WILLIAM H. ANGOFF MEMORIAL LECTURE Series

NATIONAL TESTS AND EDUCATION REFORM: ARE THEY COMPATIBLE?

BY

LYLE V. JONES

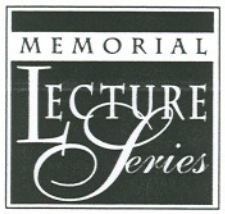
William H. Angoff
1919 - 1993

William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. *Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.*

The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

National Tests and Education Reform: Are They Compatible?



The fourth annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on October 8, 1997.

Lyle V. Jones
University of North Carolina at Chapel Hill

Policy Information Center
Princeton, NJ 08541-0001

Preface

We are pleased to make available the fourth William H. Angoff Memorial Lecture, given at Educational Testing Service by Lyle V. Jones. In July of 1995, ETS Vice President Henry Braun described this memorial lecture series:

> The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, 43 of them at ETS, Bill made major contributions to psychological and educational measurement and was deservedly recognized by the major societies in the field. As the notion of an annual lecture series took shape, the idea that these lectures should be devoted to relatively non-technical discussion of public interest issues related to educational measurement struck us all as eminently suitable. This was an aspect of our field in which Bill was keenly interested and into which he made several successful forays. I know he thought it part of our professional obligation to encourage and support reasoned public debate on important topics.

Professor Jones addresses a matter now being vigorously debated, the creation of a national test. Such a test was proposed by President Clinton, with initial test development work already under way. At this writing its fate is still uncertain as the matter is debated within the Congress.

This examination by Professor Jones is indeed timely, and we think it contributes to "reasoned public debate," the goal set for this lecture series by Henry Braun.

Paul Barton
Director, Policy Information Center

Preamble

It is my privilege to be invited to present a William H. Angoff Lecture. The lecture series is an appropriate vehicle for remembering and honoring Bill Angoff. His professional contributions have influenced several generations of students of psychometrics. Bill's personal grace and warmth and his total integrity continue to inspire his associates and friends. I am grateful to be among that number.

Ten years ago, Bill invited me to present a paper at an American Psychological Association (APA) symposium that he organized and chaired. There, I urged improvement of procedures for measuring educational achievement (Jones, 1988); I emphasized the importance of alternatives to multiple-choice testing, to encourage student mastery of understanding as opposed to the simple recognition of right answers. Then, at the ETS Invitational Conference of 1990, I prepared a critique of a paper that supported plans for mandatory national testing in the schools of Great Britain (Jones, 1991). Now, I shall address some issues surrounding another set of plans for national tests to be given to school children, this time right here in the U.S.A.

It is always a pleasure to renew acquaintances at ETS. I first visited as ETS was nearing five years of age. I was a young assistant professor at the University of Chicago. L. L. Thurstone, my mentor and senior colleague, arranged an appointment for me to meet Harold Gulliksen and Ledyard Tucker at ETS on Nassau Street. My next visit brought me to this impressive campus, quite a change from the earlier ETS office suite. In many visits since then, I profited immensely from discussions with ETS researchers and with other visiting research advisors, and have enjoyed friendships with many staff members, including your presidents, Henry Chauncey, Bill Turnbull, and Greg Anrig.

Beginning in the 1950s, ETS has employed a number of alumnae of graduate and postgraduate programs of the Psychometric Laboratory at Chapel Hill. Among more recent examples is that of Nancy Cole. Some years ago, I recommended Nancy's admission to graduate school. That decision was among my most astute, as became clear within the earliest days of her arrival at the University of North Carolina. Her many career successes repeatedly have revalidated that judgment.

Education Reform

The term "education reform" appears in my title. I mean by this the kind of enrichment of teaching and active engagement of students in learning that are endorsed by the National Council of Educational Standards and Testing (1992) and fostered by the NCTM Standards (National Council of Teachers of Mathematics, 1989) and by standards developed by the National Research Council (1993, 1996). For one thing, such reform leads to the establishment of appropriate content for learning grade by grade, to build greater breadth and depth of knowledge and skill as students progress from one grade to the next. My understanding of "school reform" is quite different from that of William Bennett (1997), who perverted the term to simply denote parental choice of school.

The President's Proposal

In his State of the Union address in February of 1997, President Clinton proposed national tests in mathematics for eighth graders and in reading for fourth graders. His intent appears to be grounded in the vague notion that national tests somehow will improve learning in mathematics throughout elementary school and will improve the acquisition of reading skills in the earliest school years. He challenged "every school, every state, every student to participate by 1999" (The White House, 1997a). By latest count, 7 states and 15 urban school districts have signed on, constituting about 20 percent of the nation's pupils at each grade. Reportedly, some urban school districts now are having second thoughts and may withdraw.

Less than a year earlier, President Clinton had agreed with the nation's governors that both educational standards and assessments should be developed by the states. Clinton told the governors, "We can only do better with tougher standards and better assessment and you should set the standards. I believe that is absolutely right." (The White House, 1996).

One may wonder what prompted the President to execute an abrupt about-face on state versus national responsibility for accountability in education. The scope of the federal government is being reduced; both Congress and the Executive Department espouse a shift in control of governmental programs to state and local levels. But for education — an endeavor traditionally controlled by states, districts, and individual schools — we now see efforts to move in the opposite direction. Consideration of a stronger federal role in education might be more defensible were U. S. school achievement to be declining. However, the National Assessment of Educational Progress (NAEP) and other recent evidence show rising levels of achievement in recent years. The President said in Martha's Vineyard only last month, "Our schools are offering broader and deeper curricula now, our students are taking more challenging courses now, our schools, by and large, are much better run now" (The White House, 1997b). Under the control of states and school districts, both average reading achievement at age 9 and average math achievement at age 13 are significantly higher than in the early 1970s (Campbell, Voelkl, & Donahue, 1997), despite demographic shifts that might have been expected to make any improvement difficult to achieve.

The President included two components in his proposal — developing national standards and developing national tests. Very recently, however, the President seems to have suggested that he views the formulation of standards and the establishment of national tests as inseparable. In his radio address to the nation on Saturday, September 20, the President said, "(T)he House of Representatives ... voted against developing the national standards we need to challenge students, improve teaching, empower parents and increase accountability in our schools" (The White House, 1997c). He went on to say he would veto legislation "that denies our children high national standards." As I understand the legislation, the House voted to deny funds for national testing (except for NAEP and the Third International Mathematics and Science Study (TIMSS)), but did nothing that would inhibit the development of standards.

Content standards for mathematics now have been developed by NCTM and are widely employed by school districts and states. Science standards have been introduced by the National Research Council (1996), and standards for other subjects are being discussed by other groups. While still controversial, it does now appear to be appropriate and constructive for national bodies (as distinguished from federal bodies) to develop curricular content standards for consideration by states and school districts.

Content standards are one thing; performance standards are another. Establishing national performance standards raises a higher level of controversy. Indeed, for NAEP it already has done so. Efforts to fix cut scores in NAEP to separate achievement levels, Basic, Proficient, and Advanced, have not been successful. The procedures employed resulted in cut points being set too high (U. S. General Accounting Office, 1993; National Academy of Education, 1993). For example, consider mathematics at the fourth grade. Recent results from TIMSS show that average math performance for U. S. fourth-graders is significantly above the international average (Mullis, Martin, Beaton, Gonzalez, & Smith, Kelly, 1997). Yet, the NAEP report for 1996 shows only 18 percent of U. S. fourth-graders Proficient and only 2 percent Advanced in math. Thirty-eight percent are reported to be Below Basic. (Reese, Miller, Mazzeo, & Dossey, 1997). When U. S. fourth-graders perform reasonably well in an international comparison, isn't it unreasonable that only 20 percent are reported to be "proficient" or better?

Justification for National Tests

How do the President and the U.S. Department of Education justify "voluntary national tests?" (Note that these are voluntary in one sense only. A state or district is free to adopt or not to adopt the tests. Having adopted a test, a state or district is likely to mandate its use for every student in every classroom in every school district.) The President has supported the development of national tests, but has failed to state a clear purpose for his proposal. He has noted that NAEP tests are given only "to a sampling of students in states, and we only know what ... the state scores are. So we have to do this for the whole nation." (The White House, 1997a). He emphasizes that these will be standards-based tests, in contrast to "normal grading. With the idea of standards you want everybody to clear at least the fundamental bar." Even with a generous translation of the President's words, he has failed to justify the need for a national test.

In Congressional testimony earlier this year, Secretary of Education Richard Riley cited a study showing that many more students scored "proficient" on state assessments than on the national assessment and concluded that "state standards are still not high enough." Riley went on to say, "This is why these proposed national voluntary tests are important" (Riley, 1997). In fact, the NAEP achievement levels were set too high.

Deputy Secretary of Education Marshall Smith claims that the purpose of the tests is "to change odds for kids in the two critical areas of basic skills, reading and mathematics" (U.S. Department of Education, 1997). He notes that for students who don't read independently by grade 4, the odds of high school graduation and entrance to college "go way down," and that much the same can be said for eighth grade students who can't "grapple with reasonably complex mathematics." Smith suggests that test results will lead to interventions for students who are deficient, to help them succeed. He envisions the tests, then, as sort of national minimum competency exams, and he trusts that states, districts, and schools will do what may be required to help low-scoring students. But as stressed recently by Robert Stake (1995), while "` many people expect standardized achievement tests to have diagnostic properties, ... most teachers are skeptical. ... The tests seldom inform teachers of previously unrecognized student talents and seldom identify deficits in a way that directs remedial instruction (Koretz, 1987)'."

Secretary Riley has said, "I believe these tests are absolutely essential for the future of American education" (Riley, 1997a). Can he really mean that? According to the Department of Education, the tests will "provide parents and teachers with information about how their students are progressing compared to other states, the nation, and other countries. An individual student's achievement level will be related directly to information obtained from two state-of-the-art educational assessment surveys — NAEP and TIMSS. Parents can see how their own children measure up to the highest standards of performance at the national and international levels" (U.S. Department of

Education, 1997). Isn't this an overly ambitious objective for a single set of items in a 90-minute test? Just how will links be established to NAEP, TIMSS, and to existing state tests? (See Baker & Linn, 1997.)

Consider eighth-grade math. (Fourth-grade reading tests will follow the same general rules.) It is proposed that 90 minutes of testing will be split in half between multiple-choice and short-answer test items to be based on content frameworks developed by NAEP. When recently asked on the *The News Hour with Jim Lehrer* whether the same test would be given to all students, all over the United States, Secretary Riley replied. "Absolutely, yes." (Specifications do call for the development of alternate forms to be used in subsequent years.) The new exam "will be made available to states, districts, schools, teachers, and other individuals to assess students in the spring of every year, beginning in spring 1999. Many commercial test publishers will likely offer the tests as a component of their product line. ... States, school districts, and test publishers will be responsible for administering and scoring the tests. After each administration, the entire test along with answers, scoring guides and other materials will be released to the public and placed on the Internet" (U.S. Department of Education, 1997).

A number of serious impediments would be encountered were such a scheme to be implemented. Some pertain to logistics, but first, consider some deeper issues.

Some Arguments Against the Proposal

In a recent article in *Science*, Iris Rothberg (1995) presents a cogent critique of the use of mandatory tests to enhance learning. She concludes that a testing program does serve to increase test scores from year to year, but at costs that are too great to bear, a narrowed curriculum and an increase in student dropout. Rothberg suggests that the imposition of required testing serves to excuse society for its inadequate support of poor children.

Some of the arguments recently presented by Robert Stake (1995) mirror my own: "Mathematics test scores ... that do a good job of indicating which students are doing best and which are doing relatively poorly do not necessarily provide a valid indication of subject-matter mastery. One test alone will not provide valid measurement of the mathematics achievement of individual students or of a group as a whole. Test content almost always is too narrow. Just as ... a few books do not represent all the books in a library, twenty or thirty test items do not represent the broad range of mathematics skills and knowledge that teachers are teaching. For measurement of subject matter attained, the simplicity of testing is at odds with the complexity of teaching and learning."

"The public does not understand how there could be sincere objection to using standardized achievement tests to represent what should be taught. People ... presume that tests valid for one educational purpose are valid for others. As students for many years themselves, they have experienced teaching and testing; the question of alignment almost never came up. Today, when a mismatch is claimed, often they presume that the teachers are wrong."

In its report to Congress and to the nation, the National Council on Education Standards and Testing (1992) supported the development of national standards and, emphasized the importance of "*Multiple assessments - not a single test.* It will be up to the states, individually or in groups, to adopt assessments linked to the national standards" (p. 15). The National Council's Assessment Task Force is more explicit: "The new assessment system must be designed to ensure that states and local districts have the primary responsibilities for creating and implementing assessments for the purposes of accountability, school evaluation, student certification, reporting to parents, and instructional improvement. Such responsibilities should be state and local because decisions about schooling are made primarily at the state and local levels. Furthermore, there is no single best method of assessment. We need to provide for the creative development of multiple alternatives to assess the national standards," (p. F-15). These recommendations are essentially repeated in a report earlier this year to the National Education Goals Panel prepared jointly by the National Research Council and the National Council of Teachers of Mathematics (1997). We must wonder why the President and the Department of Education have ignored this good counsel. Shouldn't he and his advisors understand that it is unreasonable to measure with a single 90-minute test a child's mastery of a rich array of achievements?

Potential for Misuse

Misuses of test results would plague national tests as it continues to be troubling for some state and local testing programs. A current court battle in North Carolina illustrates one kind of misuse. Johnston County has chosen to use student scores on the North Carolina end-of-grade examinations as the sole basis for promotion to the next school grade. Parents whose children have been retained in grade despite having satisfactory classroom achievement records have sued the county school board. A judge has ruled that the school board has the right to use the State test scores in this way; an appeal is expected.

That the misuse of test scores can do harm to individual students is undeniable. Probably all of us can cite examples from personal knowledge. Consider one case, that of a member of my family. Based on a very low score on a math test at the end of the first grade, my nephew was assigned to the lowest math group in the second grade. During the first grade, his mastery of math had been unquestioned. His mother sought more information from the school. The boy's first-grade teacher said of the test result, "That can't be! He did very well in math. Those tests and answer sheets are still on file. I'll look into it." The teacher discovered that on his answer sheet, from test item 4 to the end, answers to items k-1 were the correct answers to items k. Having skipped an answer to item 4, all of the later answers were out of line. The mother brought this to the attention of the school principal, who told her "There is nothing I can do." This boy not only remained in the low track, he greatly enjoyed being top dog in that group, without exerting much effort. Indeed, throughout elementary school, he learned ways to manipulate the system so as to be assigned to the lower tracks, where competition was less challenging. Of course, he was disadvantaged in high school and college, not by low ability but by the lack of a solid preparation. The primary culprits here appear to be those who misused a single test score.

My intent here is to emphasize and support advice such as that included in ETS testing manuals for many years, that decisions about individuals should not depend solely on a single test score. Other relevant information also should be considered. Might the perceived high status of a national test increase the likelihood that it would be misused for decisions about individual students? And might decisions about individual students be contested on grounds of disparate oppportunities to learn the materials covered by the national test? (Baker & Linn, 1997).

Lessons from Other Countries

Stake (1995) notes that, "In Sweden and Iceland, objections to control by Stockholm and Reykjavik have resulted in more support for the teacher, less national specification of instruction, and less reliance on standardized testing. But the movement in the United States and most of the world is toward greater control by the government, less honoring of professional experience (particularly as to subject matter conceptualization by the teachers), and more emphasis on formalized student assessment."

Larry Cuban (1997) presents arguments against national tests and notes some pertinent results from TIMSS. Countries that employ mandatory testing seem not to display systematic differences in average achievement scores — better or worse — when compared with countries that do not.

A study of the effects on teaching of a high-stakes testing programs in British Columbia is especially revealing (Wideen, O'Shea, Pye, & Ivany, in press). Based on videotaped classrooms and teacher interviews from 10 school districts, the study compared classroom activity in randomly selected science classes at grade 12, where an end-of-year provincial examination was mandated, with science classes at grades 8 and 10. The mandatory exams at grade 12 were found to have a strong effect on teaching. While teachers at grades 8 and 10 reported that the 12th grade exams had little or no effect on their teaching, teachers at grade 12 reported the opposite. Moreover, the only exemplary science teaching observed by the researchers occurred at grades 8 and 10; strict lecturing, non-involvement of students, and an emphasis on tests and testing characterized the classroom at grade 12, where the prevalent question heard was "will this be on the exam?" The study concludes that the mandatory exam tends to freeze innovative teaching practices and to discourage inquiry and active student learning.

Regarding societal factors that are correlated with school achievement, Jaeger (1992) reported some interesting data *vis a vis* comparisons of the United states with Germany and Japan. (See also Behrman, 1997.) For example, among families with children, the percentage of single-parent families, about 25 percent in the U.S., is almost twice that of Germany and four times that of Japan. The poverty rate for children, now nearly one out of four in the U.S., is more than twice that of Germany. Note that in the U. S., nearly 30 percent of poor children have repeated a school grade, twice the percentage for non-poor children, and twice the percentage of poor than non-poor children had been expelled or suspended (Brooks-Gunn & Duncan, 1997). According to Al Beaton, the most recent data from TIMSS show that within every one of the nations that participated, poverty is related to achievement. Economic and societal factors are found to be good predictors of average student achievement scores in mathematics for industrial nations. Further, classroom variables predict only small portions of variability. Based on these findings, there can be little basis for surprise when we discover that the U.S. may lag behind Japan, Germany, and some other developed countries in average school achievement in mathematics.

While findings from TIMSS rank Japanese students among the best in the world, the Japanese complain that their schools are too regimented and too harmful to creativity. Especially in secondary schools, where the preparation for college-entrance exams

dominates student learning, it is said that students memorize lots of facts but never really learn how to think. Kristof (1997) cites a Japanese critic who attributes the excellent test scores of Japan's students to "endless drills by trained seals". Many Japanese parents reportedly are urging educational reforms to decrease regimentation in middle schools and high schools.

Learning from the British Experience

There are cogent lessons that we might learn about national tests from experiences in Great Britain.

The Education Reform Act of 1988, instituted by the Thatcher government and enacted by Parliament, established a national curriculum and required a "system for reporting of individual pupils' achievement in the National Curriculum subjects ... introduced in stages from 1991, under which parents in England and Wales will receive a progress report each year" (British Foreign and Commonwealth Office, 1989).

Paul Black (1990) suggests that the demand for external examinations in England arose from three sources: (1) the public distrust of teachers, (2) the belief that external exams are technically superior to assessments by the teacher, and (3) the public demand for common standards and for fairness of comparisons between schools. One is reminded of the urgent appeal of the late Albert Shanker (1985) to combat the first two of these concerns by making teaching into a more respected profession.

Early critiques of the British assessment plans include commentaries by Nuttall (1990), Stobart (1990), and Jones (1991), each of which anticipated the exorbitant costs and the ultimate demise of the ambitious program that required annual performance testing of every student in selected grades in England and Wales. (In Scotland, by a national referendum supported by school teachers and endorsed by voters, the national assessment program was rejected.)

To fulfill initial plans for student testing in England and Wales, standard achievement tasks — referred to as SAT's — were developed between 1989 and 1992. These were to be administered and scored by teachers. Because the SAT's required as much as 10 hours of teacher time per child (Lofty, 1993), the initial plan was modified and, in a national tryout examination, the bulk of the teacher assessment was replaced by three hours of paper-and-pencil testing. Also, testing became limited to core subjects.

A consensus among teachers was that, first, these changes narrowed the curriculum, not only to core subjects but to what was expected to be assessed for those subjects; second, it served to re-establish ability-group classrooms, so that teaching could focus on attainment targets considered to be appropriate to a student's ability; and third, children with special needs were being short-changed in some schools (Silvernail, 1996). Teachers also reported teaching to the test and cited instances of unethical elevation of test scores in some schools.

In 1993, the national exam was scheduled, test scores were to be reported to parents and raw scores for each class were to be publicized for all schools. Despite threats to withhold teacher pay, a teacher boycott of the tests was largely successful, and became generally supported by parents and the public (Black, 1994).

Implications for NAEP

The seeds for the National Assessment of Educational Progress were sown some 35 years ago by Ralph Tyler, Francis Keppel, and John Gardner. They were nourished by funds from the Carnegie and Ford Foundations, and in 1965, detailed planning was under way. For the next four years, Lee Cronbach, John Tukey, Ralph Tyler, Bob Abelson and I met often to establish specifications for a workable system for assessing precollege education in America. Recently, I published accounts of the basic principles that were agreed upon as a foundation for that system (Jones, 1996, 1997). Foremost among those: many exercises — far more numerous than reasonably could be presented to a participant in one or two hours — would be required to cover specified objectives and subobjectives for a given age, 9, 13, 17, or young adult, and subject area, math, science, or reading, for example. Also, no score or measure of achievement would be associated with any participating child, or indeed with that child's classroom or school. Were these features not to have been maintained, I believe that NAEP would have become so controversial that it would not have survived to be the useful indicator of educational progress that it is today.

In defending the proposed national tests, Secretary Riley recently stated in Congressional testimony, "We are simply taking the National Assessment of Educational Progress (NAEP) tests one step further. Right now, NAEP does not test all students, and it provides no information at all on individual students, schools, or districts. We want to change that, and that is why I call the new national tests a 'personalized version of NAEP'" (Riley, 1997b).

Clearly, as I've noted more than once, no single test can cover the breadth of content that has characterized NAEP with its matrix design. And, were the national test to be implemented, isn't it likely that the justification for NAEP as we know it would be likely to be rejected? If a state were to adopt Riley's "personalized version," that state surely would be reluctant to participate in the sampling version of NAEP. The unique values of NAEP as an indicator of long-term educational progress surely would be jeopardized.

Some Additional Concerns

About 20 years ago, there was a serious Congressional effort to adopt national testing. It was a well-intentioned effort, spearheaded by Congressman Albert Quie from Minnesota, who offered an amendment to the Elementary and Secondary Education Act. Quie proposed a reform of the formula by which schools receive Title I funding. Rather than having funds depend on proportions of students from poverty families, he proposed that funding be based on numbers of low scores on achievement tests. In that proposal as in the current one, test administration was to be entrusted to local school personnel.

In a rare bout of letter writing to public officials, I sent letters to Congressman Quie and to other Members of Congress. I noted that substituting a criterion of low achievement for a criterion of poverty was philosophically attractive, but was fraught with practical difficulties. Principals and teachers, parents and school children would be tempted to produce low test scores — by any of a variety of rather simple means — to increase federal funding for their school. Only at the last minute was that legislation withdrawn.

It is somewhat more difficult for schools to produce high test scores than low ones. Yet, as noted by Bob Linn (1996) and stressed by Lorrie Shepard (1997) in prior Angoff lectures, there is ample evidence that for high-stakes tests, artificially high scores sometimes can be generated. Cannell's "Lake Wobegon effect" (Koretz, 1988), where with repeated use of the same test, most children became above average, is a case in point. Just as is true for commercially developed achievement tests, the president's national test would not be immune from coaching. With increasing use of the Internet, where the unauthorized sharing of test items surely would occur, it would be feasible to elevate test scores without much effort. Even in the unlikely event that test items could be kept secure, violations of rules for test administration — providing extra time or giving clues to correct answers — probably could not be prevented. Resulting publicity of even rare violations certainly would impugn all results in the eyes of the public.

A Recapitulation

In recent years, the vast majority of states have initiated a variety of educational reforms. Results from NAEP and evidence from other sources suggest that these efforts are serving to improve teaching and learning and to elevate average student achievement.

No adequate justification has accompanied the proposal for national tests. The suggestion that test results would be useful to diagnose learning problems of specific children is not tenable. There are better instruments to meet that need, e.g., tests designed for that purpose, and evaluations conducted by classroom teachers and other school personnel.

It might be claimed that by adopting a national test, public interest in improving education will be sustained; yet, such interest seems unlikely to decline in the absence of a national test.

Some emphasize the importance of a common measure to compare achievement among individual students, classrooms, schools, and school districts. Scores on a single paper-and-pencil test, however, do not provide the basis for fair comparisons.

The comprehensive NAEP asssessments have been useful by providing results by state, and have proven especially useful to indicate change over time. A national test would threaten the continuation of NAEP.

Advisory groups consistently have recognized the importance of student assessments in support of educational reform. Without exception, they recommend multiple forms of assessment, and they urge that responsibility remain with states and local districts.

Evidence from abroad, especially from England and Canada, should lead us to be wary of high-stakes assessment based on a single test. Quite understandably, teachers and students will concentrate on content anticipated to be included in the restricted sample of items on the test. That can be detrimental to constructive teaching and learning.

With control of test administration and scoring vested in local school personnel, it must be expected that proper procedures would not always be followed, leading to legitimate questioning of results. In addition, test security might be almost impossible to maintain; cheating scandals could undermine the testing program.

Based on all these considerations, I conclude that, if they were implemented, the currently proposed national tests would do more harm than good. Such a test program is more likely to inhibit than to support constuctive educational reforms.

References

Baker, E. L. & Linn, R. L. (1997). National tests in reading and mathematics. *The CRESST Line*, Summer, pp. 1, 8-11.

Behrman, R. E., Ed. (1997). Children and poverty. *The Future of Children*. 7 (2), 1-160.

Bennett, W. J. (1997). School reform: What remains to be done. *The Wall Street Journal*, September 7.

Black, P. J. (1990). *Issues in public examinations, social and educational imperatives for changing examinations*. Paper presented at 16th International Conference, International Association for Educational Assessment, Maastricht, The Netherlands, June 20.

Black, P. J. (1994). Performance assessment and accountability: the experience in England and Wales. *Educational Evaluation and Policy Analysis*. 16, 191-203.

British Foreign and Commonwealth Office. (1990). *Education reform in Britain*. London: Central Office of Information.

Brooks-Gunn, J. & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children*, 7 (2), 55-71.

Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *Report in brief, NAEP 1996 trends in academic progress*. Washington, DC: National Center for Education Statistics.

Cuban, L. (1997). National testing: Wrong answer. *Los Angeles Times*, April 6.

Green, B. F. (1996). *Setting performance standards: Content, goals, and individual differences.* Princeton, N.J.: Educational Testing Service.

Jaeger, R. M. (1992). *"World class" standards, choice, and privatization: Weak measurement serving presumptive policy.* An address at the annual meeting of the American Educational Research Association, San Francisco, April 23.

Jones, L. V. (1988). Educational assessment as a promising area for psychometric research. *Applied Measurement in Education*, 1, 233–241.

Jones, L. V. (1991). Discussion of "The British experience with national educational goals and assessment." In *Educational Testing Service 1990 Invitational Conference proceedings, The assessment of national educational goals*. Princeton, NJ: Educational Testing Service, pp. 81–85.

Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher*, 25, No. 7, 15-22.

Jones, L. V. (1997). The National Assessment of Educational Progress, origins, and prospects. In National Academy of Education (Ed.), *Assessment in transition: Monitoring the nation's educational progress, background studies*. Stanford, CA: National Academy of Education.

KORETZ, D. M. (1987). *Educational achievement: Explanations and implications of recent achievement trends.* Washington, DC: U. S. Congressional Budget Office.

KORETZ, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator,* 12, 8-15, 46-52.

KRISTOF, N. D. (1997). Where children rule. *The New York Times Magazine,* August 17, 40-44.

LINN, R. L. (1995). *Assessment-based reform: Challenges to educational measurement.* Princeton, N.J.: Educational Testing Service.

LOFTY, J. S. (1993). Can Britain's national curriculum show America the way? *Educational Leadership,* 50 (5), 52-55.

MULLIS, I. V. S., MARTIN, M. O., BEATON, A. E., GONZALEZ, E. J., KELLY, D. A., & SMITH, T. A. (1997). *Mathematics achievement in the primary school years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

NATIONAL ACADEMY OF EDUCATION. (1993). *Setting performance standards for student achievement.* Stanford, CA: Author.

NATIONAL COUNCIL ON EDUCATION STANDARDS AND TESTING . (1992). *Raising standards for American education.* Washington, DC: U. S. Government Printing Office.

NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS . (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

NATIONAL RESEARCH COUNCIL. (1993). *Measuring what counts: A conceptual guide for mathematics assessment.* Washington, DC: National Academy Press.

NATIONAL RESEARCH COUNCIL. (1996). *The national science education standards.* Washington, DC: National Academy Press.

NATIONAL RESEARCH COUNCIL AND NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS. (1997). *Improving student learning in mathematics and science: The role of national standards in state policy.* Washington, DC: National Academy Press.

NUTTALL, D. L. (1990). *National curriculum assessment in the UK.* Paper presented at the annual meeting of the American Educational Research Association, Boston, April.

REESE, C. M., MILLER, K. E., MAZZEO, J., & DOSSEY, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states.* Washington, DC: National Center for Education Statistics.

RILEY, R. W. (1997a). *Voluntary national tests for reading and math.* Statement to the House Subcommittee on Early Childhood, Youth and Families, Committee on Education and the Workforce, April 29.

RILEY, R. W. (1997b). *Testimony of Secretary Richard W. Riley*. Statement to the Senate Labor, Health & Human Services & Education Subcommittee of the Senate Appropriations Committee, September 4.

ROTHBERG, I. C. (1995). Myths about test score comparisons. *Science*, 270, 1446-1448.

SHANKER, A. (1985). The revolution that's overdue. *Phi Delta Kappan*, 66, 311-315.

SHEPARD, L. A. (1997). *Measuring achievement: What does it mean to test for robust understandings?* Princeton, NJ: Educational Testing Service.

SILVERNAIL, D. L. (1996). The impact of England's national curriculum and assessment system on classroom practice: Potential lessons for American reformers. *Educational Policy*, 10, 46-62.

STAKE, R. E. (1995). The invalidity of standardized testing for measuring mathematics achievement. In Thomas A. Romberg, Ed., *Reform in school mathematics and authentic assessment*. Albany, N.Y.: State University of New York Press, pp. 173-235.

STOBART, G. (1990). *Issues in public examination of the national curriculum in England and Wales: A case of over-assessment?* Paper presented at 16th International Conference, International Association for Educational Assessment, Maastricht, The Netherlands, June 20.

U.S. DEPARTMENT OF EDUCATION. (1997). *Voluntary national tests*. Transcript from Public Meeting in Washington, DC, May 19.

U. S. GENERAL ACCOUNTING OFFICE. (1993). Educational achievement standards: NAGB's approach yields misleading interpretations. Report No. GAO/PEMD-93-12. Washington, DC: Author.

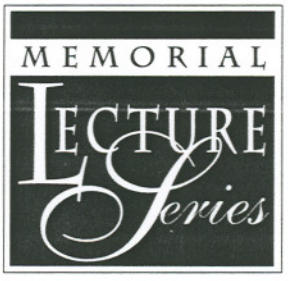
THE WHITE HOUSE. (1996). *Remarks by the President to the National Governor's Conference*. Washington, DC: The White House, March 27.

THE WHITE HOUSE. (1997a). *Remarks by the President in education town meeting*. Clarksburg, W.V., May 22. Washington, DC: The White House.

THE WHITE HOUSE. (1997b). *Remarks by the President at Oak Bluff School*. Martha's Vineyard, MA, September 3. Washington, DC: The White House.

THE WHITE HOUSE. (1997c). *Radio address of the President to the nation*, September 20. Washington, DC: The White Hiouse.

WIDEEN, M. F., O'SHEA, T., PYE, I, & IVANY, G. (in press). High-stakes testing and the teaching of science. *Canadian Journal of Education*.



ETS

POLICY INFORMATION CENTER
Educational Testing Service
Princeton, New Jersey 08541-0001

04202-13515 • S117M3•204927 • Printed in U.S.A