New ABCs program undermine efforts to increase control at the school level. For example, Wildy and Wallace (1997) discuss how it is particularly important in science education that accountability programs follow professional models that (1) maintain long-term perspectives (2) promote a culture of trust and support, and (3) emphasize professional development. Unfortunately, such approaches appear incompatible with the high-stakes accountability agenda of the New ABCs. As stated by Jones et al (in press), "when the State Board of Education has the power to shut the doors of the school based on end-of-grade test scores, there is no local control of education."

The stakes associated with the accountability component of the New ABCs are substantial. In North Carolina schools, teachers receive bonuses when test scores are high (schools receive as much as $1,500 per teacher). Severe sanctions are applied when schools fail to meet their "expected growth" standards, especially for schools that "earn" the "low-performance" label. These sanctions include the following:

- publication of performance measurements (the distinctions range from "school of distinction" to "low performing school"),
- mandated assistance from state-provided teams (comparable to a hostile take-over in the business world),
- competency tests for teachers (teachers are given three chances to pass the test before being discharged), and
- removal of principals and teachers who are "not willing to improve their practice."

The stakes continue to rise as the program is implemented. For example, recent legislation in North Carolina has simplified the procedures for dismissing teachers by "streamlining" the appeals process. Additionally, test results are playing an increasingly more significant role in decisions concerning student placement, advancement, and retention.

Proponents and critics of high-stakes testing both advocate that the high stakes have a powerful impact on motivating teachers and school administrators to do what is necessary to bring about higher scores. However, there is less agreement about what can be inferred from those test scores. Proponents advocate that the results demonstrate success. Consider, for example, the meaning North Carolina's Governor Hunt attributed to data showing a rise in test scores for the 1997-98 school year:

> The results show us that North Carolina's schools are working. . . . Through the ABCs of Public Education, our schools are working like never before to put children and their education first (North Carolina Public Schools Infoweb, 1998).

In contrast, critics question the assumption that high scores equate with improved schools. Additionally, the literature on standardized testing raises concerns about the desirability of programs that may motivate teachers and administrators to do "whatever is necessary" to bring about higher scores (Jones et al, in press; FairTest, 1999a; CNN.com, 1999; Shapiro, 1999; Neill, 1998; Darling-Hammond, 1991;

Haladyna et al. 1991; Madaus, 1991; Neill and Medina, 1989; Brandt, 1989; Smith, 1991a; Smith, 1991b). In fact, for at least a decade researchers have argued that using standardized test scores as the primary basis for any policy decision-making is "reckless," given what is known about the limited validity, accuracy, and reliability of the tests (Neill and Medina, 1989). In a survey of state programs used to establish accountability in the public school systems (not all of which were high-stakes programs), FairTest concluded that 2/3 of the programs impeded rather than promoted educational reform (FairTest, 1999a). In this study, North Carolina's ABC program received the lowest possible rating (1 on a scale of 1 to 5), which distinguishes it as a program "requiring a complete overhaul" (FairTest, 1999b).

The New ABCs and Standards-based Reform Goals

In assessing the New ABCs in terms of the Standards, this paper considers the two areas of emphasis in the Standards--equity and excellence.

Equity Issues

System Standard E of the National Science Education Standards states that, "science education practices must be equitable" (National Research Council, 1996; p 232). In explaining this standard, the Standards emphasize the need to ensure that programs overcome, rather than compound, "well-documented barriers" to learning science for selected groups of students, including those from economically disadvantaged populations. One of the objections to accountability testing is that the testing may promote such inequities. As stated by Darling-Hammond (1991):

> Applying sanctions to schools with low test scores penalizes already disadvantaged students. Having given them inadequate schools to begin with, society now punishes them further for failing to perform as well as students attending schools with more resources (p. 222).

A number of other serious equity issues have been raised in the literature on standardized testing. For example, there is evidence that the tests are biased to middle class, white, male worldviews (CNN.com, 1999; Darling-Hammond, 1991; Neill, 1998). As the high-stakes testing movement builds momentum, the legal implications associated with these issues is drawing increasing recognition and attention. For example, the U.S. Education Department's Office for Civil Rights has recently begun the process of developing a policy that would restrict testing practices. A draft policy statement, which is being circulated within the educational community for comment, would ban "the use of any education test which has a

significant disparate impact on members of any particular race, national origin, or sex. . . unless it is educationally necessary and there is no practicable alternative form of assessment" (CNN.com, 1999).

The New ABCs program purports to address these issues by using a "complex formula," which focuses on improvements, to determine each school's required performance goals. Proponents of the program claim that, "the decision to focus on progress removes the nettlesome problem of unfairly expecting poor rural schools or inner-city schools to do as well as their counterparts in wealthy suburban areas" (Simmons, 1997).

Given the complexity of equity issues of concern, the efficacy of this relatively simplistic solution is less than self-evident. For example, there is no reason to assume that the practice of focusing on improvement would remove racial or cultural biases in tests. Additionally, concerns have been raised that schools may have difficulty in recovering from a "low performance" rating because the label might scare away highly qualified personnel--including principals and teachers who might otherwise be recruited to help turn around schools in disadvantaged districts (Kurtz, 1998). We believe that education supervisors should be wary of "silver bullets" in the details of high-stakes testing programs, such as the practice of measuring improvement discussed here, which purport to resolve complex issues through strategies that appear to be relatively simplistic and largely invalidated.

Excellence Issues

The National Science Education Standards outlines numerous changes in emphasis that will occur as the Standards' vision is realized, some of which are summarized in Table 1. The evidence reviewed here indicates that high stakes testing programs appear to drive changes in the opposite direction of those envisioned in the Standards.

Insert Table 1 about here

As shown in Table 1, the Standards envisions a shift of focus away from one in which instruction and assessment focus on a broad body of discrete knowledge. Instead, the Standards advocates a narrower focus on key concepts directed towards deeper, richer understanding. The central role of inquiry drives a shift of focus from lower-level thinking to problem solving and other higher-order thinking skills.

Importantly, the changes envisioned in the Standards call for substantial increases in the amount of time allocated for science instruction. Without exception, the New ABCs appears to be driving science instruction in North Carolina's elementary and middle school classrooms in the opposite direction as advocated by the Standards, that is, away from inquiry.

High stakes assessments programs in general, and the New ABCs in specific, clearly are antithetical to the Standards' goal of decreasing the emphasis placed on "standardized assessments unrelated to Standards-based programs and practices." One consequence of this shift is that teachers are spending more time teaching the tested subjects, at the expense of other subjects. In grades K-8, where the testing focuses on the "basic" subjects of mathematics, reading and writing, science often is marginalized. Additionally, teachers are spending more time teaching test-taking skills and having students take "practice tests."

The changes in how instructional time is allocated can be substantial. For example, in a survey of North Carolina elementary school teachers, Jones, et al (in press) report that 80% of the teachers indicate that they spend over 21% of their total teaching time practicing End-of-Grade (EOG) tests. Additionally, over 28% of the teachers indicate that they spend from 61% to 100% of their teaching time practicing for the tests. The mean amount of time devoted to science instruction among these teachers is 99 minutes per week. The teachers also report that science instruction was often radically marginalized as test time grew closer.

There is reason to doubt science instruction would be aided by the addition of a science test to the testing schedule. Within the arena of the tested subjects, teaching practices appear to have been degraded to strategies focused on "teaching to the test," which are antithetical to Standards-based practices. For example, in mathematics instruction, at least one county system provides teachers with a database of math questions representative of end-of-grade math test questions. Teachers are also provided with a breakdown of the questions that organizes them by objective and identifies the percentage of questions per each objective that were present in previous EOG tests. We have observed teachers being instructed (pressured) by county-level and school-level administrators to adjust the emphasis of their instruction to match the pattern of emphasis identified on past years' tests.

Although evaluative literature on the New ABCs is just beginning to become available,

questionable practices such as those described above appear to be prevalent (Jones et al, in press; Jones, 1997; FairTest, 1999b). Additionally, concerns have been raised in the literature about the tendency for high-stakes testing programs to encourage school administrators and teachers to engage in practices that are questionable in terms of both pedagogy and ethics. For example, the literature on standardized testing raises substantial concerns about the how widespread practices of "teaching to the test" lead to unethical teaching practices that invalidate test results (Haladyna et al, 1991). The insidious nature of these problems and the evidence to date on the New ABCs suggests that numerous undesirable practices might well follow in the wake of a science accountability test, should one be implemented in the future. We believe that science education supervisors should be proactive in addressing high-stakes testing and that they should not wait for mandated science testing to reach their schools before taking action.

High-stakes testing in general, and the New ABCs in particular, also appear to work against the Standards' goals involving affective domain learning. Test anxiety replaces an open atmosphere of exploration where diverse ideas are respected and risk-taking is valued (Hill and Wingfield, 1984). Competition flourishes at the expense of community (Shapiro, 1998). A love of science--and of learning in general--is anything but nurtured. For example, Jones et al found that teachers were six times more likely to report that the New ABCs program resulted in a negative impact on students' "love of learning" than a positive impact.

Finally, the New ABCs appears inconsistent with the Standards' goals regarding what children are taught about the nature and purposes of science itself. The Standards calls for a shift of emphasis that de-emphasizes science as a body of factual knowledge and emphasizes science as a way of structuring and using inquiry to answer real questions and investigate real problems. This shift of emphasis is a move away from science as the accumulation of factual knowledge separate from exploration and experimentation (with experimentation often limited to the closing activity for a unit of study). The shift of emphasis is a move towards a model of science as "argument and explanation," involving ongoing, repeated, and public investigation and experimentation in which students "combine process and scientific reasoning and critical thinking to develop their understanding of science" (National Research Council, 1996; p. 105).

It seems unlikely that high-stakes testing programs, such as the New ABCs, will further these

goals for at least two reasons. First, as suggested above, high stakes testing tends to promote an emphasis on teaching what is easily measured with objective (e.g., multiple choice) tests. Objective tests are a poor tool for testing the ways in which a student has developed the values and attitudes conducive to being able to truly apply scientific inquiry to real world problems. Secondly, once again the realization of the Standards' vision takes time, which is all too often a scarce resource with end-of-grade tests only a matter of a few months or weeks ahead.

Conclusion

As a case study, North Carolina's New ABCs of Public Education provides compelling evidence that high-stakes testing is not a "silver bullet" that will cure all the ills that beset our schools. In fact high-stakes testing is problematic. Nonetheless, it is reasonable for the public to expect that schools and teachers be held accountable to high professional standards. Further, as recognized in the Standards, assessments of student learning can be used as a valid tool for establishing such accountability. However, such testing will only be effective if it is implemented properly. Towards this end, The National Science Education Standards is a useful guide in that it provides a model and a vision of recognized best practices.

Importantly, the Standards also provides a guide to potential false starts and pitfalls in educational reform. For example, one major weakness of the New ABCs appears to be that it has been implemented under a cloud of urgency. Also, many of the concerns raised here may well stem, at least in part, from failures of the New ABCs to accomplish its objective of providing increased control of educational policies to local schools and teachers. As a consequence of these shortcomings, teachers may have had less, rather than more, control in ensuring that the sweeping changes wrought by the New ABCs are in students' best interests. As noted above, the Standards provided warnings relevant to both of these apparent shortcomings.

There are no simple solutions to the complex problems associated with accountability testing. However, research and standards of best practice can inform decisions about how to move towards viable solutions and sound practices. Science education supervisors can play an important part in helping to guide research and policy development. It is the professional educator's responsibility to help ensure the established knowledge base on assessment practices is not disregarded.

References

Brandt, R. (1989, April). On the Misuse of testing: A conversation with George Madaus. Educational Leadership. 26-29.

CNN.com (1999, June 14). Standardized tests under fire. [On-line]. Available: http://cnn.com/US/9906/15/standardized.tests.

Corbett, H. D., & Wilson, B. (1991). Testing, Reform, and Rebellion. NJ: Ablex.

Darling-Hammond, L. (1991, November). The Implications of testing policy for quality and equality. Phi Delta Kappan. 220-225.

FairTest (1999a) Testing our children: [On-line]. Available: http://www.fairtest.org/states/survey.htm ("Introduction" link).

FairTest (1999b) Testing our children: [On-line]. Available: http://www.fairtest.org/states/nc.htm.

Haladyna, T., Nolen, S., & Haas, N. (1991). Raised standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20(5), 2-7.

Hill, K., & Wingfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. The Elementary School Journal, 85(1), 105-126.

Jones, G. M., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T. and Davis, M. (in press), The impact of high stakes testing on teachers and students, Phi Delta Kappan.

Jones, T. (1997, August 17). ABCs warrant an F. Raleigh News and Observer, section 29-A.

Kunen, J. S. (1997, June 16). The test of their lives. Time, 149(24), 62-63.

Kurtz, M. (1997, Nov. 6). State poised to approve teacher tests. Raleigh News and Observer. section A-1.

Madaus, G. F. (1991, November). The effects of important tests on students: Implications for a national examination system. Phi Delta Kappan. 226-231.

National Research Council. (1996). National Science Education Standards. (1 st ed.). Washington, DC: National Academy of Sciences.

North Carolina Department of Public Instruction. (1996) A guide to the ABCs for teachers,

Raleigh, NC.

North Carolina Public Schools Infoweb (1998). ABCs results show strong growth in student achievement K-8; high schools post first year's results. [On-line]. Available: http://www.dpi.state.nc.us/news/abcs_results_98.html.

Neil, M. (1998, March). National tests are unnecessary and harmful. Educational Leadership. 45-46.

Neill, M. D., and Medina, N. J. (1989, May). Standardized testing: Harmful to educational health. Phi Delta Kappan. 688-697.

Shapiro, S (1998). Public school reform: The mismeasure of education, Tikkun, 13 (1), 51-55.

Simmons, T. (1997, August 8). 43% of schools in N.C. fall short. Raleigh News and Observer, section A-1.

Smith, M. (1991a). Meanings of test preparation. American Educational Research Journal, 28, 521-542.

Smith, M. (1991b). Put to the test: The effects of external testing on teachers. Educational Researcher, 20(5), 8-11.

Wildy H. & Wallace, J. (1997). Improving science education through accountability relationships in schools. Science Educator, 6(1), 11-15.

Table 1: Shifts in emphasis called for in the National Science Education Standards (based on National Research Council, 1996, pp. 52, 72, 113, 239).

Less Emphasis On. . .	More Emphasis on. . .
Standardized tests and assessments unrelated to Standards-based programs and practices.	Assessments aligned with the Standards.
Assessments aligned with "traditional" content of science education. (Science directed towards memorizing scientific facts and "getting an answer.")	Assessments aligned with the Standards expanded view of science. (Science as argument and explanation directed towards using evidence and strategies for developing or revising an explanation.)
Students doing relatively few experiments and using experiments primarily as a means of concluding an inquiry.	Students doing more experiments in order to develop understanding, ability, values of inquiry, and knowledge of science content. Applying results of experiments to scientific arguments and explanations.
Focusing on student acquisition of information and assessing "what is easily measured"—discrete factual information.	Focusing on student understanding on the use of scientific knowledge, ideas, and inquiry processes. Assessing "what is most highly valued"—rich, well-structured knowledge including scientific understanding and reasoning abilities.
Assessing at the end of learning to determine what students don't know.	Ongoing assessment to inform decision-making throughout the period of instruction.
Development of external assessments by measurement experts alone.	Teachers developing authentic assessments of student science learning.
Instructional activities that ignore or marginalize affective domain learning. Using competition among students as a motivational tool.	Instructional activities emphasizing affective domain learning and social skills.

Teacher as technician, follower (including follower of established curriculum), and target of change.	Teacher as intellectual, reflective practitioner (who is constantly adapting curriculum), leader, and source/facilitator of change.