DOCUMENT RESUME

ED 362 551 TM 020 585

TITLE	Hearing on H.R. 6: Assessment. Hearing before the Subcommittee on Elementary, Secondary, and Vocational Education of the Committee on Education and Labor. House of Representatives, One Hundred Third Congress, First Session.
INSTITUTION	Congress of the U.S., Washington, D.C. House Committee on Education and Labor.
REPORT NO	ISBN-0-16-040958-6
PUB DATE	18 Feb 93
NOTE	62p.; Serial No. 103-2.
AVAILABLE FROM	U.S. Government Printing Office, Superintendent of Documents, Congressional Sales Office, Washington, DC 20402.
PUB TYPE	Legal/Legislative/Regulatory Materials (090)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Accountability; *Compensatory Education; *Educational Assessment; Elementary Secondary Education; Evaluation Methods; Federal Programs; Hearings; Learning; Models; *National Competency Tests; National Surveys; Research Methodology; State Programs; Testing Problems; Testing Programs; *Test Use
IDENTIFIERS	Congress 103rd; *Education Consolidation Improvement Act Chapter 1; Performance Based Evaluation; Proposed Legislation

ABSTRACT

At one of a series of hearings on H.R. 6, the Elementary and Secondary Education Amendments of 1993, several experts discussed issues related to national assessment, assessment in Chapter 1, and state efforts to develop new forms of assessment. Eleanor Chelimsky, Assistant Comptroller General for Program Evaluation and Methodology of the General Accounting Office reports that respondents to a national survey of educators do not favor a national test for accountability purposes, and that many have reservations about the use of national examinations. Richard Mills, Commissioner of Education for Vermont, also discussed a national system of examinations and reviewed some assessment developments in the Vermont state program. Thomas A. Romberg, Director of the National Center for Research in Mathematical Sciences Education, indicates that a new assessment paradigm centered on children's learning is needed for Chapter 1 assessment and for all students. Sylvia T. Johnson, professor of statistics and research methodology at Howard University (Washington, D.C.), spoke about testing issues in relation to the proposed national assessment. Prepared statements of each of these witnesses and Lynn C. Woolsey, a Representative in Congress from California, follow the remarks. (SLD)

HEARING ON H.R. 6: ASSESSMENT

HEARING

BEFORE THE

SUBCOMMITTEE ON ELEMENTARY, SECONDARY, AND VOCATIONAL EDUCATION

OF THE

COMMITTEE ON EDUCATION AND LABOR HOUSE OF REPRESENTATIVES

ONE HUNDRED THIRD CONGRESS

FIRST SESSION

HEARING HELD IN WASHINGTON, DC, FEBRUARY 18, 1993

Serial No. 103-2

Printed for the use of the Committee on Education and Labor

U.S. GOVERNMENT PRINTING OFFICE
67-897 WASHINGTON : 1993

For sale by the U.S. Government Printing Office
Superintendent of Documents, Congressional Sales Office, Washington, DC 20402
ISBN 0-16-040958-6

67-897 O - 93 - 1

COMMITTEE ON EDUCATION AND LABOR

WILLIAM D. FORD, Michigan, *Chairman*

WILLIAM (BILL) CLAY, Missouri
GEORGE MILLER, California
AUSTIN J. MURPHY, Pennsylvania
DALE E. KILDEE, Michigan
PAT WILLIAMS, Montana
MATTHEW G. MARTINEZ, California
MAJOR R. OWENS, New York
THOMAS C. SAWYER, Ohio
DONALD M. PAYNE, New Jersey
JOLENE UNSOELD, Washington
PATSY T. MINK, Hawaii
ROBERT E. ANDREWS, New Jersey
JACK REED, Rhode Island
TIM ROEMER, Indiana
ELIOT L. ENGEL, New York
XAVIER BECERRA, California
ROBERT C. SCOTT, Virginia
GENE GREEN, Texas
LYNN C. WOOLSEY, California
CARLOS A. ROMERO-BARCELÓ, Puerto Rico
RON KLINK, Pennsylvania
KARAN ENGLISH, Arizona
TED STRICKLAND, Ohio
RON DE LUGO, Virgin Islands
ENI F. H. FALEOMAVAEGA, American Samoa
SCOTTY BAESLER, Kentucky

WILLIAM F. GOODLING, Pennsylvania
THOMAS E. PETRI, Wisconsin
MARGE ROUKEMA, New Jersey
STEVE GUNDERSON, Wisconsin
RICHARD K. ARMEY, Texas
HARRIS W. FAWELL, Illinois
PAUL B. HENRY, Michigan
CASS BALLENGER, North Carolina
SUSAN MOLINARI, New York
BILL BARRETT, Nebraska
JOHN A. BOEHNER, Ohio
RANDY "DUKE" CUNNINGHAM, California
PETER HOECKSTRA, Michigan
HOWARD "BUCK" McKEON, California
DAN MILLER, Florida

PATRICIA F. RISSLER, *Staff Director*
JAY EAGEN, *Minority Staff Director*

SUBCOMMITTEE ON ELEMENTARY, SECONDARY, AND VOCATIONAL EDUCATION

DALE E. KILDEE, Michigan, *Chairman*

GEORGE MILLER, California
THOMAS C. SAWYER, Ohio
MAJOR R. OWENS, New York
JOLENE UNSOELD, Washington
JACK REED, Rhode Island
TIM ROEMER, Indiana
PATSY T. MINK, Hawaii
ELIOT L. ENGEL, New York
XAVIER BECERRA, California
GENE GREEN, Texas
LYNN C. WOOLSEY, California
KARAN ENGLISH, Arizona
TED STRICKLAND, Ohio
DONALD M. PAYNE, New Jersey
CARLOS A. ROMERO-BARCELÓ, Puerto Rico

WILLIAM F. GOODLING, Pennsylvania
STEVE GUNDERSON, Wisconsin
HOWARD "BUCK" McKEON, California
THOMAS E. PETRI, Wisconsin
SUSAN MOLINARI, New York
RANDY "DUKE" CUNNINGHAM, California
DAN MILLER, Florida
MARGE ROUKEMA, New Jersey
JOHN A. BOEHNER, Ohio

CONTENTS

HEARING ON H.R. 6: ASSESSMENT

THURSDAY, FEBRUARY 18, 1993

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON ELEMENTARY, SECONDARY,
AND VOCATIONAL EDUCATION,
COMMITTEE ON EDUCATION AND LABOR,
Washington, DC.

The subcommittee met, pursuant to notice, at 10:45 a.m., Room 2175, Rayburn House Office Building, Hon. Dale E. Kildee, Chairman, presiding.

Members present: Representatives Kildee, Miller, Sawyer, Becerra, Green, Woolsey, English, Strickland, Romero-Barcelo, Goodling, Gunderson, Molinari, Cunningham, Roukema, Boehner.

Staff present: Susan Wilhelm, staff director; Lynn Selmser, professional staff member; Jeff McFarland, subcommittee counsel; Jane Baird, education counsel; Jack Jennings, educational counsel; Diane Stark, legislative specialist; Margaret Kajeckas, legislative associate; Tom Kelley, legislative associate; June Harris, legislative specialist.

Chairman KILDEE. The Subcommittee on Elementary, Secondary, and Vocational Education convenes this morning for the third of its hearings on H.R. 6, the Elementary and Secondary Education Amendments of 1993. This morning's topic is assessment. As many of you know, during the last Congress this committee struggled with the issue of a voluntary system of national assessments and what the appropriate Federal role might be in that.

At the same time while we were wrestling with that, many States, such as California and Vermont, have been moving forward in developing new forms of assessment. The reauthorization of the Elementary and Secondary Education Act provides an opportunity to further discuss the issue of national assessment, as well as to consider options for improving assessment in Chapter 1.

As members of the subcommittee are very well aware, Chapter 1 contains requirements which have resulted in extensive testing throughout our schools. Questions have been raised about the usefulness of these assessments for improving instruction. One of my concerns has been how can we use assessment to really improve education and the delivery of education, rather than just getting statistics and figures.

Our witnesses this morning will discuss issues related to national assessment, assessment in Chapter 1, and State efforts to develop new forms of assessment. Our witnesses are Ms. Eleanor Chelimsky, assistant comptroller general for program evaluation and

methodology, U.S. General Accounting Office; Dr. Richard Mills, commissioner of education, the State of Vermont.

I saw your good senator last night as he walked in—both of your good senators, I should say, as they walked in. They are both good friends of mine and a bipartisan representation of a very great State.

Dr. Thomas Romberg, University of Wisconsin. Dr. Romberg is appearing this morning in his capacity as a member of the Department of Education's Advisory Committee on Chapter 1 Assessment; and also Dr. Sylvia Johnson, professor and coordinator of research methodology and statistics of Howard University.

Before we begin the testimony this morning, I would like to recognize my very good friend, Mr. Bill Goodling, the ranking Republican member of both this subcommittee and of the full Committee on Education and Labor, and an unquestioned, devoted, zealous friend of education.

Mr. Goodling?

Mr. GOODLING. After hearing that, what can I say? Only that I am here to try to do whatever we can do to make sure that Head Start and Chapter 1 are far better programs than they presently are. I am the one who keeps saying quit saying it's ice cream, apple pie, and all these wonderful things. But it has to be a darn sight better than it presently is. Hopefully, we can have as our guiding motto: "Excellence rather than access." Then I think we can give to the youngsters what we had hoped to originally.

Thank you, Mr. Chairman.

Chairman KILDEE. Thank you. If there are no other opening statements, we will begin with our first witness, Ms. Eleanor Chelimsky.

STATEMENTS OF ELEANOR CHELIMSKY, ASSISTANT COMPTROLLER GENERAL FOR PROGRAM EVALUATION AND METHODOLOGY, U.S. GENERAL ACCOUNTING OFFICE, WASHINGTON, DC; RICHARD MILLS, COMMISSIONER OF EDUCATION, STATE OF VERMONT, MONTPELIER, VERMONT; THOMAS A. ROMBERG, DIRECTOR, NATIONAL CENTER FOR RESEARCH IN MATHEMATICAL SCIENCES EDUCATION, UNIVERSITY OF WISCONSIN, MADISON, WISCONSIN; AND SYLVIA T. JOHNSON, PROFESSOR AND COORDINATOR, RESEARCH METHODOLOGY AND STATISTICS, HOWARD UNIVERSITY, WASHINGTON, DC

Ms. CHELIMSKY. Thank you very much, Mr. Chairman, members of the subcommittee. It's a great pleasure to be here to talk to you about the work we've been doing on student achievement standards and testing.

I did want to present the people who are here with me who have worked on the study, if that's possible. Fritz Mulhauser, who is our study director; Gail McCall, back there; and Kathleen White.

As you know, our evaluation has three parts and asks three questions. First, where are we today in student testing with respect to time and cost, and what would be the likely price tag for a national examination? Our first report that was issued in January—which you have, Mr. Chairman—responds to this two-part question.

Second, what has been the experience outside the U.S. in linking testing to standards, and what can we learn from that experience? We're issuing a report shortly which responds to that question.

Finally, how are we doing with regard to our own efforts in the United States to develop national standards and use existing tests to measure adherence to those standards. This is the NAGNI effort to use the NAEP test. Now, we issued a letter on this in March 1992, and our final report will be published in April.

In the interest of time, let me just give you some short answers to the first two questions with the understanding that you will find supporting detail in my full statement in the published report.

Chairman KILDEE. Very good.

Ms. CHELIMSKY. First, the current status of testing: Well, it seems our students can hardly be called overtested. We looked at systemwide testing; that is, not specialized tests, but tests that are given generally to children in a school district. We found that the average student spent only 7 hours annually on such testing, that includes preparation, the actual test-taking, and all related activities. If you look at test-taking alone, that brings the time down to 3.4 hours annually.

Now, the cost of this testing came to $15 per student, on average, including the cost of the test and staff time. In total, we estimate that in the Nation as a whole systemwide testing in 1990-1991 cost us about $516 million. What then would a national test cost? Well, the estimated cost for two types of tests, multiple-choice tests and performance tests, the multiple-choice tests we projected at about $160 million, and the performance test we've projected at $330 million, plus about $100 million more for a one-time test development.

Now, these costs would be additional to the $516 million currently being spent only if schools did not eliminate tests they are presently giving. Both our estimates assume three grades, totalling 10 million students tested annually.

Finally, with regard to current status, we asked State and local testing officials and educators how they felt about a number of issues now being raised. Four points seem especially important to report here. First, our respondents considered that multiple-choice tests, which constitute 71 percent of current testing, are notably inferior to performance tests. We have only seven States currently with experience on these tests.

Our respondents did not favor using tests for accountability purposes; that is, high-stakes decision-making. You know, certifying whether individual students meet standards or not, deciding on whether they will go on to other education, and so forth. They saw this to be likely inaccurate and potentially counterproductive to real improvement in the classroom.

Now, others, as you know, believe that accountability is perhaps the most important purpose of testing, but that was not what our respondents told us. Instead, they called for tests of high technical quality that are used for two other purposes: monitoring, that is, measuring progress relative to standards over time with low stakes or no stakes, and diagnosing teaching and learning needs in the classroom.

Finally, about 40 percent, the significant percentage of our respondents, were against a national examination, fearing data of poor quality and results that could be misused.

As I mentioned earlier, our second question dealt with experience outside the U.S. in linking testing to standards. We looked at Canada because of its U.S.-like decentralization and because of other similarities as well. We have a rich set of findings to report, and let me just give you five of them here.

First, the question of testing purpose that emerged as so important among our U.S. respondents seems to have been absolutely critical in Canada. Indeed, it channeled the current Canadian system into two kinds of tests. They have examinations that feature high-stakes decision-making and accountability, but only for some courses, and they have assessments that monitor progress but without stakes, any stakes, for students or teachers.

Second, we found that there has been enormous funded involvement of educators at every phase of standard and test development, execution, and revision. Third, notable efforts have been made to address questions of fairness and potential misuse of test results. Fourth, we found that many difficulties are now being encountered in moving from a decentralized province-based testing system to a centralized national one.

Finally, although it's important to understand the evolution of the Canadian testing system, it's also important to recognize that beyond a few one-shot surveys, anecdotal information, and polls, no formal evaluation has yet assessed this system. So, we don't know whether or not it has been effective in increasing student achievement.

Our conclusions based on these two studies are that we need to pay attention to a number of things as we move forward to develop national standards and link tests of them. Standards need to be developed and tests created to support them in an objective, careful, and highly skilled process that ensures the technical quality which brings measurement accuracy. Pressures for speed and excessively tight time lines are just not conducive to valid and reliable measurement.

A second point is that our respondents' lack of interest in using test for accountability and their objections to national testing are important. They mean that educators need to be brought along in any test effort and intimately involved along with technical experts and others in the whole standard and test development process.

Third point: if we're going to have high-stakes tests, issues of fairness and accuracy will need to be confronted and addressed and safeguards should be developed against misuse of results.

Finally, the ability to develop well-specified tests of wide application whose results can properly be used for policymaking is closely tied to the purpose of those tests. Local purposes differ from national ones, just as monitoring and accountability differ and the tests that implement them must reflect those differences.

The Canadian experience shows that all of these purposes are feasible to implement; however, it would be nearly impossible to accommodate them all, I think, with the same test. Whatever purpose or purposes we eventually decide on in a national testing effort, the results must inform what is happening at the local

school. Nothing will really improve in education unless what is actually taught and learned in the classroom increases in both quantity and quality across America.

That's why it's so important not only to ensure that a feedback loop exists between national or regional testing and local teaching and learning, but also to provide for the full involvement of America's educators in this effort. Programs of the 1960s taught us that not much will happen without their support.

That concludes my remarks, Mr. Chairman. Thank you very much.

[The prepared statement of Eleanor Chelimsky follows:]

United States General Accounting Office

GAO

Testimony

Before the Subcommittee on Elementary,
Secondary, and Vocational Education
Committee on Education and Labor
House of Representatives

For Release on Delivery
Expected at 10:00 a.m.
Thursday
February 18, 1993

Student Achievement Standards and Testing

Statement of Eleanor Chelimsky
Assistant Comptroller General
Program Evaluation and Methodology Division

GAO/PEMD-T-93-1

Mr. Chairman and Members of the Committee:

I am pleased to be here today to discuss GAO's work in the broad area of student achievement standards and testing. At your request, we have done three studies: one on the extent and cost of testing in this country, another on the experience with standards and tests in Canada, and a third on the initial efforts to set standards for judging student performance on the National Assessment of Educational Progress (NAEP). A report on the first is available and a report on the second will be published soon.[1] We issued an interim paper on the NAEP work last March.[2] We expect the final report to be issued in 90 days.

I will focus today on the main themes of our findings and conclusions from the first two studies. Our reports describe the scope of the work and methods we used in detail. In brief, we gathered data on the present extent and costs of testing in the United States (and the views of education officials on testing issues) by surveying all the states and a national sample of school districts. We also estimated likely costs for a national test. With regard to the Canadian experience with standards and tests, our effort involved reviewing provincial evaluations and other data, visiting provincial and district offices in several provinces, and interviewing officials in the provinces that we could not visit. The Canadian experience is relevant to the current U.S. effort to establish standards and related tests for school learning because some provinces have for some time had testing systems similar in various ways to plans suggested for the United States and because standards play a large role in those systems.

I will turn first to the information we produced on current testing and our forecasts of resources required for a national test and then discuss the Canadian experience.

TESTING TODAY

In 1990-91, students in the United States did not seem to have been overtested--the average student spent only 7 hours annually on systemwide testing (including preparation, test-taking, and all related activities)--and the cost totaled, on the

---

[1]U.S. General Accounting Office, Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination, GAO/PEMD-93-8 (Washington, D.C.: January 1993), and Educational Testing: The Canadian Experience With Standards, Examinations, and Assessments, GAO/PEMD-93-11 (Washington, D.C.: February 1993).

[2]U.S. General Accounting Office, National Assessment Technical Quality, GAO/PEMD-92-22R (Washington, D.C.: March 1992).

average, $15 per student, including the cost of the test and staff time.[1] The bulk of this testing was traditional in format (71 percent of tests consisted of multiple-choice questions only). Newer test types, such as performance tests in which students write out some answers, were much less common: tests with more than just a writing sample element were in use in only seven states. The performance tests also cost more. In the states where we had the best comparative data we found that multiple choice tests averaged less than half the cost of performance tests--$16 versus $33 per student, respectively. We estimated that in the nation as a whole, systemwide testing in 1990-91 cost about $516 million.

ESTIMATES OF NATIONAL TESTING OPTIONS' COSTS

We used our data on current costs for different kinds of tests to estimate what it would cost for a national-level test (assuming three grades tested in a year, totaling 10 million students). Since multiple-choice tests currently average about $16 per student, a national multiple-choice test would cost about $160 million. Because performance tests cost more (an average of $33 per student), national implementation of such a test, again at three grade levels, would cost a total of $330 million. Also, our data showed that these tests are expensive to develop: we estimate a national system would cost as much as another $100 million in one-time development costs.

The new costs of a national testing plan would vary, however, depending on whether schools added the test or used it to replace others currently in use. The multiple-choice option would add the least new cost in money and time, since (from data we gathered on past decisions) we predict three quarters of the districts would drop an existing test and replace it with the national test. Because many fewer districts use performance tests now, a national performance test would add more new costs in money and time: $209 million and another half hour per student per year. Regional state clusters of performance tests, the option recommended by the congressionally mandated National Council on Education Standards and Testing (NCEST), would add slightly less: $193 million of costs and 25 minutes more for the

[1]We define systemwide tests as those given to all, almost all, or a representative sample of students at any one grade level in a school district. This definition covers most standardized tests, except those given to certain groups. We did not include tests given to students under the requirements of the federal Chapter 1 program (unless districts gave such tests systemwide, which is common, according to Department of Education officials).

average student in testing time.[4]

TESTING OFFICIALS' VIEWS

Cost is not the only issue in comparing the options, of course. Multiple choice tests are familiar and provide strong comparative data but--according to opinion data from our survey--are least valued by state and local testing officials. State clusters of different performance tests are the least-developed method, cost twice as much as multiple choice tests, and would not necessarily be comparable among themselves or over time. They may, however, be better linked to local teaching and--again according to our survey--are viewed by testing officials as better measurements of what students know and can do.

Testing officials saw continued benefit to testing in general, even if there were to be more tests, but in discussing trends in the field, they expressed concern over the purpose, quality, and locus of control over further tests. With regard to purpose, our respondents voiced their preferences for more performance-based assessment that can help diagnose learning and teaching needs at the lowest levels, and they also recognized as valid the purpose of producing national data that are comparable over time. However, a third purpose of testing--accountability--was downplayed by our respondents, and concerns about possible misuses of tests in this regard (to compare unlike schools and districts, for example, or to reach unwarranted conclusions about students), were cited quite often as well. Quality and locus of control issues were expressed in respondents' preferences for tests that are of high technical quality, measure diverse skills in diverse ways, and cover what their teachers teach.

On the question of a national test or system of tests, our survey revealed significant opposition to the concept. Forty percent of local respondents and 29 percent of state respondents saw no advantages to a national system, and they forecast some disadvantages, particularly the potential for misuse of results. (Thirty-two percent of local respondents and 53 percent of state respondents did, however, specifically cite the potential for comparing test scores nationally as an advantage of a national testing system, although this purpose is to some degree in conflict with the local utility they also wanted.)

CONCLUSIONS

The costs of a national examination system may be less than anticipated. Assuming a hybrid system of testing for a number of

---

[4]The Council's recommendations are in its final report, Raising Standards for American Education (Washington, D.C.: January 1992).

potential purposes--testing all students in three grades--our estimates of the cost are higher than those of some national test proponents but lower than those of some opponents. Our projected figure of $330 million annually (for the most likely type of performance test, similar to tests in use in some states now) is about one tenth the amount some have suggested. The new costs would be less than that (about $200 million) and the added student testing time (increasing by up to 30 minutes the average amount of systemwide testing time per student, to a national average of about 7.5 hours total time and 4 hours of actual test-writing per student per year) does not seem unduly burdensome.

More specific forecasts or predictions will require making some decisions about the purpose or purposes that national tests can be expected to serve. Our data exemplify this need to choose in two ways. First, tension exists between our correspondents' preferences for two distinctly different emphases in testing: tests developed under local control and tests used principally for monitoring progress over time. Local control suggests a wide diversity of tests matched, in order to be most useful, to local variations in what is taught and learned; however, the goal of monitoring across classrooms, schools, districts, or states sets limits to the variation in tests that can be allowed without losing comparability. Second, tension exists between both local control and monitoring, on the one hand, and accountability, on the other. Although our respondents were not greatly concerned with accountability, others--chiefly outside the schools--have suggested that this purpose may be the most important: that is, using test results for high-stakes decisionmaking about students, teachers, or schools, and thereby emphasizing the importance of teaching and learning the material to be tested. Since it is not clear that one test can serve all three purposes, we conclude that decisions about test purposes are a high priority.

A final point is that the opposition we found to national testing, although abstract in the sense of not being linked to a particular proposal, should be carefully considered and addressed. The cooperation of state and local administrators and educators is important for any national testing effort. It seems reasonable to believe that if their knowledge, skills, and involvement can be effectively harnessed to the national testing effort, the success of the enterprise is more likely to be achievable.

EDUCATION STANDARDS AND TESTING IN CANADA

Turning to our second study for the committee, Mr. Chairman, let me present some observations drawn from the experience of Canadian provinces with education standards and testing, discussed in detail in our full report. As an affluent "high-tech" industrial society, Canada resembles the United States in many ways, and it also has considerable experience with a

decentralized student testing system presenting features recommended by NCEST for future adoption in the United States. Such features include measuring progress in relation to standards, using performance tests and other methods, and involving teachers intimately in all phases of testing. The United States does not lack experience with testing, of course, but what has happened in Canada affords useful contrasts on some key dimensions, as well as interesting information with regard to the development of incentive systems to counter various problems and pitfalls.

In brief, the major instructive contrasts and important elements we found are as follows.

Province-Level Standards

In Canada, educational standards are currently set at the province level, with major involvement of educators, especially teachers. (A recent effort there to set some national standards in basic learning areas, as a prelude to a national test of minimum competencies, has also included extensive involvement of teachers.) This differs from current efforts in the United States to set national standards chiefly by groups of experts, with only modest teacher involvement.

Different Tests for Different Purposes

In most Canadian provinces, two entirely different testing systems are dedicated to the separate purposes of certifying whether individual students meet standards (accountability) and tracking whether learning in general across a province is in line with what is expected (monitoring). (We refer to these in our report as examination and assessment systems, respectively; five provinces have the former and eight the latter.) This contrasts with the views of some in the United States who have proposed a single test or assessment method to serve many purposes.

Tests Linked to Standards

Both examinations (for accountability) and assessments (for monitoring) are developed within a province based on the standards for what should be learned in a particular course (in the case of the examinations) or in a particular subject and grade (in the case of the assessments). Both kinds of tests are revised often with major teacher involvement to reflect constant changes in those standards. This contrasts with the large U.S. use of commercially developed tests (customized in some cases to reflect state requirements) to measure students' cumulative knowledge of broad subject areas. In addition, both types of provincial tests use multiple methods, including the common use of essays but other tasks as well. The predominant format of U.S. tests, in contrast, is multiple-choice questions.

Stakes Differ for Different Tests

The idea often heard in recent U.S. testing debates that it is necessary to attach high stakes to all or many tests to emphasize the importance of learning to teachers and students does not seem to be reflected in the Canadian experience. For example, examination scores do not stand alone but are blended with teacher evaluations to form students' final grades, and the weight given to the exam score has been declining. (Assessments have no stakes for students or teachers.) Canada seems to rely instead on the continuous funded involvement of teachers in all phases of standard-setting and design of both examinations and assessments, as well as in test administration and scoring, to emphasize the importance of provincial standards.

Safeguards Associated With Tests

Canadian officials have employed a variety of safeguards to prevent misuse of test results. Safeguards in the examination system (where the accountability purpose means that results will have some consequences for students and teachers) include distributing the test specifications widely in advance, ensuring multiple opportunities for success, allowing for rescoring, and accommodating students with disabilities. In the assessment system--that is, tests designed to monitor or give an accurate picture of how all students are doing--other safeguards such as requirements that all students be tested and that reporting be both delayed and aggregated help ensure, on the one hand, data undistorted by biased participation and, on the other, fewer possibilities that results can be misused in decisions about individual students or teachers. Again, where data on all students are not needed in the assessments, sampling is increasingly used to permit multiple methods of testing (such as more expensive performance methods) without increases in cost.

Resources for Learning

Provincial funding formulas have been used in Canada to level resources among schools in a province and thus enable teachers generally to have comparable resources to implement the curriculum requirements. This is in contrast to sometimes large resource disparities among districts in the United States which give rise to the complaint that testing is inherently unfair since students may have experienced major differences in opportunity to learn. Thus, the issue discussed in the United States concerning "delivery standards," which some believe should accompany learning or achievement standards, is mitigated in Canada, because a degree of equalization of resources has been achieved.

Inadequate Evidence of Results

It is important to note that the effects of Canada's efforts to set standards and link tests to them have not been established. It is not known whether the elaborate strategy Canadian provinces have put in place has in fact caused better student achievement. No independent yardstick--no set of data, no national evaluation--affords such a measurement. There is some information on other effects of the effort, but it is scattered and of varying quality. For example, there are assertions by teachers that there has been some narrowing in what they teach and how, and there are survey data showing that high stakes on examinations elicit both anxiety and increased motivation in students. Increased fragmentation or stratification of student groups in some provinces has also been suggested to be the result of the isolation from others of those taking courses for which there are exams. Also, a rise in the number of students taking an extra year of high school is attributed, in part, to some staying longer in order to do better on the last set of examinations. In the view of some, and to the degree that they are accurate, all these results--from greater teacher focus on the content to be examined to heightened emphasis by students on academics--may be useful correctives to past problems of too much diversity in what is taught and too little student time and attention; others see them more negatively.

Positive Response to Testing from Teachers and Others

We found that Canadian teachers respond to the incentives offered them: many of them seek out the opportunities to be involved in provincial activities of setting curriculum standards and designing or grading tests of all kinds; they see these as valuable professional development efforts. Provincial authorities see them as building commitment to the results. Surveys and public opinion polls show that teachers and the Canadian public manifest general approval of the examination systems and believe that education has benefited.

Uncertainties Concerning Canadian National Test

Finally, we were interested to discover that the Canadian provinces have initiated a project to develop a national test. Because of the extensive province-level systems of standards and tests I have just described, the national project has encountered many objections. Agreement on the standards to be used has been elusive, and one province has decided that its disagreements are so fundamental that it will not take part at all. Extensive work has been done to define what is expected and how it should be measured. There is consensus that the purpose of the effort is monitoring and that there will thus be no stakes for students. Reporting is planned to extend no lower than the province level:

67-897 O - 93 - 2

Canadian officials believe school or district-level monitoring by this national test would raise the stakes too high and compromise participation, as well as being much more costly owing to the larger samples needed. The present plan is for several provinces to work together to develop, on behalf of all the provinces involved, a new test to measure the standards emerging from the multiprovince conversations. However, many key matters remain unsettled, including disagreements within professional groups about the emphasis to be given different topics within a subject and the testing methods to be used, the level of difficulty of the test, and disagreements between educators and employers about the balance of academic and real-world skills to be tested.

Summary of Observations on Standards and Testing in Canada

In short, in Canada we found a coordinated set of standards, course specifications, and tests that are well-regarded by both educators and the public. Monitoring and accountability purposes are separated, and teachers are extensively involved in the activities of deciding what should be learned and of measuring the results. However, we could find no strong information on the effectiveness of Canada's system; it is implemented essentially at the provincial level; and efforts among the provinces to gain consensus on a plan for common standards and a national test have proven to be of great difficulty and uncertain feasibility.

CONCLUDING OBSERVATIONS

Let me conclude, Mr. Chairman, with some general observations that link the details of our work to broader questions of testing policy facing the Congress.

National Testing Design Must Flow From Purpose

First, is national testing feasible? Although our data show some skepticism on the part of officials and educators, they were reacting only to a general concept of expanded national testing. The Canadian experience suggests that the key determinant of feasibility may be deciding on the purpose to be achieved by testing. This is because most issues of technical quality (for example, validity and reliability) and cost must be addressed in a specific context of purpose. For example, if the purpose is monitoring, samples can be used that afford great flexibility in the type of test and large cost savings, even if expensive testing methods are used. If the purpose is accountability, such as certifying students, tests must have safeguards and other properties that will be expensive, including security; also, equitable exposure to the tested material is of critical importance to a fair use of the results. Maximizing one purpose may degrade another: the research shows that the higher the stakes of a test, the more effort individuals will put into assuring high scores quite apart from genuine learning, which in

turn makes the data less valid for monitoring. Our sense is that the debate over national tests has not yet distinguished clearly among the purposes to be served, nor has it drawn the appropriate conclusions concerning the technical difficulties involved in reconciling the conflicting requirements of a multipurpose test. We found the Canadian observations helpful in showing the feasibility of separate testing systems clearly specified to serve different purposes.

Finally, and again with respect to feasibility, our estimates suggest that students are not currently overtested and that the likely resource expenditures for various national test options are not exorbitant. However, these expenditures could vary considerably, depending on the purposes that are chosen for the test. At present, we have an open field of options before us, with none foreclosed. Yet it is not clear that we can achieve all purposes with one test.

The Desire for Rapid Development Must Not Constrain the Technical Quality of Measures

Second, will measurement be accurate? Both policy decisions, in general, and decisions affecting individual students and teachers should rest on sound data. Here, the key question is how we intend to test. For now, our hopes outstrip our capacity. As our respondents showed us, there is a yearning for better ways to test, so that we do full justice to students' learning, yet there is uncertainty over the state of the art in testing once we go beyond the familiar methods and a recognition of danger in the overeager use of unproven measures. We do not know whether the intense pressure first seen in 1990 and 1991 for the immediate implementation of a national test has abated, but we do know that high-quality innovative measurements, especially if adapted to many different regions (NCEST's clusters of states), will not be done quickly. Funding and governance arrangements need, therefore, to include careful monitoring of the technical aspects of the work, so that eagerness for rapid results does not supplant quality as the prime goal.

Standards Raise Many Tough Issues Worth Considerable Effort in Design and Implementation

Third, and last, where should we begin? Just the initial step of setting standards for student learning is quite difficult and it raises procedural, conceptual, and technical issues such as what roles should be played by educators and others in setting standards, what is needed by all or only some students, and how can such efforts guard against setting standards that are technically unmeasurable? Groups are at work in the United States on precisely this, some with federal support, for many different subjects, but the work has just begun and not all the

issues are on the table yet. And we must acknowledge that, in general, our schools do not now hew to high standards for rigorous academic work for all students. That is, the set of new content standards will pose considerable challenges for teachers and students, quite apart from the measurement problems we have just discussed. How will time be found to teach all the new material likely to be urged by each subject-matter group? What about schools that lack the instructional resources needed or teachers who lack the knowledge to be covered? And as implementation begins to affect measurement, what happens to test comparability if states and districts cannot handle all the material required by new standards and make different choices among the new requirements?

Given so much complexity, it may be wise to begin by emphasizing work on standards for the next several years, including some of the thornier issues of how they will come alive within the schools, while allowing the many promising state experiments with testing methods and formats to yield their results. In this way, we can learn much more about what works before we take too many major decisions about what and how to test at national levels. That would allow the debate over purpose to catch up as well, which is critical because practical choices will be difficult without a resolution of that debate and because the final answers to the questions of feasibility, cost and measurement accuracy also depend on purpose.

MATTERS FOR CONSIDERATION

We would emphasize two matters. First, because of the sizable knowledge base in current testing programs, because of the voices of opposition and uncertainty we heard from our survey respondents, and because of the successful Canadian experience with regard to teacher involvement, we believe it would be important for the Congress to consider specific ways to encourage the participation of teachers as well as state and local education administrators in further steps of developing standards and all aspects of increased testing (including development, administration, and scoring).

Second, we believe that the Congress should carefully consider how to ensure the technical quality of any tests in a national examination system. This is not only because technical quality was a frequently reported concern of our respondents but also because of the combined popularity and newness of large-scale performance testing. Popularity often results in time pressures and compressed schedules, whereas newness requires the development of valid, reliable tests and efficient and reliable scoring methods, all of which need trial, effort, and time. Seven states have performance tests now, and nine others told us they are 3 years away from such tests; creating a national system will be an effort of unprecedented scope and novelty and yet

enthusiastic prompting for immediate action seems likely. Money and time can always be saved at the expense of quality: for example, by doing less pilot testing, creating fewer test forms, shortening the test, or relaxing test security. In view of the lasting effects of incorrect decisions based on flawed test data, we urge explicit and proactive consideration to quality assurance in any national examination system implementation plan.

Mr. Chairman, this concludes my statement. I will be happy to answer any questions.

(973736, 740, 741)

Chairman KILDEE. Thank you very much, Ms. Chelimsky.

Our next witness is Dr. Mills. Dr. Mills, in our bill last year, the Neighborhood School Improvement Act, we gave high profile to commissioners of education. We stuck with that, even though we got into trouble with the Governors Association. So we're happy to have you here this morning.

Mr. MILLS. Well, thank you very much. I'll try to stay out of trouble.

I very much like what I just heard. I think the substance of this research is that it is doable, a national examination system—not a Federal test, but a national examination system. I think that you have an amazing opportunity in front of you right now, in front of the Congress, to create the conditions or really to take the most decisive assessment policy decision that could be taken because it would remove the pressures that currently exist to focus almost entirely on a particular testing, multiple choice testing.

When Vermont was beginning its development of an assessment, the State board and I spent a lot of time traveling the State, talking with citizens—actually not making presentations, but asking questions—asking people what they wanted to know about their school, what they knew already, what they would do if they knew more. We spent a lot of time listening to teachers and students.

And I particularly remember the teachers who described weeks of preparation for multiple-choice standardized tests that they didn't believe in, another week of the administration of that test, and then the presentation of results to a public that didn't know what the arcane jargon meant and so could not have a sensible discussion about performance.

I think what is at issue here is having and provoking a sustained national, State, and local conversation about performance. We have to not only talk with people, we have to listen, and we have to show them what high performance looks like. I think that we need assessment results at several different levels; a teacher needs to know certain things.

There has to be a conversation between teacher and student and parent; there has to be a conversation among governors and commissioners and State boards and legislators; and there has to be a national conversation. It has to go far beyond the educational community. It's very important. It's a practical issue and it's a policy issue to make certain that all these pieces fit together in some reasonable way.

That means reaching an agreement on goals, reaching agreement on the principles that drive assessment, making clear that we really mean it when we say high skills for every single student—no exceptions, no excuses. We say that, I don't think many people yet mean it and understand the consequences of not meaning it.

I don't think that just any assessment system would do. We need an assessment system that not only measures results, but inspires continuous improvement. It's not about finding out who didn't get something, it's about sustaining continuous improvement. It's about an assessment system that's linked to high, shared common standards.

We need assessments that are founded on genuine opportunities to get good at what is being assessed: opportunities for teachers

and opportunities for all students. I found in the Vermont experience two other things. I'm convinced, from what I've seen and what I've heard in classrooms all across Vermont, that assessments have to include something more than just multiple-choice tests. They have to include portfolios and other actual encounters with real problems.

Secondly, I'm convinced that the kinds of assessments that are going to drive change are assessments that emerge out of very extensive collaboration. That doesn't mean developing something and showing it to a few teachers or even a few hundred teachers. It means giving the keys to people on the front line and saying, "What does high performance look like? I don't get it. Tell me again. That isn't right yet either. Tell me again." Keep pushing the design decisions back to the very best teachers we can find. That's what we did.

It's great to see Tom Romberg here. He was present at the beginning. We created some expert panels of Vermont teachers and then we enriched them with the best thinking that we could find from around the country and we supported them in their thinking and their work.

I think a national system is important for another reason. We have to be able to benchmark performance against some high common standards. The idea is not to provoke competition, although that's probably going to happen, it's to enable us all to put aside this commonplace that we all hear: "Our school is fine; there is a problem, however, down the road."

There's a tremendous amount of denial. If you listen to students, they will almost to a person say, "We are not challenged by the work we are given to do in schools." But we adults don't seem to confront that fact. By putting in place a coordinated system of assessments that rests on State components and local components and that has at its heart shared standards of performance, we can have a sensible discussion about what children really know and can do.

I also think that a national system is important because no one can afford to do it right acting alone. I cited in my printed testimony the New Standards Project, I think it's a revolutionary effort. It's very much along the lines of the State of Vermont and the State of Kentucky, the work that California is doing and many others, but it's a shared venture.

The partners together represent 46 percent of the enrollment in the Nation. We are together going to be spending about $30 million to develop a system of assessments in the subject areas covered by the national goals. No State can do a credible job on that acting alone. Kentucky needs Vermont, Vermont needs Kentucky. We are all working together.

Finally, lest someone doubt, I really believe, I think, there is a very powerful emerging consensus behind this belief that we do need new assessments. We are on the cusp of change—we are beyond the cusp of change. I cited just one study that came to hand in my remarks.

George Maddaus and a team from the National Science Foundation reviewed commonly used assessments, assessments commonly used in American classrooms. They found that tests commonly

taken by students, both standardized tests and textbook tests, emphasize low-level thinking and knowledge.

We are just midway in a process in Vermont to define a common core of learning. So far, we have listened to 2,000 people explain in plain language what they want their children to know and be able to do. They are not asking for basic skills; they are asking for world-class performance. And we have to have assessments that match that. I could go on about the Vermont experience. I've written to you about it. Let me stop there and yield to someone else and take questions later.

[The prepared statement of Richard P. Mills follows:]

Summary of Testimony before the Subcommittee on Elementary, Secondary and Vocational Education on February 18, 1993 by Richard P. Mills, Vermont Commissioner of Education

A word about perspective

I have served as Vermont's Commissioner of Education since March of 1988. During that time the State Board of Education and I have acted on a systemic reform agenda that includes a state-wide student assessment based on portfolios. I am also a member of the National Assessment Governing Board, and the board of the New Standards Project. I am here to speak on assessment to the Subcommittee from the perspective of a state commissioner of education.

We need a national system of examinations

The reauthorization proposal now before the Congress could be the decisive national policy decision on assessment, because it would remove pressures to rely solely on multiple choice tests and would provoke research, design and implementation of assessments more likely to inspire high student performance.

We need information on results at several levels -- classroom, school, state and nation. A thoughtful and complete structure will include local, state and national assessments. It is both a practical and a policy issue to see that the pieces do in fact fit together. That means that assessments at all levels should reflect similar goals, standards and principles.

Not just any assessment system would do. We need assessments that support continuous improvement in learning, that link to high standards -- standards that apply to all. We need assessments that are built on genuine opportunities for teachers and all students to learn. Vermont's experience has convinced me of two other things. Assessments should include portfolios and other actual performances rather than multiple choice alone. And assessments that drive change emerge from extensive collaboration with teachers.

A national system will enable us to benchmark the performance of our state or community against a high common standard. This is important to the State Board, to many local boards and to the Legislature. Everyone is aware of -- but hard pressed to counter -- the commonplace reaction: "my school is fine, the problem is elsewhere." The purpose behind this benchmarking is not to provoke competition but to help everyone confront the facts about performance so that all can engage in concerted action to improve.

A national system permits sharing of development costs. The New Standard Project is an illustration. New Standards envisions a combination of portfolio and performance assessments in all the subjects, with professional development and quality control elements built in that will cost over $30 million in the first three years. Vermont wants access to that system but could never build it alone.

And we do need new assessments. A recent study supported by the National Science Foundation concluded that "tests commonly taken by students -- both standardized tests and textbook tests -- emphasize low-level thinking and knowledge." * Those tests have driven curriculum and improvement efforts. It is common for improvement programs to be cast in terms of raising test scores rather than boosting skills. If we accept that research, those tests have contributed to a diminished educational opportunity for students.

A major barrier to change is the absence of sustained public demand for high performance. Regular reports on results could build that demand and give schools everywhere permission for dramatic change.

Very high skills for every student; no exceptions, no excuses

Vermont envisions very high skills for every student; no exceptions, no excuses. A similar vision appears to motivate the reauthorization effort. In Vermont, the portfolio assessment has been one lever to move us toward our vision. Here is our story in brief.

Vermont's assessment has three parts. The portfolio represents a students work product over a year. It reveals the whole scope of work, and has the potential to measure growth. The uniform test is a more traditional writing sample and a series of multiple choice and open ended mathematics problems. It offers an anchor as we develop the newer portfolio approach. The best piece is the student's selected example of peak performance. It answers the dreaded question that teachers asked each of us at one point or another: "Is this your best work?" It reveals a student's own standards. We think that multiple measures are more likely to show the whole picture.

There are several other features. Scoring criteria are public and let students, teachers and the public know what is expected. Benchmarks pieces are examples of actual student work that match the criteria; they make expectations concrete. Vermont teachers built 17 networks to deliver continuous professional development to enable teachers to coach students to higher performance.

One goal of our assessment was to encourage sensible discussions about performance. We created "School Report Nights" where the focus is on the student work and the standards against which it ought to be judged. These town meetings about student performance are now held in most communities. The business community helped launch that idea with us, and also took the message directly to students with a program they called Performance

* George Madaus and others, The Influence of Testing on Teaching Math and Science in Grades 4-12. National Science Foundation, October 1992, p 7.

Counts More than 250 employers have pledged to consider transcripts and portfolios when graduates come for the job interview. And they took that message directly to all the high schools in the state.

A distinguishing feature of the Vermont assessment is that is reflects a view that assessment must be a part of instruction, not apart from it. Most of the basic design decisions were teacher decisions. Our strategy in creating the assessment was collaborative and inclusive to an extraordinary degree. We have not had to mandate this initiative.

The RAND Corporation has conducted a rigorous evaluation of the Vermont portfolio assessment. The first RAND report of last July revealed profound --and very positive -- changes in classroom practice. For example, teachers have increased dramatically the amount of time spent on problem solving. It is clear from that report that teachers moved quickly to put in place the standards of the National Council of Teachers of Mathematics.

The second RAND report in December showed that while portfolio results were reliable at the state level in our first year, the reliability was too low to support reporting at the supervisory union level. RAND recommended changes in training and scoring which Vermont quickly adopted. And we are now mid-way through our second full year. And RAND will give us another evaluation in the year ahead.

Other resources have made it possible for Vermont to begin extending its original assessment in writing and mathematics to other subjects. Our Statewide Systemic Initiative grant from the National Science Foundation opens the way to an integrated math, science and technology assessment. A recent grant from the Jesse B. Cox Charitable Trust enables us to start building the first state-wide portfolio assessment in the arts. Our participation in New Standards and our recent sharing in a New American Schools grant through our membership in the National Alliance for Restructuring Schools will help us share the great design work yet to be done in assessment, as in so many other parts of the systemic reform agenda.

Change on this scale is stressful. There are critics who think this is too time consuming. But I spend a lot of time listening to teachers. They tell how challenging it is to change curriculum, assessment and instructional practices all at the same time. But they aren't backing down. I also listen to a lot of young people as they show me through their portfolios. Many of them have internalized these high standards. And they are not backing down either. But their persistence deserves bold leadership. Congress has the opportunity sustain this kind of effort on a national scale with the actions they about to take in the reauthorization.

Chairman KILDEE. Thank you, Dr. Mills.

Dr. Romberg.

Mr. ROMBERG. Mr. Chairman and members of the subcommittee, I am pleased to have the opportunity to share with you the recommendations of the Advisory Committee on Testing and Chapter 1. This committee was appointed by the Secretary of Education to give advice to the Department and to Congress about current testing practices in that program and make recommendations on possible improvements or alternatives to those current practices.

During the past year, the committee has met several times, heard testimony from a host of practitioners and experts, had several background papers prepared, and drafted a report containing a set of recommendations. The fourth draft of this report is now being circulated to the committee members for final review. Before proceeding to present the specific recommendations, it's important to see them in light of four things, four features of our deliberations.

First, the committee's primary purpose was to reinforce the significance of Chapter 1 for America's schools. Every fact, every conclusion, every recommendation in the report is aimed squarely at strengthening Chapter 1. As the largest Federal school aid program, it helps local school districts meet the special needs of educationally disadvantaged children in low-income areas.

Second, while the need for Chapter 1 has grown, the educational and organizational context in which it operates has changed. The need to reform American education has led to the establishment of national goals and, in concert with these goals, their new descriptions of the intellectual content all students should have an opportunity to learn, new understandings of organizational dynamics, new knowledge regarding the nature of human learning and growth, and far more sophisticated efforts in testing and measurement.

As you've just heard, several States are in the process of using many of these reform efforts; and, in fact, in most States and many school districts, a variety of reform efforts are underway. Chapter 1 programs need to be consistent with those reform efforts and every effort should be made to base these practices on the reform movement and, as a matter of fact, to lead the reform rather than follow or be a hindrance to the reform efforts now going on.

Third, based on the evidence that we gathered, the committee's conclusion is that the current assessment practices in Chapter 1 are out of balance for two related reasons: One, the needs for regulatory and financial accountability have led to an assessment paradigm concerned primarily with large-scale evaluation to ensure State and school procedural adherence to mandated policies. The time is now appropriate for Chapter 1 testing to concentrate more on promoting student learning and less on measuring regulatory compliance.

Two, although for Chapter 1 programs there are five different functions for which student performance information is used, a single source, test scores from a norm-referenced standardized test have become the indicator of student performance for all five functions in too many situations. These functions are selection of stu-

dents for the program, progress during instruction, and accountability at the local, State, and national levels.

In fact, it is our judgment that the norm-referenced tests serve none of the five functions well, and in some instances their use has led to detrimental effects. Our conclusion is that a new assessment paradigm is needed that matches information needs with appropriate sources of data.

Finally, the fundamental research understandings, organizational agreements, technological developments, and so forth, necessary to attain a new paradigm are awesome. Consequently, one cannot expect an immediate transition toward an outcome orientation for all Chapter 1 projects and its State and local components. It will almost assuredly take a transition period to develop and make operational the new paradigm. Yet, even if it is time consuming, we believe that the transition to a new paradigm must eventually occur, lest Chapter 1 lose its present effectiveness and fail to meet future challenges. Now let me turn to the committee's recommendations.

I have attached the draft of "Executive Summary" from the fourth draft of our report. The Committee Chair, James Guthrie, from the University of California at Berkeley, authorized me to distribute this to you, as long as you have the full understanding that this is still due for editorial revisions and even the possibility of some dissenting comments from members of the committee.

The recommendations by the committee are based on five principles. These principles are: Chapter 1 should continue to have strong accountability, but the balance should shift to emphasize how well students are learning and how effectively they are being taught.

Second, Chapter 1 testing should no longer be a separate domain, but should be linked with the educational reforms that States and school districts are undertaking for all children. Per this end, the committee proposes that Chapter 1 accountability be based on assessments that are aligned with high standards for the content of all children that all children should learn and the performance all children should attain in reading, writing, oral language, mathematics, and, to the extent possible, the other subjects associated with the national educational goals.

Third, the National Chapter 1 accountability functions should be decoupled from State and local ones. Fourth, multiple forms of assessment should be used for Chapter 1 at all levels, including performance assessments that require students to undertake an action or create a product demonstrating their ability to use knowledge and skills. Fifth, Chapter 1 assessment should acknowledge the different developmental stages of children, thus the committee recommends different assessments strategies for different age and grade levels.

To implement these principles, the committee has gone ahead to make eight specific recommendations, five associated with the specific functions in Chapter 1 and three general recommendations. They are as follows:

Recommendation No. 1: the committee recommends that the Federal Government design a national assessment to meet national accountability needs based on the sampling, quality control, and

other technical procedures used by national assessment of educational progress.

No. 2, the committee recommends that States assume a stronger leadership role in Chapter 1 assessment and accountability. States should be the linchpin of the new paradigm for accountability. Much as we have just heard from Vermont, States should develop and implement high standards for Chapter 1 content and student performance that are the same as State standards for all children.

Recommendation No. 3: The committee recommends that local accountability be closely intertwined with State accountability.

No. 4, the committee recommends that Chapter 1 explicitly endorse the use of continuous, intensive, and varied assessments for instructional feedback and diagnosis. The aim is to give teachers the encouragement and the tools they need to incorporate good assessment practices into their everyday classroom experiences and operations.

No. 5, to identify eligible students and select those with the greatest needs for Chapter 1 services, the committee recommends the use of multiple indicators, including informed teacher judgment. Thus, with respect to each of the five functions, there are specific recommendations associated with each of those functions.

Then overall, we have three additional recommendations: one, the committee recommends the schoolwide project approach in which Chapter 1 funds are used to upgrade instruction for all children attending high poverty schools as a highly desirable option for Chapter 1 services.

Next, the committee recommends a five-year transition period to commence upon reauthorization of Chapter 1. During this time, States would develop and put in place standards, assessments, and procedures.

Finally, being fully aware that staff development and research and development are two steps that are critical to the success of all the committee recommendations, the committee recommends that Chapter 1 include a set-aside for funds for staff development related to assessment and that the Federal Government also lead and fund a national research and development effort to expand our knowledge about and techniques associated with assessment and standards.

In summary, like Ms. Chelimsky and Dr. Mills, we're convinced that a new assessment paradigm centered on the children's learning in light of high standards is needed and is needed for Chapter 1 as well as for all students. Thank you.

[The prepared statement of Thomas A. Romberg follows:]

Thomas A. Romberg

Sears Roebuck Foundation-Bascom Professor in Education

University of Wisconsin-Madison

TESTING IN CHAPTER 1

The Advisory Committee on Testing in Chapter 1 was appointed by the Secretary of Education to give advice to the Department of Education and to Congress about current testing practices in the program and make recommendations on possible improvements or alternatives to those current practices. During the past year the Committee met several times, heard testimony from a host of practitioners and experts, had several background papers prepared, and drafted a report containing a set of recommendations. The title of our report, REINFORCING THE PROMISE, REFORMING THE PARADIGM, captures the two purposes our work - strengthening the program for at-risk students and suggesting a new assessment paradigm.

A fourth draft of the report is now being circulated to the committee members for final review. I have attached the draft "Executive Summary" of the report to this statement. It needs to be understood that the summary is subject to editorial changes and possible dissenting views of committee members.

Before proceeding to present the specific recommendations it is important to see them in light of four features of our deliberations: the importance of Chapter 1 in American education, the relationship of Chapter 1 programs to National Goals and reform, the need for a new paradigm, and the need for a period of transition.

CHAPTER 1

The committee's primary purpose was to reinforce the significance of Chapter 1 for America's schools and for our society. Every fact, every conclusion, every recommendation in this report is aimed squarely at strengthening Chapter 1.

Chapter 1 of the Elementary and Secondary Education Act of 1965, the largest federal school aid program, helps local school districts meet the special needs of educationally disadvantaged children in low-income areas. With an appropriation of $6.1 billion for fiscal year 1993, Chapter 1 is a common presence in American schools, touching urban, suburban, and rural schools, and children from the full spectrum of socioeconomic, racial, and ethnic backgrounds. Nearly every school district in the country receives Chapter 1 funds, which are used to mount programs in about half the nation's public and private schools, including 71% of elementary schools. Over 5.5 million students--or about one in nine U.S. school-age children--receive supplementary instruction, mostly in reading and mathematics, through Chapter 1. The needs of these at-risk students remains the priority of the program.

NATIONAL GOALS AND REFORM

While the need for Chapter 1 has grown, the educational and organizational context in which it operates has changed. During the past quarter century a growing awareness of the need to reform American education has led to the establishment of National Goals by the Governors. In concert with these goals and new descriptions

of the intellectual content all students should have an opportunity to learn, new understandings of organizational dynamics, new knowledge regarding the nature of human learning and growth, and far more sophisticated efforts in testing and measurement, a variety of reform efforts are now underway in most states and school districts. Given this impetus, every effort should be made to render Chapter 1 programs models of instruction such that they, instead of being outside the mainstream of American education, actually contribute to the improvement of instruction throughout United States schools. In effect, Chapter 1 should represent the best educational practices consistent with current knowledge and an effort should be found to use these practices as leverage on the remainder of the programs in America's schools. In fact, the same high standards should be held for Chapter 1 students as is held for the larger body of American school children.

THE NEED FOR A NEW TESTING PARADIGM

The committee's second purpose was to provoke a careful and practical reconsideration of the fundamental premises currently undergirding Chapter 1 testing and measurement. It is our conclusion that current assessment practices are out of balance for two related reasons. First, the needs for regulatory and financial accountability have led to an assessment paradigm concerned with large scale evaluation to ensure state and school procedural adherence to mandated policies. The time is now appropriate for Chapter 1 testing to concentrate more on promoting student learning and less on measuring regulatory compliance. Second, although for Chapter 1 programs there are five different decisions for which student performance information is used, a single source, test scores from a norm-referenced standardized test, has become the indicator of student performance for all decisions in too many situations. The five decisions are: selection; progress during instruction; local, state, and national accountability.

A new assessment paradigm is needed that matches information needs with sources of data. Practically this translates to the use of assessments which are operationally linked to instructional objectives for students. Chapter 1 assessments, eventually, should be tightly tied to what students are expected to know and be able to do. In effect, tests, under this new vision, would be so fundamentally integrated into regular instructional activities that students would frequently not be able to distinguish assessment from the regular flow of teaching in a classroom or school. Also, assessments should be designed with careful consideration of their appropriateness for the age, grade level and developmental stages of the students from whom they were intended. Finally, assessment would be sufficiently linked to instructional purposes that a school's professional staff could regularly rely upon their results to inform them of the degree to which their instructional strategies were succeeding, both with individual students and with groups of students.

TRANSITION PERIOD

The fundamental research understandings, organizational agreements, and technological developments necessary to attain the above-described ends are awesome. Consequently, one cannot expect an immediate transition toward an outcome orientation for all of Chapter 1 and its state and local components. As wise and well intentioned as executive and legislative branch officials may be, it will nevertheless almost assuredly take a transition period, perhaps as long as five years, to strike a creative balance between the present financial and procedural regulatory format and new, badly needed, student achievement orientation.

During this transition, proponents of change will no doubt at times become frustrated, and advocates of the status quo no doubt will feel vindicated. Nevertheless, even if time-consuming, we believe that the transition to a new paradigm must eventually occur, lest Chapter 1 lose its present effectiveness and fail to meet future challenges.

RECOMMENDATIONS

The draft "Executive Summary" from the report is attached.

EXECUTIVE SUMMARY

As the largest federal school aid program, the Chapter 1 program for disadvantaged children is an influential force in American education. Testing is one particularly strong area of influence. Millions of school children take standardized tests every year because of Chapter 1 testing requirements. Standardized tests, primarily the norm-referenced kind, are used to help determine which children should be served, assess how much participants are learning, and evaluate whether the program is effective in individual schools and for the nation as a whole. Most of these functions relate in some way to the goal of "accountability"--ensuring that Chapter 1 funds are used well, to help improve the achievement of disadvantaged children.

Few would disagree that there should be strong accountability for Chapter 1, and that this accountability should include an appraisal of student progress. But the world has changed considerably since the current Chapter 1 testing system was put in place. Knowledge about teaching and learning has expanded. New approaches to testing have been piloted or implemented. Demands for higher educational standards for all students have emerged. Consequently, new questions have arisen about whether the current Chapter 1 testing requirements are keeping pace.

The Advisory Committee on Testing in Chapter 1 was established to help answer these questions and advise the U.S. Department of Education on improvements or alternatives to the current testing system. After analyzing existing testing procedures, the Committee has concluded that Chapter 1's over-reliance on a single testing method--aggregated gain scores on standardized, norm-referenced tests--does not provide adequate information by which to judge student progress, school-level program quality, or national program effectiveness. Rather, the Committee has concluded, the current testing requirements tend to reinforce some of the more ineffective or outmoded approaches to teaching disadvantaged children, such as drilling students on low-level basic skills or giving them less challenging subject matter than their peers receive. The weaknesses of the current system have become more pronounced since enactment of the

1988 Amendments to Chapter 1, which raised the stakes attached to Chapter 1 testing by requiring schools that showed insufficient test score gains to engage in a "program improvement" process.

The Committee therefore recommends a new approach to Chapter 1 assessment and accountability, based on several important principles.

First, *Chapter 1 should continue to have strong accountability, but the balance should shift to emphasize how well students are learning and how effectively they are being taught.* The current emphasis in Chapter 1 testing is on compliance with evaluation procedures and mandates. After twenty-seven years of experience with Chapter 1, states and local districts understand and respect its basic goals and are ready for a new degree of flexibility and creativity in assessment. In exchange, however, they should be able to demonstrate that Chapter 1 children are progressing toward ambitious expectations for learning, and that schools are providing high quality instruction. To make these determinations, states and districts will need to use multiple measures aligned more closely with the types of student learning outcomes being sought.

Second, *Chapter 1 testing should no longer be a separate domain but should be linked with the educational reforms that states and school districts are undertaking for all children.* Right now several professional associations and study groups, including the panels following progress toward the National Education Goals, are in the process of developing high, voluntary national standards for what American students should know and be able to do in key subjects. Chapter 1 students should be prepared to reach those standards, or whatever high expectations states set for all children. Toward this end, the Committee proposes that Chapter 1 accountability be based on assessments that are aligned with high standards for the content all children should learn and the performance all children should attain in reading, writing, oral language, mathematics, and to the extent possible the other subjects in the National Education Goals.

Third, *national Chapter 1 accountability functions should be decoupled from state and local ones.* This would give states, districts, schools, and teachers greater flexibility to

design Chapter 1 accountability approaches that are better aligned with educational and assessment reforms for all children. It is the need for aggregated national data that has driven much of the reliance on a single form of testing.

Fourth, *multiple forms of assessment should be used for Chapter 1 at all levels*, including performance assessments that require students to undertake an action or create a product demonstrating their knowledge or skills.

Fifth, *Chapter 1 assessment should acknowledge the different developmental stages of children*. The Committee supports the concept of early intervention but recognizes that care must be taken in assessing young children, defined in this report as children below grade 3. Therefore, the Committee recommends different assessment strategies for different age and grade levels.

How can these principles be implemented? The Committee offers several specific recommendations, five pertaining to the major functions of Chapter 1 testing at the national, state, local, classroom, and student levels, and three that cut across all levels and functions.

The Committee recommends that the federal government design a national assessment to meet national accountability needs, based on the sampling, quality control, and other technical procedures used by the National Assessment for Educational Progress (NAEP). It is not necessary to test every Chapter 1 child every year to obtain a reliable national picture of Chapter 1's effectiveness. In fact, the current system of aggregating millions of test scores upward through the district, state and federal levels is an inefficient and sometimes imprecise way of doing a national evaluation. Through a NAEP-like sampling approach, a national assessment should evaluate the achievement of a representative sample of Chapter 1-eligible children in reading, writing, oral language, mathematics, science, history and geography. The assessment could be conducted on a multi-year cycle, rather than annually, and could be implemented in selected grades, beginning with grade 3. The assessment should also collect background information about Chapter 1 students and programs and analyze the long-term effects of Chapter 1 participation. A

well-designed assessment of this nature could provide Congress with better information than it receives now.

Children in prekindergarten through grade 2 should not be included in the national assessment. There should, however, be special national studies at grade 2, using performance based assessments that meet other strict criteria to ensure appropriate, sensitive assessment of young children. In addition, the Secretary should review data on program delivery for prekindergarten through grade 1.

The Committee recommends that States assume a stronger leadership role in Chapter 1 assessment and accountability. States would be the linchpin of the new paradigm for accountability. As a first phase, States should develop and implement high standards for Chapter 1 content and student performance that are the same as state standards for all children. As part of this process, states should consider whatever voluntary national standards exist for key subjects, as they become available. In the next phase, states should design and implement a system of multiple assessments for Chapter 1, including alternatives to conventional standardized tests, that are aligned with the state content and performance standards. The standards and assessments resulting from this process would be submitted to the U.S. Secretary of Education for approval and would guide Chapter 1 assessment and accountability at the state and local level.

Because it is not fair to expect students to perform at a certain level without also ensuring that they receive meaningful opportunities to learn, the Committee also recommends that states develop "delivery standards" addressing the elements, practices, and inputs that contribute to a high quality Chapter 1 program. States and local districts would use these delivery standards as a basis for evaluating program quality at the school and classroom level. As a final component of the state role, the Committee recommends that states develop procedures for local reporting of Chapter 1 assessment results and for state monitoring of program effectiveness, which should include classroom observations of program delivery.

For state and local accountability purposes, programs at the prekindergarten and kindergarten levels would be assessed on the basis of delivery standards only. At grades 1 and/or 2, there would be an assessment using both delivery standards and some student content and performance standards, provided that assessments were performance based and developmentally appropriate.

The Committee recommends that local accountability be closely intertwined with state accountability. At the option of the state, local school districts could be allowed to modify state standards and assessments or develop their own standards and assessments that met similar criteria. The accountability system of content and performance standards, delivery standards, assessments, and monitoring procedures could form a basis for determining the effectiveness of programs at the school level, as well as the progress of individual students. The Committee stresses, however, that these determinations should be based on multiple measures.

Teachers need a range of information to monitor student learning, diagnose student needs, and inform their own teaching. This "instructional feedback" function of Chapter 1 testing is one of the most important functions, but is among the most neglected by the current testing requirements. *The Committee recommends that Chapter 1 explicitly endorse the use of continuous, intensive, and varied assessments for instructional feedback and diagnosis.* The aim is to give teachers the encouragement and the tools they need to incorporate good assessment practices into their everyday classroom operations. This function of assessment should be controlled by teachers. The results of state/local accountability assessments would be just one source of feedback; teachers would determine what others were needed, which could include informed teacher judgment, classroom observations, performance assessment, and more.

To identify eligible students and select those with the greatest needs for Chapter 1 services, the Committee recommends the use of multiple indicators, including informed teacher judgment. Children at the prekindergarten through grade 2 level should be

selected for Chapter 1 primarily on the basis on poverty, with consideration of other factors that may place children at educational risk and with informed teacher judgment.

Special care should also be taken to include limited-English-proficient (LEP) children and special education children in Chapter 1 programs and assess them appropriately. For LEP children, assessments for both accountability and eligibility purposes should include an assessment of oral language.

The Committee recommends the schoolwide project approach, in which Chapter 1 funds are used to upgrade instruction for all children attending the highest-poverty schools, as a highly desirable option for Chapter 1 services. When well implemented, the schoolwide project approach fits well with the paradigm of linking Chapter 1 assessment with educational reforms for all children.

A great deal of research, development, training, consensus-building, and other work will need to be done to bring about a paradigm shift of the magnitude proposed in this report. For this reason, *the Committee recommends a five-year transition period, to commence upon reauthorization of Chapter 1.* During this time, states would develop and put in place standards, assessments, and procedures. The federal government would develop and begin to implement the national accountability assessment, and would provide staff development, research, and technical assistance to state and local agencies. By the end of five years, all elements of the new accountability system should be in place.

Upon enactment of new amendments to Chapter 1, the Committee recommends that the current system of nationally aggregated norm-referenced test data be discontinued. Instead, we recommend that states immediately develop transition plans for ensuring Chapter 1 accountability during the interim five-year period, until the new system is ready. These transition plans should be approved by the U.S. Secretary of Education and should include multiple measures of student performance and program effectiveness.

Staff development and research and development are two steps that are critical to the success of all the Committee recommendations. *Therefore, the Committee recommends that Chapter 1 include a set-aside of funds for staff development related to assessment and that the federal government also lead and fund a national research and development effort to expand knowledge about assessment and standards.*

Chairman KILDEE. Thank you very much, Dr. Romberg.

Dr. Johnson?

Ms. JOHNSON. Thank you, Chairman Kildee, Congressman Goodling, and members of the subcommittee. I thank you for the invitation to come before you this morning to address this area of concern in American education, the appropriate use of assessment.

Any consideration of assessment should be done in the context of its value to the advancement of student learning either directly or indirectly. Our goal is to enhance the quality and excitement and value of the learning experience. An assessment should be itself assessed for its contributions towards this goal.

My comments are direct l towards three general issues: the need for a national system of ssessments, the equity and validity in such an assessment, and the role of assessment in school reform. I want to begin, though, with some general background regarding assessment and measures in American society, which although perhaps a little old hat, I think has particular importance when considering the equity and fairness implications of assessment.

Americans like numbers. We depend on them for help in decisionmaking at levels ranging from national policy to personal choices, whether we're talking about the gross national product or the median family income, the daily change in the Dow Jones or the 2.1 children in the average American family, they are an important part of our daily lives. We view them as quick, reliable informers to help us cut through verbiage and make sense of it.

Assessment numbers or assessment scores are numbers with inherent egalitarian appeal. It seems logical that an assessment generates or operates as a mental yardstick so that the outcome accurately reflects the level of knowledge of the person being assessed. However, careful steps are needed in interpreting these results, and as we move towards more complex extended time assessment tasks, the problems of making reliable and valid conclusions becomes correspondingly complex.

The move towards more open-ended, instructionally related assessment tasks, rather than multiple-choice tests, has drawn renewed interest to assessment issues. Assessments are increasingly used in educational and employment decisions at all levels. These decisions involve not only the educational programming or vocational placement of the individual students or job applicants who take the tests, but also are increasingly used in the evaluation of schools, school systems, and training programs in terms of their efficacy in providing educational experiences.

In addition to assessments at all levels of elementary and secondary schools, regional college and university accrediting agencies now consider the measurement of student learning as a major parameter for determining or renewing the accreditation of higher education institutions.

This broader usage of tests and assessments has several positive outcomes. Community groups have responded to reports of low test scores in local public schools by developing school- and community-based strategies to improve curriculum. Because of these efforts, improvements in the average scores on standardized tests have been achieved by these systems.

The knowledge the test scores will be viewed as barometers of a school's progress probably sensitizes teachers and administrators to the "fit" between curriculum and test and encourages them to systematically cover curriculum. Such a use of assessments does not necessarily validate their quality as measurement devices, but rather speaks to their effectiveness as spurs to curriculum development and instructional improvement.

In a historical sense, the findings from use of assessments has documented that both African-American and white children living in the south have improved their academic performance relative to northern counterparts since school desegregation began in the 1950s. Examining norm tables from the mid-1950s subsequently through the decades since then notes the narrowing of that gap to the benefit of all southern children, and thus has also benefited commerce, technology, and culture.

As the innovative Federal education agenda emerged as we go through the 1960s and 1970s, evaluation became increasingly an important component of all programs. With the proliferation of creative ideas, well-grounded evaluation plan was an essential requirement for programs submitted for funding, particularly for Federal funding.

These evaluations included a wide variety of techniques, observations, checklists, assessments, interviews and the melding of these to provide answers on program effectiveness and helping decisions to raise, lower, or eliminate funding. But more importantly, evaluation meant a dramatic increase in the testing and assessment of children at all levels and the firm and extensive planting of a range of assessments as hardy perennials in the crowded garden of school activity. Essentially, that's how we've come to the point where we are now with the focus on assessments and their role.

The fairness of testing for minority students then has become a thorny issue not just because of the tests themselves, because we want to consider the fact that the seeds of test use in school desegregation was often scattered by the same as by opposing hands. African-American and white liberals want a test to demonstrate the effectiveness of model programs to achieve—to increase achievement in poor, often black youngsters, while reactionaries saw test scores as providing a base for recreating what they could no longer do by law: the separation of races in the classrooms.

The fairness has become a thorny issue not just because of the test, but because of the historical and contemporary both appropriate and inappropriate use. Then you have the complicated history of test development in the early part of this century, which I won't try to go into.

With that brief background, let's look at the issues that I've noted above. Is there a need for a national system of assessments? We should first note that we do have a national system of assessments. The main component of this system is the National Assessment of Educational Progress or NAEP.

NAEP provides nationally based, large sample data regarding the proficiency of students in grades 4, 8, and 12 in the basic academic areas of reading, writing, and mathematics and in other subject areas periodically. NAEP is structured so that no student is tested for extensive time periods. Essentially, they sort of get a

piece of a test, we might say, and proficiency estimates are provided for the Nation as well as all sorts of subgroups, regional, community size, and so forth. NAEP also has pioneered large-scale performance assessments and should continue to move in this direction.

The College Board SATs, the American College Test, ACTs, constitute another component of our national system. College-bound youths voluntarily take these exams and scores are provided to the colleges which they select. The advanced placement tests of the College Board are also national exams tied closely to a curriculum. Newer programs such as the Equity 2000 and Pacesetter programs of the College Board integrating instruction and assessment are additional components of a national program locally selected. Other components include the Armed Forces Vocational Aptitude Battery and various employment services examinations, as well as the nationally standardized examinations marketed by various test publishers. The issue here is whether or not the system should be further nationalized and whether the nature of the assessment should, in fact, be changed.

The National Council on Educational Standards and Testing report concluded that "National standards and a system of assessments are desirable and feasible mechanisms for raising expectations, revitalizing instruction, and rejuvenating educational reform efforts for all American schools and students."

There are critical assumptions that underlie those recommendations. The first critical assumption is that this is a desirable and healthy way to affect solid, positive learning among children; that is, using assessments for this purpose.

Two additional assumptions have been noted by Robert Linn in a recent paper. The first of these is that the establishing of clearly defined high standards and assessments with associated rewards and sanctions will motivate both students and teachers to put forth greater effort.

The second assumption is that the introduction of performance-based assessments closely aligned to national content and performance standards will, in and of itself, overcome the negative effects, particularly those experienced by teachers, with previous reform efforts based on high-stakes uses of standardized tests.

From my perspective, these assumptions are questionable. They would only have validity when students are experiencing high quality instruction and high expectations from teachers, and when teachers have been trained and have had a role in developing and implementing a system of instruction and assessment. This implies a local rather than a national base, even though national structure and formats may be used as the basis for local development.

Equity in assessment is a set of circumstances that result in measurement that is not influenced by racial, sexual, or socioeconomic background. The existence of equity in educational opportunity is an essential element of equity in assessment and cannot really be appraised in its absence. Thus, efforts to achieve equity in assessment must be simultaneously directed towards assuring that all students have equally enhancing educational experiences. This is a tall order to achieve and to evaluate.

The NCEST report notes that school delivery standards should provide the means for determining whether the school delivers to students the "opportunity to learn well" material contained in the content standards. It operationally defines this note by raising three questions: Are the teachers trained to teach the content of the standards? Does the school have appropriate and high level—and high quality instructional materials which reflect the content standards? Does the actual curriculum of the school cover the material of the content standards in sufficient depth for all students to master it to a high degree of performance?

These elements require careful evaluation before the data of student assessments, performance-based or otherwise, can be properly evaluated.

Measurement researchers are increasingly noting the need to study the consequences of the use of assessments. These include both the intended and unintended consequences of the introduction of an assessment system. Messick, in a chapter in the Educational Measurement, Volume 3, noted that evidence should address both anticipated consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness. Thus, fairness is an essential element in the determination of validity.

The change to performance assessments does not in any way avoid problems of bias or adverse impact. In fact, if children in low-income, predominantly minority schools have more limited educational experiences, the large gap between group differences in educational opportunity will likely be reflected in group differences on performance-based assessments.

We have little need to further identify these group differences. Rather, we need resources to increase the quality of educational opportunities where they are more limited. In fact, however, many schools are cutting back on teachers and closing unique multicultural schools due to lack of resources, even in the local community here in Washington.

An important equity consideration is the use of tracking children into less demanding educational experiences which essentially set limits on their eventual academic progress. The best of assessments are not likely to undo the limitations in the achievement engendered by assignment in early elementary grades to classes which do not encourage the development of thinking skills and sound academic competencies.

Some grouping practices may be well conceived and may be aimed at maximizing rigorous thinking skills among young children who may have areas of lower skill development by providing rich, heavily language-based experiences with strong performance assessments. However, most tracking is not so conceived and may have the effect of providing essentially terminal limits on student achievement.

The alluring mystique of measurement has long drawn educational reformers to its use as a technique for facilitating educational change. Movements such as minimum-competency testing and criterion-referenced testing and various thrusts towards accountability have created great flurries of measurement activity and emphasis.

While assessment should have an important place among educational activities, many of these campaigns have had the effect of demoralizing teachers and students by overemphasizing the measure to the inadequate emphasis on instructional quality, concept development, and the plain fun of learning. Fear of the consequences of lower scores may have caused teachers to decide to spend less time on the magic of discovery and more time on the discipline or drill. The outcome is that all children learn $C=2\pi r$, but few really discover what π means in the way that I discovered it in the seventh grade, and the way that students are still discovering it when teachers send them out to "measure round things," and then help them interpret what they have found.

Now, certainly good assessment, great instruction, and the excitement of learning can occur simultaneously. It seems more likely in the lower stress, local setting, but can be combined with a nationally based program. The use of well-designed assessments with varied formats, coupled with sound attention to instructional development, teacher education, and attitude change among teachers, counselors, and children are all features of some newer programs now emerging from research efforts. The Equity 2000 and Pacesetter programs of the College Board are examples.

The primary goal of Equity 2000 is to greatly increase the number of poor and minority students who graduate from college. It aims to do this by raising the expectations of teachers, students, and parents and by establishing the infrastructure to have all ninth graders successfully complete algebra. It involves institutes for math teachers and guidance counselors, awareness activities for parents, and additional information for children. Good assessments, performance-based and otherwise, are built into the program.

After considering these issues in the area of national assessment, I must conclude that there are better, more direct methods to reform education than development of a more nationally-based system of assessments than that current system. The increasing use of assessments that are more closely tied to curriculum is a sound advancement, and the move towards strong performance-based orientation is good.

The time required for such assessments and the number of assessments needed to provide reliability at the level of the individual score suggests that multiple approaches to assessment should continue to be explored. However, reform efforts should focus on instruction, teacher education, including more involvement of teachers in assessment development and use, program implementation, and also assessment, but not a primary emphasis on assessment.

Equity concerns are not met by replacing one assessment with another. In times of scarce dollars, there is a special responsibility to be sure that the educational experience of those who need the most and have had the least are not shortchanged. All children can learn. Our educational system should be directed in response to that fundamental belief. Thank you.

[The prepared statement of Sylvia T. Johnson follows:]

STATEMENT OF SYLVIA T. JOHNSON, PHD, HOWARD UNIVERSITY, WASHINGTON, DC

Good morning, Chairman Kildee, Congressman Goodling, and members of the Subcommittee on Elementary, Secondary, and Vocational Education. Thank you for the invitation to come before you this morning to address an area of genuine concern in American education today: The appropriate use of assessment. Any consideration of assessment should be done in the context of its value to the advancement of student learning either directly or indirectly. Our goal is to enhance the quality and excitement and value of the learning experiences, and assessment should be assessed for its contributions towards this goal. My comments this morning will be directed toward three general issues:

1. The need for a national system of assessments.
2. Equity and validity in assessment.
3. The role of assessment in school reform.

However, I will begin with some general background regarding assessments and measures in American society.

Americans like numbers. We depend on them for help in decisionmaking at levels ranging from national policy to personal choices, the gross national product [GNP] to the median family income, the daily changes in the Dow Jones Index to the 2.1 children the average American family rear. Numbers are an important part of our daily lives; we view them as quick, reliable informers to help us cut through the verbiage and make sense of it.

Assessment scores are numbers with inherent egalitarian appeal. It seems logical that an assessment operates as a "mental yardstick," so that the outcome accurately reflects the level of knowledge of the person being assessed. However, careful steps are needed in interpreting these results, and as we move toward more complex extended time assessment tasks, the problems of making reliable and valid conclusions becomes correspondingly complex.

The push toward more open-ended, instructionally-related assessment tasks, rather than multiple-choice tests has drawn renewed interest to assessment issues. Assessments are increasingly used in educational and employment decisions at all levels. These decisions involve not only the educational programming or vocational placement of the individual students or job applicants who take the tests, but also are increasingly used in the evaluation of schools, school systems, and training programs in terms of their efficacy in providing educational experiences. In addition to assessments at all levels of elementary and secondary schools, regional college and university accreditation agencies now consider the measurement of student learning as a major parameter for determining and renewing the accreditation of higher education institutions.

This broader usage of tests has had several positive outcomes. Community groups have responded to reports of low test scores in local public schools by developing school- and community-based strategies to improve curriculum. Because of these efforts, improvements in the average scores on standardized tests have been achieved by these school systems. The knowledge the tests scores will be viewed as barometers of a school's progress probably sensitizes teachers and administrators to the "fit" between curriculum and test and encourage them to systematically cover the curriculum. Such use of tests does not necessarily validate the quality of the tests as measurement devices, but rather speaks to their effectiveness as spurs to curriculum development and instructional improvement.

Findings from use of tests have documented that both African-American and white children living in the south have improved their academic performance relative to northern counterparts since school desegregation began in the 1950s. An examination of the norms tables for southern youths at various grade levels during the mid-1950s and through the subsequent three decades that this gap had narrowed to the benefit of all southern children, and thus has also benefited southern commerce, technology, and culture.

For example, prior to the 1954 integration of the District of Columbia Schools, the then predominately white school system had a composite mean test score for African-American and white sixth graders that was 2.3 years below the national norm. By 1958, with an 84 percent black population in these schools, sixth graders were achieving at a level equal to or above the national norm.

As the innovative Federal education agenda emerged in the 1960s and evolved throughout the 1970s, evaluation became an increasingly important component of all programs. With the proliferation of creative ideas, a well-grounded evaluation plan was a requirement for programs submitted for funding. Evaluations included a wide variety of techniques—observations, checklists, assessments, interviews—and the melding of these to provide answers on program effectiveness, and helping deci-

sions to raise, lower, or eliminate funding. But more importantly, evaluation meant a dramatic increase in the testing of children at all levels, and the firm and extensive planting of a range of assessments as hardy perennials in the crowded garden of school activity.

The seeds of test use and school desegregation were as often scattered by the same as by opposing hands. African-American and white liberals wanted tests to demonstrate the effectiveness of model programs to increase achievement of poor, often black youngsters, while reactionaries saw test scores as providing a basis for recreating what they were no longer able to do by law; the separation of races in the classroom.

Thus, the fairness of tests for minority students is a thorny issue, not just because of the tests themselves, but because of the historical and contemporary appropriate use as well as misuse of the tests.

With this brief background, let us examine the first issue noted above: Is there a need for a national system of assessments? It should be first noted that we have a system of national assessments. The main component of this system is the National Assessment of Educational Progress [NAEP]. NAEP provides nationally based, large sample data regarding the proficiency of students in grades 4, 8, and 12 in the basic academic areas of reading, writing, and mathematics, and in other subject areas. NAEP is structured so that no student is tested for extended time periods, and proficiency estimates are provided for the Nation, as well as for regional subgroups. NAEP has pioneered large-scale performance assessments, and should continue to move in this direction.

The College Board SATs, the American College Test, [ACT] constitute another component of our national system. College-bound youths voluntarily take these examinations, and scores are provided to the colleges which they select. The advanced placement tests of the College Board are also national examinations. Newer programs such as the College Board Equity 2000 and Pacesetter programs, which integrate instruction and assessment are additional components locally selected.

Other components of our national system include the Armed Forces Vocational Aptitude Battery, and various employment services examinations. The nationally standardized examinations marketed by various test publishers might also be considered as elements of our existing national examination system. The issue here is whether or not this system should be further nationalized.

The National Council on Educational Standards and Testing [NCEST] report concluded that "National standards and a system of assessments are desirable and feasible mechanisms for raising expectations, revitalizing instruction, and rejuvenating educational reform efforts for all American schools and students" (Page 8). There are critical assumptions underlying these recommendations. The first assumption is that this is a desirable and healthy way to effect solid, positive learning among children. Two additional assumptions have been noted by Robert Linn (1992). The first of these is that establishing clearly defined high standards and assessments with associated rewards and sanctions will *motivate* both students and teachers to put forth greater effort. Linn further notes that the NCEST recommendation assumes that the introduction of performance-based assessments closely aligned to national content and performance standards will overcome the negative effects, particularly those experienced by teachers, with previous reform efforts based on high-stakes uses of standardized tests.

From my perspective, these assumptions would only have validity when students are experiencing high quality instruction and high expectations from teachers, and when teachers have been trained and have had a role in developing and implementing a system of instruction and assessment. This implies a *local* rather than a *national* base, even though national structure and formats may be used as the base for local development.

Equity, Validity and National Assessment

Equity in assessment is a set of circumstances that result in measurement that is not influenced by racial, sexual, or socioeconomic background. The existence of equity in educational opportunity is an essential element of equity in assessment, and cannot really be appraised in its absence. Thus, efforts to achieve equity in assessment must be simultaneously directed toward assuring that all students have equally enhancing educational experiences.

This is a tall order to achieve and to evaluate. The NCEST report notes that school delivery standards should provide the means for determining whether the school delivers to students the "opportunity to learn well" material contained in the content standards. It operationally defines this note by raising three questions:

1. Are the teachers trained to teach the content of the standards?
2. Does the school have appropriate and high quality instructional materials which reflect the content standards?
3. Does the actual curriculum of the school cover the material of the content standards in sufficient depth for all students to master it to a high standard of performance?

These elements require careful evaluation before the data from student assessments, performance-based or otherwise, can be properly evaluated.

Measurement researchers are increasingly noting the need to study the consequences of the use of assessments. These include both the intended and unintended consequences of the introduction of an assessment system. Messick (1992) notes that evidence should address both anticipated consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness. Thus, fairness is an essential element of the determination of validity.

The change to performance assessments does not in any way avoid problems of bias or adverse impact. In fact, if children in low-income, predominantly minority schools have more limited educational experiences, the large gap between group differences in educational opportunity will likely be reflected in group differences on performance-based assessments. We have little need to further *identify* these group differences. Rather, we need resources to increase the quality of educational opportunities where they are more limited. In fact, however, many schools are cutting back on teachers and closing unique, multicultural schools due to lack of resources.

An important equity consideration is the issue of tracking children into less demanding educational experiences which essentially set limits on their eventual academic progress. The best of assessments are not likely to undo the limitations in the achievement engendered by assignment in early elementary grades to classes which do not encourage the development of thinking skills and sound academic competencies. Some grouping practices may be well conceived, and may be aimed at maximizing rigorous thinking skills among young children who may have some areas of lower skill development by providing rich, heavily language-based experiences with strong performance assessments. However, most tracking is not so conceived, and may have the effect of providing essentially terminal limits on student achievement.

Assessment and School Reform

The alluring mystique of measurement has long drawn educational reformers to its use as a technique for facilitating educational change. Movements such as minimum-competency testing and criterion-referenced testing and various thrusts towards accountability have created great flurries of measurement activity and emphasis. While assessment should have an important place among educational activities, many of these "campaigns" have had the effect of demoralizing teachers and students by overemphasizing the measure to the inadequate emphasis on instructional quality, concept development, and the fun of learning. Fear of the consequences of lower scores may have caused teachers to decide to spend less time on the magic of discovery and more time on the discipline or drill. The outcome is that all children learn $C=2\pi r$, but few really discover what "π" means in the way that I discovered it in the seventh grade, and the way that students are still discovering it when teachers send them out to "measure round things," and then help them interpret what they have found.

Now, certainly good assessment, great instruction, and the excitement of learning can occur simultaneously. It seems more likely in the lower stress, local setting, but can be combined with a nationally based program. The use of well-designed assessments within varied formats, coupled with sound attention to instructional development, teacher education, and attitude change among teachers, counselors, and children are all features of some newer programs now emerging from research efforts. The Equity 2000 and Pacesetter programs of the College Board are examples.

The primary goal of Equity 2000 is to greatly increase the number of poor and minority students who graduate from college. It aims to do this by raising the expectations of teachers, students, and parents and by establishing the infrastructure to have all ninth graders successfully complete algebra.

The project involves institutes for math teachers and guidance counselors, awareness activities for parents, and additional information for children. Good assessments, performance-based and otherwise, are built into the program.

Pacesetter involves a well-designed set of curricular experiences, interspersed with assessments, to build high-level competencies in eleventh or twelfth grade level courses in the major academic areas. The courses are designed as "capstones" to a high school experience. While this nationally-based program can be adopted by local school systems, it does not carry the same high-stakes baggage that an external pro-

gram that is not so essentially connected to instruction might have. The teacher education-instruction-assessment connection also makes the issue of equity more assured.

After considering these issues in the area of national assessment, I must conclude that there are better, more direct methods to reform education than development of a more nationally-based system of assessments than our current system. The increasing use of assessments that are more closely tied to curriculum is a sound advancement, and the move toward a strong performance-based orientation is good. The time required for such assessments, and the number of assessments needed to provide reliability at the level of the individual score suggests that multiple approaches to assessment should continue to be explored. However, reform efforts should focus on instruction, teacher education program inplementation and assessment, rather than a primary emphasis on assessment.

Equity concerns are not met by replacing one assessment with another. In times of scarce dollars, there is a special responsibility to be sure that the educational experience of those who need them most, and have had the least, are not shortchanged. All children can learn. Our educational system should be directed in response to that belief.

Chairman KILDEE. Thank you, Dr. Johnson.

Since this is the first hearing of the 103d Congress on assessment, I'm going to really ask a very fundamental question of experts. Could you summarize, each one of you, what you perceive to be the purpose or purposes of testing? We hear many things, track students to improve individual performance, to improve school or delivery performance. How should testing affect learning, or how should testing affect teaching?

Dr. Mills, do you want to start with this? It's a very fundamental question, but I would like to maybe start off this year with those answers.

Mr. MILLS. It is a fundamental question, and thank you. I would begin with the title of our proposal in Vermont, back in 1988, when we began to develop, what we called, "Working Together to Show Results." I think the answer is in that title. We believe that the fundamental purpose is to boost performance; it's not to find out who is not doing well.

It's to boost performance, to provide information that enables teachers to see that they need to try other strategies, to give them opportunities to learn those other strategies, to give students the opportunity to internalize standards that matter to them. It sets in motion a series of actions involving parents, teachers, school boards, or the general public, that enables students, all students, to perform at a higher level.

I can tell you what that looks like in a school. I could tell you what it looks like when a fourth grader explains to me what his portfolio is all about. Let me put that aside and get to the other part of the question. I think that the other related purpose of assessment is to spark a sensible discussion about results, about performance. We do not have such a discussion in this country right now.

We have even something as fine as NAEP, the National Assessment for Educational Progress—and I don't want people to misunderstand this—but I think most people use it for anecdotes for speeches. I can remember doing this myself. It's full of "gee whiz" statements about "Do you know that seventh graders—or 17-year-old kids can't do the following?"

What we need in place of these anecdotes is a probing, continuous, searching discussion about results: "What does performance

look like here in our town? Is that good enough for us? If it's not good enough, what are we going to do about it?" Not who are we going to pin it on, but how do we boost performance?

That's the way high performing teams behave, high performing businesses, high performing military units, high performing classrooms. Any high performing group craves information about results, information that they can trust. They talk about these sorts of things. We need an assessment system that does both of those things, inspires performance and enables people to talk sensibly about it.

Chairman KILDEE. Ms. Chelimsky?

Ms. CHELIMSKY. Yes, thank you. I would agree with what Dr. Mills said. As an evaluator, I believe that one should always know where one is with regard to a program that is being dealt with and that is being given to many, many people. I think it's extremely important in all government programs that evaluation be done, that testing be done, not just to be able to talk about it and have a debate, but just to know whether you're doing well, or you're not doing well. I just think that's a crucial national duty.

I also think that the purpose of testing, although it's often said to be to improve student achievement, it's not really clear to me that testing alone will ever do that. Testing is important in the sense that when you find out that you're not doing well it has a shock value. I think that it's quite likely that somebody will, as Sylvia Johnson said a minute ago, start thinking how to improve local and community relationships, how to improve curricula.

Although I think testing has a role in moving toward improved student achievement, I don't think it's the only thing, and I think you need to do other things as well. I think there is teacher training. In other words, I'm concerned about applying that purpose as a monolith be-all and end-all of testing will improve student achievement. As I mentioned earlier, the Canadians haven't been able to show that. They don't have an evaluation that says it has improved achievement, despite the fact that they have done admirable things.

The other thing I feel about testing is that it should be able to inform a diagnosis about what is the matter in the way we're doing things, not incriminate children who can't, who haven't, who are not successful, as you said a minute ago. But rather, if this isn't working, then what should we do to change it? So I think it inspires also some questions about what can be done.

Chairman KILDEE. Dr. Johnson?

Ms. JOHNSON. Well, I agree with what has been said. I do think that anything that happens in a school should be evaluated in terms of its contribution to student learning, and so assessment and testing therefore in that context must essentially be viewed in the same light. I do see an important monitoring role for assessment, but that in and of itself also should contribute. The fact of that monitoring should contribute in some way to the improvement of educational opportunities and therefore to student learning. I don't see assessment as the primary tool for change, as I outlined earlier.

Chairman KILDEE. Dr. Romberg?

Mr. ROMBERG. Yes. In our deliberations, we thought there were three primary purposes for gathering test data. No, there are decisions that need to be made, and you would like to have some reliable, valid information to help you make those decisions. One has to do with teachers monitoring student progress.

As we're talking about setting new goals and standards, one of the things we find is that many teachers have been de-skilled to such an extent that they don't believe their own judgments about listening to kids, about observing what students are doing are valued, and what really is valued is a test score that someone else has written. One of the big tasks is to get teachers to be better monitors of their own students' progress.

Second, information is needed about students' profiles, about their ability and their capability of doing different things for a variety of selection purposes. Whether it's for college or military or whatever, there are needs for having information about students on a variety of different attributes or characteristics.

Third, there is the important aspect of accountability, of external tests to judge how resources being allocated at the local or State or national level are being used and the effect of that. I agree with Dr. Mills that much of that needs to be as a part of a discussion in relationship to goals. It needs to be said that this is a direction that we need to go, rather than simply saying this is your rank in relationship to others.

Chairman KILDEE. I guess I ask the question for this reason: Can assessment help us maybe determine whether standards in a delivery system need changing, whether even a certain teaching methodology might be better than another in transmitting a certain type of education? I'll give you an example. I used this before.

When I was teaching Latin, th sequence of tenses, when you move from the indicative mood to the subjunctive mood which tense to use, was very clear to me, and I would teach that to my students. I began to realize that really not many of them were getting it.

I got that through my testing, you know, through assessment. They just weren't getting what tense to use there so I kept going back, changing my methodology and finally hit upon a method—that I felt I should have patented—that really did teach even my slower students the sequence of tenses. I did go back and had to change my methodology.

I'm just wondering whether assessment can help us in our teacher training institutions or to identify what methods might work best, or even assess the standards in a certain school system and change, upgrade maybe some standards. I can predict, for example, how students will do in certain schools in this country. Some will do very well, and in other school districts, they are going to do very poorly because maybe the standards are different. Could you comment on the methodology and standards?

Mr. ROMBERG. Your example is a perfect example of what I think is important for teachers monitoring student progress. As you begin to see students aren't achieving what you would like them to achieve, you change your methods in light of trying to get them to reach those goals. That indicates that you have a good sense of

what those goals are, what you expect students to be able to do, and so on.

You know, that's the instructional importance, that you need to be able to say these are the instructional standards, these are the kinds of things kids ought to have an opportunity to learn. You are continually monitoring progress toward those goals. That's what teachers should be doing. Unfortunately, in too many classes I don't think that's happening.

Ms. CHELIMSKY. It also shows a certain inner security that you have. That's what evaluation is for. The truth of the matter is that in so many cases when evaluation is done and you see that a methodology ought to be changed or that there is a better way of delivering services, or whatever, instead of looking at how can you change that and how can that be better, what you say is, "Oh, my God, somebody is going to look at that, and they're going see that our scores are lower than somebody else's scores."

You end up with problems of people reacting in the wrong way. That's why I feel that safeguards are often necessary so that people can feel secure enough to experiment with those things. Those standards that we're talking about are dynamic, they're iterative, they move up.

Chairman KILDEE. Dr. Johnson, you had——

Ms. JOHNSON. The kind of experience you're talking about having a teacher, I think, is vital, and it's difficult to have it in a high-stakes setting; that is, it requires a few things. It requires some background in terms of assessment. It doesn't have to be all that formal, but one does have to pay attention to the kinds of assessments that are being developed and what sort of thing you are trying to get at measuring and the extent to which you are, in fact, measuring it reasonably. One has to feel safe enough to be able to use these assessments and to modify them. I think that that is an important use of assessment and probably a use that is best done at the local level.

Chairman KILDEE. Dr. Mills, do you have——

Mr. MILLS. Just a brief addition to that. I think that standards very clearly drive instructional practices, if they are the right kind of standards, if teachers have participated in building them and believe in them. I've seen evidence of that in a Rand Corporation evaluation of our first year of assessment. In mathematics we embraced the standards of the National Council of Teachers of Mathematics, which as you know probably stress problem solving among other things.

We saw a dramatic shift in the way teachers use their time in the teaching of mathematics. They saw that standard, they helped write it, it was theirs. They said, "You know, the way we teach mathematics is not promoting problem-solving." The results of their teaching, since it went into the student portfolio, was daily apparent to them.

Also, the standards, as a couple of us have said here, have evolved and they are not fixed. That's because these standards are, or at least should be, public. In Vermont we have something called "School Report Night." It's a voluntary thing, but there are, oh, probably 140 or 150 communities that did this last year. It's kind of

a town meeting focused on results. It's show and tell on a massive scale.

What happens is that the focus is not on the teacher or on the school, it's on the work that is done in this place called school. Parents look at the work and then they look at the standards, and then there is a real intense conversation. "How do we know this stuff is good? Are these standards good enough?" There's continuous upward pressure and a pressure to make standards free of jargon—clear, plain speech. This is what quality looks like.

Chairman KILDEE. Thank you very much. There is a vote on in the House right now. Mr. Goodling, do you want to start some questions now?

Mr. GOODLING. I think I can do mine in a few minutes.

Chairman KILDEE. Okay. You take all the time you want.

Mr. GOODLING. The Chairman wasn't my Latin teacher, so I learned "amo, amas, amat, amamus, amatis, amant"——

Chairman KILDEE. Perfect.

Mr. GOODLING. [continuing] so that I could, as a matter of fact, give it back on an assessment so that I would do well. I didn't really know what it was all about. I learned about what "amo" means later on in life.

[Laughter.]

Mr. GOODLING. Just a couple of very quick comments and perhaps question. Ms. Chelimsky, I'm sure you were referring to standardized tests when you gave the 7 hours and the 3½ hours?

Ms. CHELIMSKY. Yes. Yes, sir, standardized tests.

Mr. GOODLING. Because in some places they are tested to death otherwise.

Ms. CHELIMSKY. Yes, that's right, but they are specialized, they are different. No, we just looked at standardized tests.

Mr. GOODLING. I made the mistake—the Chairman and I served on the National Education Goals panel last year when we had the assessment people before us and sometimes my choice of words is not the best—and I made the mistake of saying, "Any idiot could design the test if we knew exactly what it was we wanted to assess." Of course their response was, "Perhaps not any idiot." There's probably some truth to that.

Dr. Mills, in listening to your testimony, I'm assuming that as the result of your testing programs, remedial instruction is number one, as far as the purpose is concerned. If it is, how do you control it from becoming what some people mentioned, that possibly ranking of schools and those kind of things, things that probably don't help us?

Mr. MILLS. Well, we control for it. We've avoided the problem of ranking by having several criteria. There are seven criteria on the assessment in math and five in writing. As groups of students have explained to me, I don't know, here's the voice of an eighth grader: "I'm doing quite well on this fourth criterion. I mean, I've really got it. I'm terrific at that. But this last one here, no matter what I do it doesn't work."

It's possible to be a high performer on part of it, on part of the criteria and on some of the criteria, and perform quite differently on others. Consequently, you can't really rank schools, and we don't want to rank schools. It's analogous to—well, if you're trying

to look at investment opportunities and you—it's probably a terrible analogy. But if you're looking at a bunch of balance sheets, you might calculate several different ratios: what is their return on assets, what is the inventory turnover, and a number of other things.

You might do the same thing if you're trying to compare a Toyota to a Ford to a something else and you're looking at Consumer Reports. You look at a lot of different criteria, and it's a matter of thinking and judgment. I think that's the way it should be.

Mr. GOODLING. In answer to my question, the remedial effort is very much in evidence when you look at the result?

Mr. MILLS. No. I think remedial——

Mr. GOODLING. Our whole purpose of getting teachers to test was to see where they did poorly in presenting the material and then going back and doing something about changing the direction, et cetera, et cetera.

Mr. MILLS. Okay. I thought you meant something else.

Mr. GOODLING. That's my fear, you know, sometimes when we talk about national standards and national assessments, and so on. I want to make sure that we don't lose the whole idea of this for remedial purposes; this is to help children improve.

Mr. MILLS. Yes. I think our whole system is charged throughout with an attitude of remediation, that we're going to somehow correct problems and pull children out to do something special to them. Meanwhile, education goes on by them. I would very much like to see—and I think many, a great many of my colleagues would like to see—us learn how to do it right the first time so that we don't have this sort of "catch-up" attitude, because we can't really ever help a child catch up.

Mr. GOODLING. I'm probably going to just have to make a couple of comments because the Chairman and I can't miss this important vote. I'm sure it's approving the journal or something that's earth-shattering.

Chairman KILDEE. It does not mean you have to approve the President's speech last night, just the technical approval of the journal.

Mr. GOODLING. I always vote no anyway, because I don't read the journal, and I don't know if the Speaker does or not. I don't want to approve it if someone hasn't read it.

Dr. Johnson, the three general issues that you've talked about, if we accept No. 1, I'm assuming No. 2 and 3 are very, very critical, if we accept No. 1?

Ms. JOHNSON. If you make the decision to go with a national examination system, certainly it would be essential that one would really, I think, have to look at, though whether or not this is the assessment in and of itself, would be adequate to try to reach, to try to carry out school reform. Essentially, it would be, I think, that if you tried to do it only with assessment it would essentially be like trying to carry a Latin sentence without the accusative case, you know.

Mr. GOODLING. Very good. I would have added—in your No. 1 assumption, you said, first, there has to be high-level instruction and, secondly, high expectations from teachers. I would add parents and

community to that, if we're going to be successful. My concern, of course, as I said, with Chapter 1 is, I always thought it was supposed to be something over and above everything else that everybody else got. I think in many instances that isn't what happens. I am so glad—and, Dr. Romberg, I was saying, "Amen, amen, amen," to those first five things you said there, because I think an awful lot of Chapter 1 youngsters have been cheated over the years and didn't get everything else plus more, and that's what I thought Chapter 1 was all about.

Chairman KILDEE. If you could wait—we'll be right back, we have to dash over there for a vote—for some further questions. We will be back immediately. We'll take a little recess so you can take a seventh inning stretch.

[Recess.]

Chairman KILDEE. I thank the panel for indulging us in our vote over there in the House. We have 15 minutes ordinarily to respond, and the Speaker has sent, probably his annual message out to everybody, stick to 15 minutes. We're rushing over there for a few days more quickly than usual to make sure we get our vote in. That's probably the teacher in Bill and I, more than the politician, that we generally try to be on time and try to be there every day for all the votes. I haven't missed a vote, I don't think, in 9 years.

I think Mr. Goodling, Bill, my good friend, has a few more questions.

Mr. GOODLING. Well, I just wanted to wrap up by thanking you very much for coming, because I think the testimony that you've provided is very, very important as we look at Chapter 1, Head Start, and so on, and we talk about assessment. Again, I go back to those five suggestions that Dr. Romberg gave and they are so, so very important in my estimation.

I was telling the Chairman on the way over, No. 3 talks about flexibility, and I've been preaching flexibility until I'm blue in the face. We haven't gotten very far because we're still on this access bit, and we don't think in terms of the quality and we only think in terms of access. Hopefully, that's going to change. Then, the fifth one—I think all of you were saying the same—is very, very important because you're talking about acknowledging the different developmental stages.

Again, I said to the Chairman on the way over, my wife teaches first grade and she so many times says half of her class are not ready to learn to read until after Christmas. If we say, well, somehow or other, "Grade 1, you must know such-and-such and so-and-so"—we've got to get rid of the grade system if we're going to get into the standard assessment system, or we will run into trouble because of the different developmental stages.

Again, I thank you very much for your testimony. I think it will be very helpful as we talk about reauthorizing Chapter 1 and Head Start.

Chairman KILDEE. Thank you.

Mr. Green?

Mr. GREEN. Thank you. Thank you, Mr. Chairman.

Congressman Goodling, I'll give you an example of what happened in Texas, since your wife teaches first grade. In 1984, we passed a reform bill, and we tested everybody, starting in first

grade. We finally realized that if we could just keep the kids from chewing on the test, maybe they could take them. It took us 3 years to abolish the first grade test in the legislature because we would only meet once every 2 years. It took us a while because we were testing everyone.

I just have some questions for the whole panel. Because again, from experience and the discussion particularly on Canada because along the southwestern border we have similar bilingualism that Canada has to deal with, with Quebec. I was going to ask how Canada has addressed a bilingual testing situation that they have had with Quebec as compared to what—we don't necessarily have bilingual testing in Texas, but with the growth and the number in the number of students who are Spanish-speaking, you know, it's something that has been talked about.

Ms. CHELIMSKY. We didn't look at that in this particular study, but we did happen to look at that in the study we did about 4 years ago, so I can speak a little bit to it.

Mr. GREEN. Okay.

Ms. CHELIMSKY. We were looking essentially at the possibility of using the immersion system, the total immersion system, that they have in Canada as a possible way of dealing with bilingual instruction in the United States. What we found was that it wasn't really applicable because it was the British that were learning French, people who were basically very comfortable with their status in the society and very well educated to begin with, except in the French language. The success of that method, there is an undoubted success of it in Canada—I wouldn't, you know, question that—but the question of whether that would work in the United States is a whole different thing.

Mr. ROMBERG. Let me respond also to the problem—because we did address it in Chapter 1—and that is, there are a large number of limited-English-proficient children who are typically either included in Chapter 1 or sometimes excluded for a variety of reasons, depending upon State and so on. One of the issues, one of the central technical issues, is how to best develop assessment techniques and get teachers to use reasonable techniques that give them an opportunity to indicate their progress in their language as well as in ours.

This is not an easy task, but it's something we certainly need to address. What has happened in Chapter 1 quite often is simply to eliminate these students from the testing program because they won't test well, and, therefore, they don't get included in the kinds of services that Chapter 1 is intended to provide.

Mr. GREEN. I think that's one of the concerns because those students are there and they have to be served and we have to develop some type of test. Typically, a bilingual student will do much poorer on the standardized testing, and it's because of the language barrier.

Ms. CHELIMSKY. Exactly. They do in Canada have exams and assessments in both French and English.

Mr. GREEN. I'm sure that's a requirement from what little bit I understand about Canada.

Ms. CHELIMSKY. It is.

Mr. GREEN. Although we're not to that point, we just want to make sure that we can address it, you know, for those. Not just in the southwestern States, but even in other urban areas of the country, there is a question. Again, my experience is in Texas with a lot of testing that we have done since the mid-1980s, and we're still having what we call an exit exam. It's for the eleventh graders and the twelfth graders.

We have a problem and I'm sure it's not anything new around the country. My wife teaches high school algebra and so, you know, algebra is not always on the standardized exit exam, but they're under a great deal of pressure to teach that test. Even though it's not "high order math," it's on there. How have other States been able to deal with that?

We all read the newspapers in Texas and we want to see where our schools rank and how successful they are in having completion of these tests. Yet, we don't want, particularly in some of the different classes, to actually just teach a test. That's what is happening, and I know it's probably happening in other parts of the country. Could that be dealt with, or is there some secret that maybe Texans haven't heard about?

Mr. ROMBERG. Ms. Chelimsky gave one answer, and let me just elaborate upon it, that many people are looking at, and that is, tests associated with specific courses. One of the things the Canadians do is, they don't just give tests, it's a test associated. If kids have had algebra, then algebra is included. One of the difficulties with many of our "general tests" is that they end up being, at most, tests of seventh and eighth grade arithmetic or something of that nature.

Even at, say, the exit exam that you are talking about, having very little of the mathematics that one would typically teach in high school on those exams. As a result, you take time away, you test on a number of skills or items that may not be that important, certainly don't reflect what the students have been studying the last 2 or 3 years. One of the things that people do in other countries—Canada, Australia, England—is their tests are associated with the course of study that students have had and not some general collection of items.

Mr. GREEN. Okay. It's not necessarily a standardized, it's one based on their level that they had?

Mr. ROMBERG. Right.

Mr. GREEN. Thank you. I appreciate that.

Chairman KILDEE. Thank you, Mr. Green.

I have a question for Dr. Mills. In light of Vermont's experience, what would you advise national testing proponents about the difficulty, the time. and effort required to develop such a system?

Mr. MILLS. I would say to them it is difficult and it takes time. We cannot get a high quality system of assessments together quickly. We adopted the strategy of starting with a pilot effort in—well, we wanted to start in 1988, but we weren't able to convince the legislature of that until the following year.

We started in the summer of 1988 with lots and lots of discussion that kept getting wider and wider and wider, lots of discussion with the legislature, eventually an appropriation to begin a pilot in 40 schools. It actually involved 139 because there were a lot of volun-

teers beyond the pilot, but we've just finished the first year, the first full year where everyone was involved in writing and mathematics in two grades.

There's a certain sense that things can be accelerated. Once you get the development cycle down, you discover all kinds of problems that you make. If you conduct an evaluation, as we did, then you have those issues right in front of you so you can work on them. I would hesitate to say that you can accelerate it too much. We just won a major grant to bring a statewide portfolio assessment in the arts to Vermont, and even there we're envisioning a four-year effort to think about it, talk about it, design it, test it, perfect it, make mistakes, fix the mistakes.

One issue that is abundantly clear to me as I watch and as I reflect on our experience is that you need to think through—we need to think through how much training, professional growth opportunities we offer to teachers. Really that's the fuel that drives change. You raise the standards in a public way, and you have to enable people to reach them. That means professional development, and that takes time.

Schools, school calendars are not designed to provide for that time. You need to be thinking hard about how you're going to use the summer. We've learned that. I think there is probably a sense that people might be pressed to over-promise how fast it can be done for fear that if you don't match some imagined political timetable, then there won't be the support.

I've found in dealing with the Vermont legislature, that they are enormously patient. They are not saying, "Give us a quick answer," they are saying, "Give us a good answer. Give us a system that works. Take time and do it right." If Congress can do the same thing with a mixture of pressure and support, and part of the support is enough time to do it right, I think you would be well served.

Chairman KILDEE. I think Congress moved rather cautiously and carefully last year.

Mr. ROMBERG. Not that slow.

Chairman KILDEE. Not quite as slow as last year, right? Thank you. That's a good point you made there. We've been criticized on both sides on that. Did you experience more difficulties than what you had anticipated when you started the process there in Vermont?

Mr. ROMBERG. No, I don't think that there were more difficulties than anticipated, because it's a very small State and the expectation is that every leader—the governor, the school board, everybody, the commissioner—they are out there in the community all the time. I heard from teachers all along the way, "We need more training, we need more time, we need more this and more that." I think we're about where I thought we would be, maybe one step short.

Last year, I wanted to be able to report statewide data and supervisory union-level data on eliable basis, and we found that we could only take the statewide step, not the supervisory union step. Next year, we want to go down to a school level. So, it's a progression like that. We will have to wait and see, because we have all made a pledge that we're going to report only the data that are re-

liable that stand up against national standards of quality. You can't fake it.

Chairman KILDEE. Dr. Johnson?

Ms. JOHNSON. I think it's important to note in terms of the careful explication that Dr. Mills is doing on the Vermont experience that there's so much emphasis on professional development and on the involvement of teachers at every phase. It's very different from an assessment model that essentially is externally imposed and while dealing with standards, does not involve the people who, in fact, have to build and engender and develop those standards within children. I think that's an important feature that needs to be carefully noted, especially when there's the consideration of moving to a broader base, to larger than the State of Vermont.

Chairman KILDEE. Dr. Romberg, the National Council of Teachers of Mathematics has received, I believe, well-deserved praise for developing national content standards for mathematics. What is required to develop assessments aligned to those standards?

Mr. ROMBERG. That's a very good question, and you ought to be aware of the fact that the National Council of Teachers of Mathematics as of this last month has agreed that they are going to try to develop a set of assessment standards in relationship to the contents standards to give some sort of direction in relationship to mathematics as to what are the key features that we think are important. If you're talking about these as the content standards we want all students to have an opportunity to learn, then what are the kinds of performances we would expect from children in relationship to those.

What is it going to take? When we began this whole process of working on the curriculum standards, that was 10 years ago, as of about this time of the year. Shortly after the report, "A Nation at Risk and Educating Americans for the 21st Century," the math community began starting to deal with issues about, "Well, what is it that students really ought to know in mathematics?"

We felt at the time that it was going to take until the turn of the century to make this happen in American schools. What we argued was that you first needed to talk about what was the content that you wanted students to learn. Then you wanted to turn and talk about and what are the conditions, the delivery standards associated with that, and we produced the standards for professional teaching of mathematics.

Then the question is, if you can do—and this is what Sylvia said earlier—you need to set standards and you need to then begin talk about how you create a system to reach those standards. Then you ought to start talking about, and what is the evidence that we can gather that kids are, indeed, beginning to reach those standards, what the problems are, and so on.

I would think that another—well, of course what the committee on Chapter 1 said, 5 years of investment and time and effort is going to take as a minimum to develop the good kind of system we have in mind that will give you the kind of information that we need. It isn't going to happen overnight.

It's great to see places like Vermont and California and Connecticut. Actually, the last count, it was over 30 of the States are trying now to develop assessment procedures associated with the

NCTM standards. It's going to take time and effort and resources and, above all else, staff development work with teachers to pull this off.

Chairman KILDEE. Thank you.

Mr. Sawyer?

Mr. SAWYER. Thank you, Mr. Chairman. Thank you for your patience. I have had to be in and out of here in the course of the morning. If this panel lived up to its early promotion by the staff, it is at least the equal of the last full hearing that we had. It was remarkably useful and insightful, and I want to thank you for putting this together.

Chairman KILDEE. Well, this panel has lived up to the expectations of the staff. At this point I would like to thank the staff. It's great really, you know, for me, as a Member of Congress and as just a person, to be able to have the wisdom that is gathered before us as we have here this morning. It's very important as we embark on this that we know what we're doing, and the four of you certainly are providing us the knowledge we have to have so we will know what we're doing. Thank you, sir.

Mr. SAWYER. I would be pleased to yield to you, Mr. Chairman.

Chairman KILDEE. Thank you. I used a little prerogative there.

Mr. SAWYER. I look forward to reading the testimony, but I particularly want to thank you for the point of view that you have represented here today.

The business of assessment has taken on a momentum of its own, much of it positive and constructive and full of the potential for long-term benefit. However, I am concerned that some of it reflects a view that assessment for its own sake will create a market for improvement. To the degree that some States have, in fact, acted on that view, I think we have really created an obstacle, a hurdle, a perception that we have to overcome about the real benefits of responsible assessment.

I particularly want to thank those who have both provided the model that the National Council of Teachers of Mathematics has offered to so many others in so many fields as a pathway that we ought to follow and those who have actually undertaken the business of following it.

If what you measure is what you get, then we had better be damned sure what we are measuring. We, sure as we are sitting here, ought to provide the means to get there. This should be building a pathway, not a terminus. I think all of us who have struggled in the course of this last session to share that view ought to be grateful for the testimony that you bring here today.

Just in conclusion, let me again thank our Chairman, who in the Neighborhood Schools Improvement Act of last year really struggled long and hard to make sure that that view received its full hearing. I don't view that debate as over, by any means, and I just want to offer thanks for your leadership in that regard.

It is enormously important if there is to be a single, relatively low-cost set of reforms that we be aware of how it will affect the entire Nation. This is really at the heart of it. This will determine, as much as anything, whether we're able to really achieve reform and improvement, or whether we're simply out there measuring some small element of our failure.

Thank you, Mr. Chairman.

Chairman KILDEE. Thank you very much, Mr. Sawyer. I personally appreciate your involvement in this issue, your contributions to the issue. I appreciate the fact that you are a member of this committee. You have been very helpful to me and to the committee.

Are there any summary statements either one of you would like to make, collectively, individually? I really appreciate your testimony. I'm dedicated to improving education for every child in this country. When I came to Congress 17 years ago, I dreamed someday of chairing this very subcommittee, and a couple of years ago I did chair it. I feel I have an enormous responsibility in this position, and I don't want to fail. I want to really have made a difference for education in this country.

I have always summarized saying education is a local function, it is a State responsibility, and a very important Federal concern. It's a Federal concern for a variety of reasons. First of all, we live in a very mobile society. A person educated in Vermont may wind up in Michigan or California or vice-versa.

Now we are also competing in a global economy, and sometimes not competing as well as we could. Education is a very, very important part in that competition so we can have a quality of life in this country. It has to be a very, very important Federal concern. I take this responsibility quite seriously. It is people like yourself that help us address that responsibility in a very meaningful and knowledgeable way.

I very much appreciate your testimony here this morning. It has been excellent. I'm sure we will be calling upon you individually again, too, as we move along this path to make some differences in American education. Unless there are any further comments?

[No response.]

We will keep the record open for 2 additional weeks for any additional submissions you may have. With that, then this subcommittee will stand adjourned.

[Whereupon, at 11:50 a.m., the subcommittee was adjourned, subject to the call of the Chair.]

[Additional material submitted for the record follows:]

STATEMENT OF LYNNE C. WOOLSEY, A REPRESENTATIVE IN CONGRESS FROM THE STATE OF CALIFORNIA

Thank you, Mr. Chairman. I would like to welcome our panel of guests and to thank you for holding this important hearing today. Because performance on tests are tied so closely to opportunities in American society—opportunity for academic scholarships, for advancing through high school, for acceptance into selective college, and of course for development of self-confidence—we must be vigilant in guarding against possible unfairness.

It is essential that great care be taken to ensure that assessment mechanisms do not reflect bias based on sex, race, ethnicity, national origin, socio-economic status, or other similar factors. Particularly now, as new assessment systems are being developed and marketed to the tax-supported public school system, we must require that these systems be valid and non-discriminatory.

The often cited example of girls having higher grades than boys in high school but doing worse on the SAT really only scratches the surface of a very complex issue. I am currently looking at ways to ensure that gender-specific needs of boys and girls are addressed in this reauthorization.

I welcome our panel and am looking forward to hearing their testimony.

67-897 O - 93 (61)

ISBN 0-16-040958-6