DOCUMENT RESUME

ED 351 386 TM 019 219

AUTHOR Koretz, Daniel M.; And Others
TITLE National Educational Standards and Testing: A Response to the Recommendations of the National Council on Education Standards and Testing. Congressional Testimony.
INSTITUTION Boston Coll., Chestnut Hill, MA. Center for the Study of Testing, Evaluation, and Educational Policy.; Rand Corp., Santa Monica, CA. Inst. for Education and Training.
REPORT NO CT-100
PUB DATE 19 Feb 92
NOTE 26p.
PUB TYPE Legal/Legislative/Regulatory Materials (090)

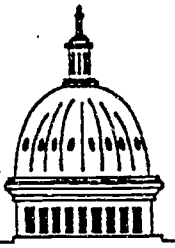
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Cost Effectiveness; Curriculum Development; *Educational Assessment; Educational Change; Educational Objectives; Elementary Secondary Education; *National Competency Tests; National Programs; School Effectiveness; Scoring; *Student Evaluation; Test Bias; *Testing Problems; Test Results; Test Use
IDENTIFIERS *National Council on Educ Standards and Testing; Standard Setting; Testimony

ABSTRACT

In January 1992, the National Council on Education Standards and Testing (NCEST) issued a report "Raising Standards for American Education," which called for the establishment of a national system of educational standards and assessments as a basis for comprehensive reform of U.S. education. This statement is a facsimile of the written testimony that was submitted to the House of Representatives subcommittee. The statement is presented as it appears in the Congressional record, with a preface and references added. It is contended that although the NCEST recommendations appear to be a matter of common sense, they are unlikely to work and may have serious side effects. The recommendations do not adequately address issues of feasibility, fairness, validity, and reliability. Although the new standards are expected to be a common core, the NCEST does not explain why the proposed tests will not narrow the curriculum. Proposed innovative assessments have not yet been adequately tested, and many practical issues must be resolved. Many problems exist in building a set of assessments that will do all that the NCEST proposes. These problems concern the following areas: (1) providing diversity and protecting local initiative; (2) providing comparability of scores; (3) equity; (4) costs; and (5) gauging school effectiveness. While efforts to move ahead with a national debate on educational standards must be encouraged, several alternatives to the proposals must be discussed to ensure feasibility, validity, fairness, and desirability of new assessment programs. (SLD)

CT-100

CONGRESSIONAL TESTIMONY



National Educational Standards and Testing: A Response to the Recommendations of the National Council on Education Standards and Testing

Daniel M. Koretz
RAND

George F. Madaus
Boston College

Edward Haertel
Stanford University

Albert E. Beaton
Boston College

Institute on Education and Training

RAND

Institute on Education and Training

RAND's Institute on Education and Training, established in 1991, performs policy analysis and research to help improve education and training for all Americans.

The institute examines all the forms of education and training that people may get during their lives. These include formal schooling from preschool through college; employer-provided training (civilian and military); postgraduate education; proprietary trade schools; and the informal learning that occurs in families, in communities, and with exposure to the media. Reexamining the most basic premises of the field of education, the institute goes beyond the narrow concerns of individual components to view the education and training enterprise as a whole. It pays special attention to how the parts of the enterprise affect one another and how they are shaped by the larger environment. The institute

- examines the performance of the education and training system;
- analyzes problems and issues raised by economic, demographic, and national security trends;
- evaluates the impact of policies on broad, system-wide concerns; and
- helps decisionmakers formulate and implement effective solutions.

To ensure that its research affects policy and practice, the institute conducts outreach and disseminates its findings to policymakers, educators, researchers, and the public. It also trains policy analysts in education and training.

RAND is a private, nonprofit institution, incorporated in 1948, engaged in nonpartisan research and analysis on problems of national security and the public welfare. The institute builds on RAND's long tradition—interdisciplinary, empirical research held to the highest standards of quality, objectivity, and independence.

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

PREFACE

At the end of January, 1992, the National Council on Education Standards and Testing (NCEST) issued its report, *Raising Standards for American Education*, which called for the establishment of a national system of educational standards and assessments as a basis for comprehensive reform of American education (National Council on Education Standards and Testing, 1992). On February 19, 1992, the Subcommittee on Elementary, Secondary, and Vocational Education of the Committee on Education and Labor, U.S. House of Representatives, held hearings at which the general issues of national standards and tests and NCEST's specific proposals were debated. Governor Roy Romer of Colorado, Co-Chair of NCEST (with Governor Carroll A. Campbell, Jr., of South Carolina), testified in support of the NCEST recommendations. We prepared a critique of the NCEST proposals, which Dan Koretz presented to the Subcommittee. The statement that follows this preface is a facsimile of the written testimony that we submitted to the Subcommittee and that appears in the Congressional Record, with the exception that we have added references.

NCEST was established by act of Congress (Public Law 102-62) to "provide advice on the feasibility and desirability of national standards and testing in education." In its report accompanying the authorizing legislation, the Committee identified a wide range of specific questions that NCEST should address (House Report 102-104, 1991, pp. 5-8). In a letter sent to other NCEST members during the Council's deliberations, Senator Orrin G. Hatch and Representatives William F. Goodling and Dale E. Kildee, three of the four Congressional members of NCEST, reminded the other members of the Council of nine specific questions NCEST was charged with addressing, including:

- The "benefits and liabilities...of imposing uniforms national standards, a national test, or a system of national examination[s]";
- "Whether there is any evidence that national education standards and a national test or system of national examinations promotes improvements in educational achievement...";
- "Whether uniform national standards are appropriate when there are wide variations in the resources available to school systems...";
- "Whether national tests or examinations that are intended for instruction[al] improvement can be used for unintended purposes (e.g., sorting and tracking of students)";

- "Whether, given the wide variations in the levels of student performance across the country...it is feasible to develop education standards which challenge all students to do their best without penalizing those with lesser educational opportunity"; and

- "At a minimum, the Council is *required* to assess the feasibility of an appropriate system of national tests or examinations in the context of four factors--validity, reliability, fairness, and cost. The Committee's intent is that such a system of tests, if it cannot pass muster on *each* of these factors, would not be 'appropriate'." [Emphasis added.]

In its report, NCEST strongly endorsed a national system of educational standards and assessments, arguing that they are both desirable and feasible. The Council suggested a system of two components, both aligned with national standards: one to measure the performance of individual students, and the second comprising "large-scale sample assessments, such as the National Assessment of Educational Progress" (National Council on Education Standards and Testing, 1992, p. 4) The system would have three other attributes as well: it would consist of multiple "methods" of measuring progress rather than a single test; it must be "voluntary, not mandatory"; and it must be "developmental, not static" (National Council on Education Standards and Testing, 1992, p. 4).

Although the NCEST report was necessarily quite general, some of the essential elements of its proposed system of examinations were made clear:

- Although they might not begin as such, the assessments would become external, "high-stakes" examinations (that is, examinations with serious consequences for students or educators);

- The examinations could serve a variety of fundamentally different functions while providing comparable information at all levels of the educational system; and

- The examinations were expected to have a very powerful salutary effect on American schools.

The Council's desire for external, high-stakes examinations is explicit. In the summary to its report, the Council noted:

> "The Council finds that the assessments eventually could be used for such high-stakes purposes for students as high school graduation, college admission, continuing education, and certification for employment. Assessments

> could also be used by states and localities as the basis for system accountability" (National Council on Education Standards and Testing, 1992, p. 5).

Later in the report, the Council clarified that the use of these tests for accountability, far from being an open question, was a key element of their justification. For example:

> "The Council finds a need to shift the basis of educational accountability away from measures of inputs and processes to evidence of progress toward desired outcomes....A nationally coordinated initiative would result in high-quality outcome measures that can be used for accountability....Unprecedented national attention...could be focused on the system of assessments and use of the results for accountability" (National Council on Education Standards and Testing, 1992, pp. 17-18).

In this context, what are "voluntary" examinations? The report did not fully clarify for whom the assessments would be voluntary and for whom mandatory, but high-stakes examinations used for accountability cannot be voluntary for those held accountable. We infer that a system in which the examinations would be voluntary for governors or chief state school officers but obligatory for all students and teachers working under their authority would pass NCEST's muster.

As the quotes above indicate, the Council's expectations for the new examination system were not modest. NCEST expected the examinations to be able to serve very different functions, such as certifying that students have met the minimum requirements for high school graduation, measuring the readiness of the most able students for selection into elite universities, and establishing that individuals have the skills required for diverse entry-level positions in the work force. Moreover, the Council expected the examination system to measure the performance of educational systems as well as the achievement of students. The Council also expected the system to produce "comparable" results at all levels of the educational system, from individual students to the nation as a whole (National Council on Education Standards and Testing, 1992, p. 15).

Lastly but most important, NCEST concluded that new national standards examinations could have sweeping positive effects on education while avoiding the often serious negative effects of previous systems of test-based accountability. The Council argued that "standards and assessments linked to the standards can become the cornerstone of the fundamental, systemic reform necessary to improve schools" (National Council on Education Standards and Testing, 1992, p. 5). Indeed, NCEST maintained not only that the system could serve

this function, but apparently that nothing else could substitute: "[establishing] national content and performance standards and assessments of the standards....will constitute *an essential next step* to help the country achieve the National Education Goals [emphasis added]" (National Council on Education Standards and Testing, 1992, p. 5). Its report noted in passing that "care must be taken to avoid the unintended negative and undesired effects of some testing practices, such as narrowing instruction and excluding certain students from assessments," but this essential and difficult step is to be accomplished only by unspecified "sufficient safeguards" (National Council on Education Standards and Testing, 1992, p. 29).

In our view, the NCEST report holds out a false and even dangerous promise. Historical experience and research evidence suggests that the Council's proposals are unlikely to meet the Council's expectations for positive effects and could indeed have serious negative effects on students, teachers, and the educational enterprise. The Council gave insufficient attention to this evidence and to the basic questions--such as validity, fairness, feasibility, and cost--the Congress specifically charged it with addressing. Some of the bases of our concerns are briefly spelled out in the following testimony.

STATEMENT OF

Daniel M. Koretz
Senior Social Scientist
RAND

George F. Madaus
Boisi Professor of Education
Director, Center for the Study of Testing, Evaluation, and Educational Policy
Boston College

Edward Haertel
Associate Professor, School of Education
Stanford University

Albert E. Beaton
Professor of Education and Senior Research Fellow,
Center for the Study of Testing, Evaluation, and Educational Policy
Boston College

before the

Subcommittee on Elementary, Secondary, and Vocational Education
Committee on Education and Labor
U.S. House of Representatives

February 19, 1992

Preparation of this testimony was supported by the RAND Institute on Education and Training and The Center for the Study of Testing, Evaluation, and Educational Policy at Boston College. The opinions expressed here, however, are solely those of the authors.

Mr. Chairman and members of the Subcommittee, thank you for the opportunity to speak with you about the recent report of the National Council on Education Standards and Testing (NCEST) and about the critical issues of national standards and tests.

Today I will present a statement prepared jointly with Dr. George Madaus and Dr. Albert Beaton, both of the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College, and Dr. Edward Haertel of Stanford University. Dr. Madaus and I both served on NCEST's Assessment Task Force. Drs. Madaus, Beaton, and Haertel are nationally recognized scholars of educational measurement, and I am privileged to present this joint statement on their behalf. Drs. Madaus and Beaton are here today; Dr. Haertel was unable to attend.

We share some of the premises and goals of the National Council, and we commend the report for giving them a rare prominence in political debate. Standards are too low in many American schools. We share the Council's concern that expectations have been especially low for groups that have historically done poorly in school. A nationwide debate on educational standards, carried out properly, could have substantial positive effects on education and should begin promptly.

Nonetheless, we are deeply troubled by the NCEST report and the policies it recommends. Although the NCEST recommendations may appear commonsensical, they are unlikely to work, and they may well have serious negative side effects. Moreover, in our view, the NCEST report does not adequately address key issues of feasibility, fairness, validity, and reliability--precisely the issues emphasized in the Congressional charge to the Council.

We will use the limited time today to discuss only a few critical aspects of the NCEST recommendations, and thereafter we will discuss very briefly a few alternatives. We believe that standards and assessment can play a key role in educational reform, but to fulfill that potential and avoid negative side-effects, they must be used in ways quite different from NCEST's recommendations.

A HISTORICAL PERSPECTIVE ON THE NCEST RECOMMENDATIONS

We know from many years of research that the problems of American education have many causes and are often very difficult to ameliorate. Yet NCEST's recommendation is far simpler than these difficult problems warrant: standards linked to "high-stakes" tests that have serious consequences for individuals. True, the report notes that tests and standards are not "panaceas" and lists in passing some other needed elements of reform, such as improved professional development and "the reduction of health and social barriers to

learning" (The National Council on Education Standards and Testing, 1992, p. 7). But it puts its faith in tests, maintaining that tests and standards "can be the cornerstone of the fundamental, systemic reform necessary to reform schools" (The National Council on Education Standards and Testing, 1992, p. 7). Moreover, its specific proposals for national action are largely limited to tests and standards. Most of the remaining and more difficult aspects of school improvement, such as professional development and family and community supports (The National Council on Education Standards and Testing, 1992, p. 7), are left for states and localities, and the report offers no specific proposals for dealing with them.

Using test-based accountability to drive education is hardly a new idea. This approach has been tried many times over a period of centuries in numerous countries, and its track record is unimpressive . Most recently, it was the linchpin of the educational reform movement of the 1980s, the failure of which provides much of the impetus for the current wave of reform, including the Council's report (see Madaus, 1988 and Madaus and Kellaghan, 1992 for a general review of this history). Holding people accountable for performance on tests tends to narrow the curriculum. It inflates test scores, leading to phony accountability. It can have pernicious effects on instruction, such as substitution of cramming for teaching. Evidence also indicates that it can adversely affect students already at risk--for example, increasing the dropout rate and producing more egregious cramming for the tests in schools with large minority enrollments (see, e.g., Darling-Hammond, 1991; Jaeger, 1991; Koretz, Linn, Dunbar, and Shepard, 1991; Madaus, 1991; Shepard, 1991; Haertel, 1989; Haertel and Calfee,1983; Haertel, Ferrara, Korpi, and Prescott, 1984; and Stake, 1991). NCEST has proposed some departures from current testing practice in the hopes of doing better, but those departures are unproven and hold as much peril as promise.

Despite its Congressional charge, the Council report did not discuss the evidence about test-based accountability, but it acknowledges past problems fleetingly in asserting that "care must be taken to avoid the unintended and undesired effects of some testing practices, such as narrowing instruction and excluding certain students from instruction" (p. 29). Indeed, but how? The burden rests with the Council to explain why we should have confidence that this hoary prescription will work so much better this time. In our view, the Council has not made a persuasive case.

By analogy, it is as though the Council came to you and said: we want your support to medicate 9 million children a year because we don't like their being ill, but the medicine we propose using bears an unsettling similarity to some that have failed and had serious side effects in the past, and we have not yet finished designing the new

medicine or testing it. In no field other than education would we consider, let alone accept, such a proposal.

HOW WOULD THE NCEST APPROACH DIFFER FROM PAST PROGRAMS?

The NCEST report proposes a system of testing that differs in several respects from current practice. Specifically:

- The standards would not be a curriculum that would constrain what people teach; rather, they would be a "core" upon which educators would elaborate.
- The proposed system would rely substantially on different types of tests: "performance assessments."
- NCEST is proposing a system of tests that would be under local, state, or "cluster" control and would be free to differ but that would nonetheless be linked to national standards.
- A new entity, the National Education Standards and Assessment Council (NESAC), would provide "quality assurance."

The Council maintains that these four attributes will make the old prescription of test-based accountability work better than previously. We are skeptical of all four.

WILL THE NEW STANDARDS BE JUST A "CORE?" THE RISK OF NARROWED INSTRUCTION

Advocates of the new testing often say that "what you test is what you get." Their first premise is that when tests are made to matter, people teach what yc test. Their second premise is that we can get them to teach better by testing better things and giving the tests serious consequences. This is the core logic of the NCEST report.

We agree with the first of these premises. The historical and research evidence, both in the United States and elsewhere, consistently shows that when people are held accountable for performance on tests, teachers focus on the tests' content. The problem is that to spend more time teaching what is tested, teachers generally spend less time teaching what is not tested, and what they give up is often important (see Madaus, 1988 and Madaus and Kellaghan, 1992 for a general review of this issue).

That is, external tests coupled with serious consequences generally narrow the curriculum. The NCEST report deals with this

problem only obliquely. It maintains that the new standards will be only a "common core" that would be enhanced by local elaboration. Perhaps, but the essential problem lies not with the standards but with the tests that would be used to implement them, and the Council report does not explain why these tests will not narrow the curriculum.

One consequence of narrowed instruction is the inflation of test scores. That is, students do much better on a particular test than their actual mastery of the subject matter warrants, so the public is misled about students' and schools' performance. This inflation of scores appears widespread, and our own research indicates that it can be egregious (Koretz, Linn, Dunbar, and Shepard, 1991).

Inflation of test scores should have been a central concern of the NCEST report for two reasons: it undermines the validity of the tests, and it makes a mockery of the Council's goal of greater accountability. First, if a test suggests that students know more than they actually do, it is ipso facto invalid. Second, when scores are inflated on a high-stakes test, students and teachers are indeed held accountable, but policymakers, such as chief state school officers and governors, are let off the hook because performance appears better than it really is. (See Koretz, 1988; Linn, Graue, and Sanders, 1989; Cannell, 1987; and Cannell, 1989 for a discussion of test-score inflation).

Despite the critical importance of these issues, the NCEST report had little to say about them, apart from holding out a vague promise of new types of tests. To what extent can current technologies help counter the problem? How should inflation of scores be detected? What should NESAC do about the inflation of scores when deciding whether to certify a test? The Council report is mute.

WILL NEW TYPES OF ASSESSMENTS SOLVE THE PROBLEM?

The NCEST report pins its hopes on the proposed use of "innovative" types of performance-based assessments. These tests, we are told, will focus on "higher order or complex thinking skills" (p. 28) and will be "worth teaching to" (p. 6).

We do not oppose the development of innovative forms of assessment; indeed, some of us have been involved in that effort. It is essential, however, to temper our enthusiasm with a bit of realism.

First, what NCEST proposes is in some respects innovation, but in other ways it is a return to the past. Over the centuries, testing programs have evolved and become more standardized for reasons of cost, practicality, administrative convenience, and a desire for comparability and objectivity . There are ample grounds for criticizing

current objective tests--and in particular the misuse of such tests that recently has become commonplace. But to change direction and to downplay concerns such as objectivity and comparability for the sake of other goals--say, richness of assessment or better incentives for teachers--will confront us with very serious difficulties that the Council report alludes to but does not adequately address (Madaus, in press; Madaus and Kellaghan, in press; Madaus and Kellaghan, 1992; Madaus, in press; Madaus and Tan, in press).

Second, to the extent that the proposed assessments really are innovative, they are in many cases unfinished and untested. They are at a stage where they are ripe for a serious R&D effort, complete with rigorous evaluation, but they are not yet ready to be a linchpin of national policy. There are many practical, technical, and infrastructure issues that must be resolved before such techniques can safely be deployed as policy instruments on a large scale in schools. We do not believe in flying an airplane while building it, particularly when the passengers are children.

Third, the field already has considerable knowledge of how certain performance assessments have worked, and the evidence to date suggests the need for caution (Madaus and Kellaghan, in press; Nuttall, 1992; Shavelson, Baxter, and Pine, 1991, 1992; Dunbar, Koretz, and Hoover, 1991; Linn, Kiplinger, Chapman, and LeMahieu, 1991).

To put the evidence about performance assessments into perspective, we need to recall what tests are. Regardless of the types of exercises it comprises--multiple-choice questions, essays, experiments, performance tasks, or whatever--a test is only a sample of student knowledge. It is useful only if we can generalize from performance on the test to a broader domain of interest, such as "mastery of algebra" or "understanding the nature and process of science." One key to reliable and valid generalization is having an adequately large and representative sample of exercises.

If we cannot draw valid generalizations from the exercises on the test, two things follow: the public is misled about student performance, and--if the test has real consequences--people are treated capriciously. That is, some students who fail on the basis of one overly limited or non-representative sample of tasks would have passed if given an equally defensible alternative set. In other words, both validity and fairness are undermined (Shavelson, Baxter, and Pine, 1991; 1992; Dunbar, Koretz, and Hoover, 1991; Linn, Kiplinger, Chapman, and LeMahieu, 1991).

Evidence suggests that student performance often generalizes poorly across related performance tasks. The quality of student essays, for example, varies markedly depending on the type of essay

required and even the specific prompt used. Similar findings have come from recent investigations of hands-on science tests (Beaton, 1988; Shavelson, Baxter, and Pine, 1991, 1992; Dunbar, Koretz, and Hoover, 1991; Linn, Kiplinger, Chapman, and LeMahieu, 1991). In the case of multiple-choice and short-answer tests, it is relatively easy to deal with problems such as this by adding additional questions. Performance tasks, however, tend to be costly to produce and time-consuming to administer and score, so using a large set of exercises is less practical (Nutall, 1992; Madaus and Kellaghan, 1991a, 1991b, in press).

One partial solution to this problem, if we are willing to pay the large development costs, would be to use a large set of performance tasks across a large group of students--say, all those in a state or large district--but to administer only one or a few to each child. This approach, however, generally does not give us valid and reliable scores for individual children. The NCEST report calls for comparable tests at all levels down to the individual student and for serious consequences for individual students, but it leaves us in the dark about how this is to be accomplished with fairness, validity, reliability, and at reasonable cost using complex performances that show limited generalizability.

The validity of performance assessments can also be undermined by teaching to the test. NCEST argues that the new tests will be designed to be taught to. By this, proponents usually mean that teaching to these new tests will itself be good instruction. Perhaps, but that is only one of the reasons to be worried about teaching to the test. The other, as we have already noted, is that teaching to the test narrows instruction, thus inflating scores and undermining validity. To our knowledge, there is no evidence that performance tests are less susceptible to this problem than conventional tests, and there are some indications that they are more susceptible.

Switching to new forms of assessment is also unlikely to help make the system more equitable. Some proponents argue that switching to performance assessments will lessen inequities in the current testing system. Again we know of no evidence substantiating that claim, and some evidence suggests the reverse.

Finally, the proposed reliance on new performance assessments raises serious questions of feasibility that the Council did not address. Great Britain's recent experiment with Standard Assessment Tasks (SATs) provides an illuminating example. Among the specific issues that arose in England and Wales were the need for extra support and staff in schools, the need for procedures to minimize the disruption of school and classroom organization, the difficulty (and perhaps undesirability) of imposing standardized conditions of administration that would permit comparability of results across schools, and the

difficulties inherent in rating large numbers of performance-based tasks (Nutall, 1992; Madaus and Kellaghan, in press).

CAN WE BUILD A SYSTEM OF ASSESSMENTS THAT CAN DO WHAT NCEST PROPOSES?

The NCEST proposal calls for a system of examinations that will serve an extraordinarily wide range of functions. Among them are providing comparable information across jurisdictions and at all levels of aggregation; providing incentives to educators to teach better and to students to work harder; and providing valid predictive information to select students for further education or employment. Moreover, this system of tests would be "bottoms-up;" only the standards to which they are somehow linked would be "top down." Many of the Council's expectations for this system are unrealistic, and some are mutually contradictory. Moreover, the Council's list of functions focuses primarily on secondary school students, leaving unaddressed the difficult question of the uses and consequences of the tests the Council recommends administering in the elementary grades.

Providing Diversity and Protecting Local Initiative

We believe that the degree of diversification needed to preserve local initiative is far greater than the Council envisioned. Absent that diversity, the system will act as a national examination and will not be able to avoid the pernicious consequences of large-scale, external, high-stakes examination systems. European systems have often been suggested as an exemplar, but they entail more diversity than is commonly recognized. In the former West Germany, for example, with a population roughly comparable to that of California, New York, Ohio and Missouri, there are eleven state Ministries for Education that set the separate Abiturs for each state. In France, with a population about the size of that in California, New York, and Ohio, there are 23 separate academies that set the Brevet de college exams and the Baccalaureat exams for each academy. For the United States, a comparable level of diversification would imply the creation of between 18 and 40 examining boards for the secondary level alone. And with over 15,000 school districts, it is questionable whether even this level of differentiation would give teachers, schools, or communities an effective voice in curriculum (Madaus and Kellaghan, 1991a, 1991b).

Providing Comparability of Scores

Permitting a great deal of diversification would also run counter to the Council's assertion that comparability of results is essential. This is one of the most glaring and puzzling holes in the Council's position. The new tests are supposed to be different but yet the same. How different, and in what sense comparable? The Council never

clarified explicitly what it means by "comparable," but it appears to want a rigorous standard of similarity. After all, if tests are going to be used to determine high school graduation, admission to college, or employment (p. 5), they had better measure very similar knowledge and skills. Otherwise, the system will be inherently inequitable, capriciously favoring students taking one exam over those taking another.

Yet at the same time, the new system is supposed to protect local autonomy and encourage development of diverse local curricula. This implies that the tests linked to these curricula would have to be quite different. The Council report seems to be saying that the way out of this apparent contradiction is that the tests and curricula will be the same in some "core" respects having to do with national standards but will be free to differ in other respects. This notion is intriguing but still lacks any real substance, and it would be prudent to wait to see what this actually means before making it a cornerstone of national policy.

Research gives us even more reason to be cautious: even fairly minor differences between tests can produce fundamental differences in their results (e.g., Beaton and Zwick, 1990; Koretz, 1986). These differences can be a matter of equity; for example, changing tests can alter the apparent size of differences among racial and ethnic groups or between males and females. Moreover, teaching to the test--an explicit goal of the Council report--can greatly exacerbate differences in performance from one test to another.

Equity

We have serious concerns about inequities in the proposed system. NCEST recognized two potential sources of inequity, but we are not satisfied by its responses to either.

First, students cannot be held fairly to the same standard of performance if they are given unequal opportunities to learn. The NCEST report acknowledges this in its call for "school delivery standards," but it would let others figure out what these ought to be, and it says nothing about the resources that would be needed to attain reasonable equity.

Second, many low-achieving students face barriers to achievement that lie outside of school, and the NCEST proposal does not seriously address them. For example, as the Council's own Implementation Task Force noted, "Students whose most elemental shelter, food, and nurturing needs are not met are not going to perform at the minimum competency level consistently, much less demonstrate...world class academic standards. The health...system...must be restructured to better meet the needs of the

poor..." (p. G-11, emphasis added). That message appears to have been lost between the Task Force report and the final NCEST report. Many statistics show how poorly this nation addresses those elemental needs--e.g., figures on our appalling infant mortality rates, our high rate of child-poverty, the large numbers of children without access even to basic health and dental care, and so on. Yet the Council report proposes nothing specific to alleviate those inequities.

NCEST provided no persuasive rationale for using a national mechanism for holding students and teachers accountable but not for holding states and localities accountable for providing them with equitable support or educational services. Why can we trust the states to worry about delivery standards but not student performance standards? This would reverse the traditional division of responsibilities, in which the federal government's interventions in education have often been designed to create equity in the delivery of services.

Costs

The likely costs of the proposed system should also give us pause. The Council says that detailed cost estimates are unavailable but that the new assessment system should not add to the net burden of testing (p. 31). There are bases for estimating both financial and other costs of the new system, however, and it is readily apparent that the proposed system would add a very large burden indeed.

One basis for a rough cost estimate is the College Board Advanced Placement (AP) Examinations, often cited as an exemplar and noted in the Council report. AP exams currently cost $65 per subject, or $325 per student for the five-subject battery proposed by NCEST. (By contrast, a commercial standardized test battery costs about $2 to $5 per student.) Assuming that economies of scale are offset by the cost of increased reliance on performance tasks, this suggests a cost of more than $3 billion per year for testing only in the three grades suggested by NCEST.

But that is not the worst of it. It may not be practical to limit exams to three grades if they are to guide instruction (rather than simply weeding out less able students, which is the primary function of exams in some other countries). Certainly an AP-style exam system would require more frequent testing; AP exams are tightly linked to one-year course syllabi. So, let's say we test in six grades instead of three; that raises the ante to over $6 billion per year.

The AP exams are not the only basis for expecting high costs. Recent experience in Europe suggests costs of about $135 per student just for scoring 4 to 5 essay exams, each comprising 4 to 6 questions, when the exams are graded by teachers and include no performance

tasks. Our own National Assessment, which requires a mere hour per student and has traditionally been mostly multiple-choice and machine-scorable, has cost roughly $100 per student (Madaus and Kellaghan, 1991a, 1991b).

Who is going to provide these billions of dollars? Perhaps even more important, if that amount of money were made available, would examinations be the most effective way of spending it? Neither of these issues is addressed by the Council report.

The non-financial costs of the proposed system are likely to be substantial as well. One cost will be forgone instructional time. For example, in Great Britain, the recent administration of SAT's was supposed to extend over three weeks, but some local education agencies maintained that it would take them more than six weeks. In other parts of Europe, preparation for exams stretches over a period of three months to a semester. Advocates of the new testing assert that the new exams would be so challenging that preparation and testing time would be good instruction, not time taken away from instruction. Long experience to the contrary suggests that we should wait for substantiation before accepting this assertion, especially given that our nation already has an unusually short instructional year. Experience also suggests that innovative examination systems will require time-consuming and expensive inservice training in addition to substantial teacher time for preparation, administration, and logistics (Nuttall, 1992; Madaus and Kellaghan, in press).

Gauging the Effectiveness of Schools

Finally, the new system is supposed to be able to "provide evidence about the success of schools, local school systems, states, and the Nation in bringing all students...to high performance standards" (p. 13). Here we need to draw a distinction that the Council did not. It is entirely feasible to build a system that will monitor the progress of students at various levels of educational organization. The National Assessment does that for the nation as a whole and is experimenting with doing so for states, and there is no reason why that system could not be expanded and enriched.

Such data, however, generally tell us little about the effectiveness of schools or systems. They tell us which groups are doing better, but not why (see, e.g., Koretz, 1991). Current assessment systems, and the new system proposed by the Council, simply provide the wrong sort of data to evaluate programs. To evaluate educational programs, one must be able to rule out plausible alternative explanations of performance differences. This is the same, simple standard used in all of empirical science. To do this with educational programs requires collecting extensive, high-quality data on factors (such as family background) that exert powerful influences

on achievement, tracking the movement of students among educational systems, and, in most cases, tracking changes in student performance over time. None of this can be done by testing students in one grade out of four and aggregating to the district or state level. Senator Moynihan's recent tongue-in-cheek argument that states' scores are caused by their distance from the Canadian border was a humorous but powerful reminder of the riskiness of ignoring these simple rules of scientific inference.

WILL NESAC PROVIDE QUALITY ASSURANCE?

The Council report and now S. 2 would give NESAC major responsibilities for quality assurance, such as establishing guidelines for developing assessments, gauging their validity, and ensuring their comparability. The report calls for NESAC to guide the certification of standards and assessments.

If NESAC is established as proposed by the NCEST and S. 2, however, its certification of assessments would be a sham. Discharging its responsibilities would require substantial substantive and technical expertise, but the recommendations do not call for the appointment of even a single individual with expertise in measurement or evaluation. NESAC would also lack needed independence; its members would be appointed by the new National Education Goals Panel, and certifications would be made jointly by both organizations.

Equally important, the NCEST recommendations show a thorough misunderstanding of what is needed to validate tests and monitor their effects. A test cannot be validated by asking a group of individuals to examine its content, as the NCEST report implies. Moreover, validation is an ongoing process, not a one-time effort. To validate a test requires substantial empirical research, and the NESAC model does not make provisions for commissioning, funding, or using the needed investigations (Madaus, 1992).

To take one example, suppose that a test is used to screen individuals for employment. Then it is essential--for reasons of law and equity, apart from simple ethics--to demonstrate that performance on the test predicts performance on the job. This requires empirical research. The need for such research, although long established in law as well as in the measurement profession, is ignored in the NCEST recommendations.

To take one other example that we alluded to earlier: if a test is used for accountability or is for other reasons "taught to," as the Council report explicitly recommends, how does one know whether scores on the test have been inflated enough to undermine their validity? Having a committee such as NESAC examine the test to see

whether it lives up to "world class standards" will not provide a clue. Again one needs research, for example, random substitution of similar tests in ongoing testing programs. Once again, there is no provision in the NCEST recommendations or S. 2 for that type of validation.

WHAT ARE SOME ALTERNATIVE DIRECTIONS?

Despite our criticisms of the assessment system proposed by NCEST, we do believe that standards and assessments should play a role in education reform. Contrary to the accusation leveled at opponents of the NCEST report by one Council member before this Subcommittee, we do not believe that "it is better to stick with the discriminatory and educationally destructive current testing technology rather than invest in developing the new technical capacity we will need for the program outlined [by NCEST]" (Resnick, 1992). In fact, all of us have been strong critics of the present system. The question is how to do better.

Establish Standards and Curricula

We endorse the proposal to move ahead with a national debate on educational standards. This effort must go beyond generally worded standards to include the development of curricula specific enough to guide teaching and assessment. These must be the first steps; a syllabus-based examination system will have to wait until standards are established, because we cannot insure that students have a fair chance to learn what is tested until we have curricula in place.

There is more to establishing standards, however, than the NCEST proposal envisions. If we want standards that reflect, for example, skills and knowledge that are needed for certain types of jobs or for certain types of postsecondary education, we will need to validate the standards and confirm empirically that the standards actually reflect what is needed.

Support Research, Development, and the Building of Infrastructure

A serious R&D effort is precisely what is needed if we are to answer the questions of desirability, feasibility, validity, practicality, fiscal costs, opportunity costs, educational costs, and consequences that are raised by the NCEST proposals but not seriously addressed in the NCEST report.

This R&D effort must go far beyond the design of new assessments, and it will take considerable resources and time. We must, for example:

- Conduct serious empirical research on the quality of new performance assessments--for example, the reliability of scores and the extent to which they generalize enough to be meaningful.

- Conduct investigations of costs, including non-financial and indirect costs.

- Conduct research into the effects of new types of assessments--effects on quality of instruction, learning, school organization, and equity.

- Build an infrastructure capable of supporting new assessment systems.

The R&D effort would need to focus on the context and use of the new assessments. Assessments are not good or bad in the abstract; their quality depends on how they are used. A test may succeed in providing one type of information and fail utterly to provide another; it may be beneficial if used in one way and pernicious if used differently.

Learn from Smaller-Scale Implementations

During the NCEST deliberations, Governor Romer repeatedly expressed concern that states and localities are getting ahead of the national effort. State and local efforts should be seen as a gift, not a cause for anxiety. If states and localities can be encouraged to couple their assessment innovations with serious evaluations--as Vermont is now doing--their efforts will provide an invaluable source of information about what works and what doesn't, and this information can improve national efforts. Indeed, the national program should include active encouragement of diverse smaller-scale efforts, but these efforts, unlike most of the recent innovative assessment efforts with which we are familiar, must be coupled with adequate evaluation. We believe that this approach can be accommodated easily within the framework of H.R. 3320.

Specify, Implement, and Document Other Components of Reform

The NCEST report says that tests and standards are not a panacea, and Governor Romer has often noted that they are only the bread of the sandwich, worth very little without the filling. We concur, and we believe that it is time to work on the filling. First, we need further clarification of "school delivery standards:" what specifically must schools provide before we are willing to say that opportunities are equitable? Second, we need to specify what the other components of reform will be and how they will be implemented. It is in our view simply unacceptable to hold students accountable for their

performance without providing them the opportunities they need to succeed on our examinations. Third, we need to develop and evaluate the indicators of equal opportunity that the NCEST proposal presumes. We know that simple, conventional measures such as teachers' years of experience will not suffice.

Establish Workable Procedures for Evaluation

As we have already noted, we believe that the proposed NESAC would not be capable of evaluating the new standards and examinations meaningfully. We see the need for an independent, non-partisan body with sufficient expertise and credibility to evaluate the technical qualities of alternative assessments, examine the evidence about their feasibility and costs, monitor the consequences of their use, and judge the comparability of results across the various local and state components of the assessment system (Madaus, 1992) To be effective, such a body would need to differ from the proposed NESAC in many ways:

- Its members must have the needed technical and substantive expertise in measurement, evaluation, and education.
- It must be independent.
- It must have the authority and funding to commission extramural research as needed.
- Its charter must call for realistic validation and evaluation of examinations as they are actually used in specific contexts.
- Its charter must call for evaluation of the effects of assessment programs on schooling and learning.
- Its charter must call for evaluation of the effects of the programs on diverse groups of students, particularly the disadvantaged.
- Its charter must call for ongoing evaluations of the strengths and weaknesses of assessment programs, rather than unrealistic, one-time, up-or-down "certification" decisions.

Regardless of decisions about NESAC, evaluation and validation efforts should be built into federally supported education reforms that use assessment. During the 1980s, very few jurisdictions using test-based accountability evaluated the effects of their programs, and some flatly refused outside evaluations. We should not allow this to be

repeated. People who want to experiment with dangerous medicines should be required to evaluate the impact of their experiments. We suggest that all reform efforts funded by H.R. 3320 or other similar legislation be required to do the following if assessment is to be a part of the reform:

- Require grantees to specify how tests will be used--in particular, how, if at all, they will be used for accountability at any level of aggregation (from students to states or clusters). Accountability need not entail concrete sanctions; it can be sufficient to publicize scores as an index of performance.

- Require that grantees specify what evidence of reliability and validity will be collected if innovative assessments will be used.

- Require grantees to specify what steps, if any, they will take to lessen the risks of inflated test scores and narrowed instruction.

- Require that grantees planning to use test-based accountability evaluate its effects on instruction, its effects on diverse groups of students (in particular, the disadvantaged), and the possible inflation of test scores.

Because many school districts will lack the expertise to structure reasonable evaluations, it may be helpful to encourage partnerships, for example, between school districts and universities, for these purposes.

Decide About a National System

When all of the steps above have been taken, we will have developed standards and curricula; we will have produced sorely needed information about the feasibility, validity, fairness, and desirability of various types of new assessment programs; and we will have put into place a mechanism that can help protect against abuse as well as inadvertent harm. At that time, the nation will be far better prepared to make reasoned decisions about a possible national assessment system such as the one proposed by NCEST.

We cannot know now precisely what directions will be suggested by information that we do not yet have. The extensive evidence already in hand, however, suggests that a key part of the answer will be to use assessment as a partner in reform, not as its primary engine.

Mr. Chairman, thank you again for the opportunity to discuss these issues with the Subcommittee. At this time, we would be pleased to answer questions.

REFERENCES

Beaton, A. (1988). *Expanding the New Design: The NAEP 1985-86 Technical Report.* Princeton: Educational Testing Services.

Beaton, A., E., and Zwick, R. with forward by John W. Tukey (1990). *The effects of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* Educational Testing Service, Princeton, NJ.

Cannell, J. J. (1987). *Naturally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above The National Average.* Daniels, WV: Friends for Education.

Cannell, J. J. (1989). *The "Lake Wobegon" Report: How Public Educators Cheat on Standardized Achievement Tests.* Albuquerque, NM: Friends for Education.

Darling-Hammond, L. (1991). The implications of testing policy for educational quality and equality. *Phi Delta Kappan, 73*(3), 220-225.

Dunbar, S. B., Koretz, D., M., and Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education,* 4(4), 289-304.

Haertel, E. (1989). *Student achievement tests as tools of educational policy: Practices and consequences.* In B. Gifford (Ed.), *Test Policy and Test Performance: Education, Language and Culture.* (pp. 25-50). Boston: Kluwer Academic Publishers.

Haertel, E., and Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*(2), 119-132.

Haertel, E., Ferrara, S., Korpi, M., and Prescott, B. (1984). *Testing in secondary schools: Student perspectives* American Educational Research Association.

Jaeger, R., M., (1991). Legislative perspectives on statewide testing: Goals, hopes, and desires. *Phi Delta Kappan, 73*(3), 239-242.

Koretz, D. M., (1986). *Trends in Educational Achievement.* Washington, D.C.: Congressional Budget Office.

Koretz, D. M., (1988). Arriving in Lake Wobegon. Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46-52.

Koretz, D. M., (1991). State comparisons using NAEP: Large costs, disappointing benefits. *Educational Researcher*, 20(3), 19-21.

Koretz, D. M., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The effects of high stakes testing on achievement: Preliminary findings about generalization across tests. In (Ed.), Annual meeting of the American Educational Research Association (pp. Chicago, IL:

Linn, R. L., Graue, B., and Sanders, N. M. (1989). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average".* American Educational Research Association.

Linn, R., Kiplinger, V., L., Chapman, C., W., and LeMahieu, P. G. (1991). *Cross-state comparability of judgments of student writing: Results from the New Standards Project* New Standards Project, University of Pitsburgh.

Madaus, G. F. (1988). *The influence of testing on the curriculum.* In L. Tanner (Ed.), *Critical Issues in Curriculum.* (pp. 83-121). Chicago: University of Chicago Press.

Madaus, G. M. (1991). The effects of important tests on students: Implications for a national examination or system of examinations. *Phi Delta Kappan*, 73(3), 226-231.

Madaus, G. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issue and Practice, 11*(1), 26-31.

Madaus, G. F. (in press). A National Testing System: Manna From Above? A Historical/Technological Perspective. *Educational Assessment, 1*(1),

Madaus, G., F., and Kellaghan, T. (1991a). National testing: Lessons for America from Europe. *Educational Leadership, 49*(3), 87-93.

Madaus, G., F., and Kellaghan, T. (1991b). *Student examination systems in the European Community: Lessons for the United States* Contractor Report submitted to the Office of Technology Assessment, United States Congress.

Madaus, G. F., and Kellaghan, T. (1992). *Curriculum evaluation and assessment.* In P. W. Jackson (Ed.), *Handbook of research on curriculum.* (pp. 119-154). New York: Macmillian.

Madaus, G. F., and Kellaghan, T. (in press). British experience with "authentic" testing. *Phi Delta Kappan*, ,

Madaus, G., F., and Tan, A. (1993). *Growth of assessment.* In G. Cawelti (Ed.), *The 1993 ASCD Yearbook.* Alexandria, VA: Association of Superivision and Curriculum Development.

National Council on Education Standards and Testing. (1992). *Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People.* Washington, D.C.: Author.

Nuttall, D. L. (1992). Performance assessment: The message from England. *Educational Leadership*, 49(8), 54-57.

Resnick, L. B. (1992). Statement before the Subcommittee on Elementary, Secondary, and Vocational Education, Committee on Education and Labor, U.S. House of Representatives, February 4, 1992.

Shavelson, R. J., Baxter, G., P., and Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4(4), 347-362.

Shavelson, R. J., Baxter, G., P., and Pine, J. (1992). Performance assessments: Political Rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*(3), 232-238.

Stake, R., E. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan, 73*(3), 243-247.

U.S. House of Representatives, Committee on Education and Labor (1991). *National Council on Education Standards and Testing,* House Report 102-104, June 10, 1991, pp. 5-8

ENDNOTES

[1]Letter to the members of NCEST from Senator Hatch and Representatives Goodling and Kildee, September 23, 1991.