DOCUMENT RESUME

ED 341 714 TM 017 899

AUTHOR Porter, Andrew C.
TITLE Assessing National Goals: Some Measurement Dilemmas.
PUB DATE 90
NOTE 24p.; In: "The Assessment of National Educational Goals: Proceedings of the 1990 ETS Invitational Conference" (New York City, NY, October 27, 1990). New York. Educational Testing Service, 1990.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Accountability; Criterion Referenced Tests; Educational Assessment; Educational Change; *Educational Objectives; Elementary Secondary Education; Evaluation Methods; *Evaluation Problems; *Measurement Techniques; National Programs; Norm Referenced Tests; Sampling; *Student Evaluation; Test Construction; *Testing Problems; Testing Programs
IDENTIFIERS Education Summit 1989 (NGA); *National Education Goals 1990

ABSTRACT

The measurement dilemmas involved in assessing the national educational goals established by the President and governors at the 1989 education summit are discussed. The first and most important choice is what to assess and whether to align assessment to the vision of curriculum reform or to the curriculum that students are actually experiencing. Another issue includes whether assessment should be aligned to what is to be assessed or to what we know how to test. Once assessments are constructed, sampling strategies must be carefully considered. Assessments that serve accountability purposes are more expensive than those that serve only descriptive purposes. The distinction between norm-referenced and criterion-referenced assessment is another issue that must be considered. The desire for international comparisons is an additional aspect that creates real problems in national assessment. Another dilemma is whether student performance is all that must be assessed, or must inputs and procedures be assessed as well? The 6 goals and 26 objectives defined by the President and the governors will only be useful if they are widely shared and widely recognized as achievable and worth the cost. A 21-item list of references is included. (SLD)

Assessing National Goals: Some Measurement Dilemmas

Andrew C. Porter

University of Wisconsin--Madison

In: The Assessment of National Educational Goals. Proceedings of the 1990 Educational Testing Service Invitational Conference.

Assessing National Goals: Some Measurement Dilemmas

ANDREW C. PORTER
University of Wisconsin — Madison

This paper was prepared for the 1990 Educational Testing Service Invitational Conference, "The Assessment of National Educational Goals," held on October 27, 1990 in New York City. The research reported in this paper was supported by the Center for Policy Research in Education, which is funded by a grant from the U.S. Department of Education, Office of Educational Research and Improvement (Grant No. OERI-G-0086-90011), and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the U.S. Department of Education, Office of Educational Research and Improvement, the institutional partners of the Center for Policy Research in Education, or the Wisconsin Center for Education Research.

"A national achievement test is no more the answer to improving educational quality than is a national curriculum" (Tyler & White, 1979, p. v). This was the first of several points strongly asserted at a three-day national conference on achievement testing convened in 1979 by then Secretary of Health, Education and Welfare, Joseph A. Califano. The contention was reiterated by participants in a follow-up, ten-day conference on research on testing sponsored by the National Institute of Education. I was one of 31 participants at the second of those two conferences. So was Greg Anrig, then Commissioner of Education for the state of Massachusetts and now, of course, president of Educational Testing Service.

Just ten years later, President George Bush announced: "The time has come for the first time in the United States history to establish clear national performance goals, goals that will make us internationally competitive" (U.S. Department of Education, 1990), p. 1). Our state governors agreed, adding that "national educational goals will be meaningless unless progress toward meeting them is measured accurately and adequately, and reported to the American people" (U.S. Department of Education, 1990), p. 13).

What contributed to this about-face in the thinking of our political leaders in such a short period of time? What are the implications for education practice? Especially for this conference, what are the implications for practices in education assessment?

One significant event, of course, was the publication in 1983 of *A Nation At Risk*, the highly influential product of then Secretary of Education Terrel Bell's National Commission on Excellence in Education. Although at the time the information about the performance of our education institutions was sketchy, Bell's commission concluded that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people" (National Commission on Excellence in Education, 1983, p. 5). The response has been considerable hand wringing, a fair amount of political jockeying for position by states and districts, and a few serious efforts at improvement on the part of schools and professional organizations. All of these efforts have gone under the general banner of raising standards in United States education. Concern for education improvement has been great and continuing. One reform has not been enough; there have been waves of reform. Still, in terms of student performance, not much has been accomplished. A report card on the nation's education productivity, released on the anniversary of the Charlottesville summit, supports the dismal conclusion that, while policies and some practices have changed, student achievement has not (Mullis, Owen, & Phillips, 1990).

The president and governors' goals for education, and the heightened interest in national assessment, are undoubtedly born out of frustration. Our efforts thus far have failed; something bold and different must be done. Certainly, the president and governors' national goals for education represent a bold move. The goals are ambitious and are to be achieved by the year 2000. Most important, the goals are stated in terms of performance. They virtually demand assessment.

Four of the six goals are stated in terms of what students know and know how to do. Some might argue that high school graduation is a performance goal, but I would not. Graduation from high school is an indication of participation, with only weak links to student accomplishment. The sixth goal, calling for a safe and orderly school environment, clearly falls outside the focus on student performance. The 26 objectives, offered to add specificity to the goals, are less performance oriented, reflecting instead same carry-over from past enthusiasms in education to specify procedures rather than accomplishments. Still, 11 of the 26 objectives, over 40 percent, specify standards for student outcomes.

While the goals and objectives do bear on performance, they are more inspirational than specific. They stop well short of specifying what exactly is to be achieved. Perhaps more importantly, they stop short of specifying who is responsible for achieving the goals, who is responsible for monitoring progress, and what the consequences will be for success or failure.

The measurement dilemmas created by the need to assess national goals are many — more than my imagination and expertise can anticipate, certainly more than can be dealt with here. The dilemmas do not, however, appear to hinge primarily upon technical problems, the bread and butter of psychometricians. The most important issues concern what to test, who to test, and what to do with the results. These are not uniquely measurement issues; they fall at the intersection of the educational, political, and measurement arenas:

- Should assessment be aligned to instruction, or should it be aligned to a vision of what we would like instruction to become?
- Should assessment serve the purpose of describing national progress toward achieving the goals, or is assessment needed to serve accountability purposes at local levels?
- Should assessment of national goals array student performance against criteria of what we think students should know and know how to do, or should the assessments report student progress against standards of past student achievement?
- Is it possible to have assessments that are valid and useful measures of what the United States is trying to accomplish and at the same time have assessments that provide international comparisons?
- Should limited resources available for assessment be focused exclusively upon performance, or is it necessary to assess inputs and procedures of education as well?
- Most problematic of all, whose goals are these?

Goals are meant to inspire, to motivate, to focus energy toward a desired end. Goals are uplifting. Measurement is empirical, objective, and grounded in reality. Dilemmas are created when only disagreeable alternatives for action exist. My assignment is to identify and discuss the measurement dilemmas involved in assessing national goals. There are difficult and perplexing choices that must be faced, but most appear to allow for the possibility of a promising resolution.

Alignment

The first and most important choice concerns what to assess. Do we assess what we aspire to or what we have now? One of the six national goals focuses squarely on student achievement in grades K-12: "By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency over challenging subject matter, including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well so that they may be prepared for responsible citizenship, further learning, and productive employment in a modern economy" (U.S. Department of Education, 1990, p. 5). The key terms in this goal are *challenging, learn to use their minds,* and *all students*. These terms leave open the door to assessment aligned to the vision of an unprecedentedly ambitious curriculum reform, which is just beginning in our nation and which can be characterized as hard content for all students (Porter, Archbald, & Tyree, 1990).

There currently exists an enormous gulf between the vision of curriculum reform and current practice. One way to see the lofty nature of our newest curriculum reform is to place it in historical context. These may be our first national goals, but we certainly have had curriculum reforms before.

In the late 1950s and early 1960s, we were in a race with the Russians to see who could be first on the moon. More and better mathematicians, scientists, and engineers were needed. Leaders from the hard sciences joined educators in upgrading the curriculum, primarily the high school mathematics and science curriculum. For comparative purposes, this late 1950s reform can be labelled hard content for the academically elite. In the 1960s, another curriculum reform began. That reform grew out of the Great Society and concerns for equality of educational opportunity. The goal was mastery of basic skills. Every student was to learn how to read and compute, a guarantee of at least easy content for all students. Enormous energies were poured into this reform; the residue from these efforts can be found in virtually every classroom in the nation.

Today's goal of hard content for all students combines the most demanding aspects of each of the two previous curriculum reforms. In the words of the National Governors Association, what is required is a radically new view of knowledge and learning throughout our society. We must abandon traditional, but no longer useful, distinctions between thinking and doing, knowledge and action, the academic and the voca-

tional. We must abandon the view that only a small portion of our population must be well educated (National Governors Association, 1990, p. 7).

Of the various school subjects, mathematics is far out in front in making clear what teachers are to teach and students are to learn for the reform to be successful. In 1989, three reports on a new mathematics curriculum appeared (American Association for the Advancement of Science, 1989; National Council of Teachers of Mathematics (NCTM), 1989; National Research Council, 1989). Each agreed with the others in calling for what I am characterizing as hard content for all students. The National Council of Teachers of Mathematics' *Standards* articulate five general goals for all students: (1) that they learn to value mathematics; (2) that they become confident in their ability to do mathematics; (3) that they become mathematical problem solvers; (4) that they learn to communicate mathematically; and (5) that they learn to reason mathematically. "The opportunity for all students to experience these components of mathematical training is at the heart of our vision of a quality mathematics program" (p. 5). But mathematics is not yet very far along in accomplishing the goals that it has set for itself.

Careful studies of curriculum practices in schools are rare; unfortunately, the empirical procedures to provide documentation of the content of instruction (i.e., the enacted curriculum) are time consuming and expensive (Freeman & Porter, 1989). My colleagues and I have done some of this type of work for elementary school mathematics (Porter, Floden, Freeman, Schmidt, & Schwille, 1988). The results of our investigations show practice to be almost polar opposite from what NCTM's standards envision (Porter, 1989a). Between 70 and 75 percent of mathematics instruction at the elementary school level is spent on skills: computational problems involving addition, subtraction, multiplication, and division. Of the time not spent on skill development, about half is devoted to conceptual understanding and the other half to problem solving. But even the problem solving that is taught is not the type NCTM would want. Highly structured story problems are used. Most frequently the problems are presented in a list where each is like the next. All call for the same type of solution; for example, adding two single-digit numbers to get the right answer. The only consensus among elementary school teachers on which mathematics topics to emphasize is computational skills: multiple-digit multiplication, long division, number facts, and subtraction with borrowing. In fourth- and fifth-grade classrooms — where skill development involves work on percentages,

decimals, and fractions — solving problems involving percentages, decimals, and fractions receives scant attention.

Mathematics is the best example of the alignment dilemma. The vision for reform is clear and clearly different from current practice. But other subjects are joining the curriculum reform, and what information is available about practice shows the same heavy emphasis on learning facts and skills and away from conceptual understanding and application. For example, in 1989 three curriculum reports also appeared in the area of social studies ("Feature," 1989; Gagnon & Bradley, 1989; National Commission on Social Studies in the Schools, 1989). In social studies, there is less agreement about the need for hard content for all students and more arguing about how much emphasis should be given to history and geography. Still, the timing and the number of reports suggest that curriculum reform is very much an issue for social studies, as it is for mathematics. Newmann and his colleagues (in press), in their studies of high school social studies, find a facts-dominated curriculum that, in its desire to cover large amounts of content, results in a shallow curriculum, providing little opportunity for students to develop in-depth understanding, to integrate ideas and information, and to engage in learning activities that allow them to acquire knowledge having value for their lives.

The dilemma is whether to align assessment to the vision of the curriculum reform or to the curriculum that students are experiencing. Either choice has its problems. If we assess students against the reform, their performance will almost certainly be terrible, at least initially. Not only is it unlikely that students will learn what they have not been taught, but conceptual understanding and application are simply more difficult to master than are facts and skills. On the other hand, if we assess students against the curriculum they experience, results will be better but the information will provide no real knowledge of progress toward the goal. Worse, an assessment aligned to current practice will become an anchor to that practice.

There is a second part to the alignment dilemma. Will assessment be aligned to what we want to assess, or will it instead be aligned to what we know how to test for? You may judge this second part trivial or foolish, but in practice it is not. I asked a number of my colleagues in different subject matter areas for examples of what they felt were good assessment practices. Being subject matter specialists in mathematics, science, and social studies, they not surprisingly focused on the question of whether or not worthwhile content was being assessed. I had in mind

that I would share those examples with you today. I thought they would serve to illustrate what experts have in mind when they say that students should be able to reason, apply knowledge, and solve novel problems. By asking for exercises appropriate for students in grades 6 through 8, I had further hoped to illustrate that what is being asked of students is not easy, perhaps not even easy for such a distinguished group as yourselves.

The assessment exercises my colleagues provided illustrated yet two other points. They take time and space to present and require elaborate scoring procedures that are not easily described. They are also not in ready supply.

Fred Newmann's assessment exercise on understanding and using Constitutional issues took four double-spaced pages to present; another three and a half double-spaced pages were needed to describe the 5-point scoring procedure. A student is suspected of violating a school rule about smoking and actions are taken. The exercise calls for respondents to draw on principles the courts have used in making decisions about the Constitutionality of student searches to decide whether the student's Constitutional rights were violated and to write a persuasive essay of their position. In tryouts of the item with over 1,000 students, only 12 percent were able to do well on the task.

Tom Romberg provided mathematics assessment exercises like the following: A goat is tied by a rope to the corner of a barn. The barn is 20 feet long on each side, and the rope is 10 feet long (there is a picture of the goat tied to the barn). This information is followed by a series of questions from easy ("Which is longer: The side of the barn or the rope?") to harder ("Calculate the area of the grass the goat can eat") to harder still ("If the rope were extended to the length of 25 feet, describe *how* you would find the area of the grass the goat can eat by drawing a sketch and describing the shapes"). Roughly 25 percent of a general sample of eighth graders were able to get the last portion of this task correct.

Science exercises for assessing student understanding and ability to apply their understanding were similar in requiring students to produce an elaborated response that then required judgment in scoring.

One characteristic of these assessment exercises is that they require considerable response time. An assessment comprised primarily of such exercises can provide only a limited number of opportunities for a student to demonstrate what he or she knows and knows how to do. Psychometricians who worry about the traditional criterion for assessment of reliability are concerned that such limited response opportunities provide unreliable information about individual students. Multiple-

choice tests, where each item takes less than a minute to complete, allow students many independent opportunities to respond. Misunderstandings and isolated knowledge gaps that can create errors of measurement are partially overcome through the weight of answering many items; content domains are better sampled. Performance-oriented exercises, if they are to provide reliable information, must be certain to demand only essential knowledge, have absolutely clear instructions, and use scoring techniques that give partial credit for a response that shows partial understanding or ability to apply.

Reliability is an important criterion for assessment, but reliability is only important in that it is enabling for validity. Advocates for performance assessment are critical of multiple-choice testing because they believe that knowledge utilization is not validly assessed unless students are asked actually to use knowledge and do it in circumstances similar to real life. They further argue that performance assessments have greater face validity to the type of instructional activities that the curriculum reform requires and so are better influences on instructional practice. The dilemma is whether to have a multiple-choice assessment that provides reliable information that may not be valid or to have performance assessment that appears valid but may not be reliable. Psychometricians are the biggest critics of performance assessment, and subject matter specialists are the biggest critics of multiple-choice tests.

The limited number of exercises possible in a performance assessment may not be a problem for content validity. If each exercise on a national performance assessment covers essential content, teachers will need to agree on that essential content. If, in addition, a large portion of this essential content involves understanding and application, outcomes that require considerable instructional time, then instruction will need to become much more focused on a few important areas of content. While these changes are consistent with the vision of the current curriculum reform, they will not be easy to accomplish.

Paradoxically, an assessment may exacerbate the problem of achieving focus in instruction. Being clear on what is important and adding to this clarity through assessment will, by omission, also identify what is less important. The goal of breadth of coverage that characterizes instruction today will be replaced by a goal of depth of coverage. Depth will require eliminating from instruction a great deal of what is currently taught (or at least covered). And teachers will not be comfortable. In a study of elementary school teachers, my colleagues and I asked them to describe what students would know and know how to do if their instruc-

tion were 100 percent effective (Schmidt, Porter, Floden, & Freeman, 1987). Interviews took approximately two hours. At the end we asked teachers whether there were topics they would add to instruction if more time were allotted. Virtually all teachers could identify several such topics. We then asked teachers what topics that they currently teach they would eliminate given less time. Not a single teacher knew where to begin; all felt uncomfortable with the task. In another study (Floden, Porter, Schmidt, Freeman, & Schwille, 1981), teachers were asked to react to a variety of circumstances and say whether or not those circumstances would persuade them to add content to what they already were teaching in mathematics. As policies and advice mounted for the teachers to add topics, they reported themselves as virtually certain to do so. But as policies and advice mounted for teachers to delete topics from their instruction, teachers remained ambivalent to do so.

Indicators or Accountability

The topic of today's conference is assessing national goals. But what purposes are to be served by this assessment? The first purpose that comes to mind is descriptive: Is progress toward the goals being made? You might say, what other purpose could assessment serve? The answer is accountability. Assessment tied to accountability can be an important part of the strategies to reach the goals.

First consider what is necessary for assessment to serve descriptive purposes, where we stand relative to the goals we have set. Obviously, the beginning point is to come to agreement about the kinds of performances the goals imply. This was the heart of the alignment dilemma just discussed. Once those agreements have been reached and assessment exercises constructed, sampling strategies must be considered. A census approach is not necessary.

First, decisions need to be made about the size and strata of a sample of students. The achievement goal specifies grades four, eight, and twelve as the stratification. A second sampling decision is how often to assess. Assessment each year might not be necessary, but with the year 2000 as the deadline for attaining the goals, assessments once every four years are not frequent enough. Obviously, the schedule for assessment wouldn't need to be the same for every subject area, but it could be. Third, a longitudinal sampling design is not necessary, but sectional samples must be representative (i.e., probability samples). The goals are stated in

ways that promote interest in knowing how, for example, fourth-grade cohorts improve over time. There is, however, a small problem that cross-sectional comparisons present, especially at the twelfth grade. If the goal of increasing the percentage of students who graduate from high school is accomplished, that will result in a shift in the demographics of twelfth-grade students over time. Comparing a representative sample of twelfth-grade students today with a representative sample of twelfth-grade students in the year 2000 would confound demographic shifts with instructional changes. A fourth sampling decision concerns the assessment exercises themselves. Different assessment exercises could be given to different samples of students, even within the same cohort, so long as the samples are randomly equivalent. Assessments would be based on a great many exercises, even if the exercises are of the time-consuming type reviewed previously in this paper.

My key point is that assessment for descriptive purposes does not require reliable and valid information at the individual student level. It does not even require valid and reliable information at the level of a particular school, district, or state. Sampling procedures comparable to those used in our current National Assessment of Educational Progress (NAEP) would do, although there would be arguments from subject matter experts about whether or not NAEP assessments are appropriately aligned with the vision of the current curriculum reforms.

Assessment that serves descriptive purposes only tells us whether or not we are making progress toward achieving the goals. It does not influence that progress. Assessment that serves accountability purposes, however, can be a lead policy instrument in achieving the goals. This position has been adopted by such groups as President Bush's Education Policy Advisory Committee and the National Center on Education and the Economy. Both groups have recently gone on record as saying that current national, state, and local tests are inadequate to spur improvements in education. They believe a new national examination that all students take is necessary.

The idea is simple enough. If a test aligned to the national goals is given to all students, and if performance holds consequences for students, teachers, schools, districts, and states, then preparation for doing well will become very important.

To have maximum potential for influence on practice, the idea of accountability needs some elaboration. Not only must there be rewards and sanctions attached to performance on the national assessments, but teachers, students, parents, and the general public must be convinced

that what is being assessed is appropriate and worthwhile. This gives the assessments both power and authority (Schwille, Porter, Belli, Floden, Freeman, Knappen, Kuhs, & Schmidt, 1983). Over time, other education policies and instructional materials would come into alignment with the assessments. State and district curriculum frameworks and objectives would be aligned with the national assessments. State and local testing would be aligned with the national assessments. The market would drive textbook publishers to produce materials consistent with the assessments.

Unfortunately, assessments that serve accountability purposes are much more expensive than assessments that serve only descriptive purposes. First, all students must be assessed. This census approach is expensive in its own right. It also eliminates the possibility of different students responding to different assessment exercises, which in turn narrows the sample of assessment exercises. Second, since performance on the assessments would be of great importance, steps would need to be taken to standardize administration and scoring procedures. Rules would need to be enforced for students, if any, to be excused from assessment. Cheating must be prevented. New assessment exercises would need to be used at each time of assessment. Not only would it be impossible to guarantee security of a high stakes test over time, but it would be counterproductive. The point of assessments to support accountability is to move practice in desired directions. After each assessment, the items should be widely circulated so that everyone knows in specific detail the kinds of knowledge and skills students are to possess.

Criterion-Referenced or Norm-Referenced Assessment

The controversy between assessment that is norm referenced versus assessment that is criterion referenced is a classic. Certainly the bulk of achievement testing in United States schools today emphasizes the narrative reporting of results. Parents and students want feedback in terms of percentile rank. Everybody wants to be above-average students, schools, even whole states.

For the past 20 years or so there has been a persistent minority of educators who have emphasized the importance of giving feedback on performance against some meaningful criterion. The notion is to report results in terms of what students know and know how to do. Probably, if there were greater agreement on what constitutes worthwhile knowl-

edge, the criterion-referenced movement would be further along. Another stumbling block for increasing popularity of criterion-referenced feedback is the difficulty of finding parsimonious ways to provide such information. A percentile rank is a single number. Thus far, most criterion-referenced feedback comes in the form of a long list of concepts, skills, and applications, with degrees of mastery reported for each. Such detailed feedback makes answers to questions of "how'd you do" complicated and conditional.

The president and governors' goals appear to straddle the fence on the criterion-referenced/norm-referenced issue. The goal of demonstrating competence over challenging subject matter is obviously calling for criterion-referenced assessment. Somehow, agreement will have to be reached about the meaning of *competency* and the meaning of *challenging*. In contrast, the first objective under that goal takes a norm-referenced approach: "Students will increase significantly in every quartile, and the distribution of minority students in each level will more closely reflect the student population as a whole" (U.S. Department of Education, 1990, p. 5). For this objective, assessment must provide information on gains in performance over cohorts and, within-cohorts, information on contrasts between minority students and the whole population. The objective could be reached without the goal being achieved. Minority students could close the gap in achievement between themselves and majority students without majority students reaching the point of demonstrated competence over challenging subject matter. In fact, for the past 20 years, this may have been happening. Gains in academic achievement for Black students have typically been greater than gains in achievement for majority students (Mullis, Owen, & Phillips, 1990).

The president and governors have taken the right approach in calling for both criterion-referenced and norm-referenced assessments. Curriculum reform is needed; all students must be given the opportunity to learn and learn how to apply important concepts in all subjects. Only criterion-referenced assessment can tell us about progress toward these new standards that are qualitatively different from current practice. The goals also specify that all students have the right to this worthwhile content.

Practice is being challenged in two important ways. First, the purposes of schooling and instruction must be fundamentally changed to place a greater emphasis on conceptual understanding and application and away from rote memorization of facts and drill and practice on skills. But this shift, which is needed for all students, is an especially great shift for

students from low-income families, including large percentages of minorities. In short, schools must not only change what it is they try to accomplish with the academically elite, but they must also find ways to decrease or eliminate tracking and to provide usable knowledge to students who at present receive very little, if any, such knowledge. This change will require great shifts in thinking about student expectations. Society and schools would no longer be allowed to write off the 20 percent or so of students who have the most difficulty in school. If we are to have assessments useful for monitoring whether progress is being made with the academically elite, clearly norm-referenced results will not do. These are the students and schools already at the top. Their percentile ranks cannot go higher. But the equity objective and the progress objective require statistical contrasts based on norms. Is average performance of minority students coming closer to average performance of majority students, and is average performance increasing over cohorts?

At least in theory, a single assessment can serve both norm-referenced and criterion-referenced purposes. A criterion-referenced assessment must be constructed. This is not an easy task. It requires careful mapping of knowledge and knowledge-utilization domains and careful sampling of assessment exercises from those domains in ways that have content validity and that provide stable estimates of performance for each of the various domains deemed important. Once an assessment with these properties is available, norms are easily achieved through administration to appropriate samples. If a census approach is taken, such as is required for accountability purposes, then norms can provide statistical contrasts of whatever type might be valued. Contrasts can be formed between various minority groups; contrasts can be formed on gender; contrasts can be formed on socioeconomic level; and, of course, contrasts can be defined across cohorts. If the assessment is to serve only descriptive purposes, then the contrasts of interest must be built into the sampling design a priori. Clearly, contrasts are needed for ethnicity, gender, and socioeconomic status. Some might argue for geographic contrasts, as well.

There is one conceptual problem worth singling out for special consideration when norm-referenced assessment is done. Who is to be assessed, and who is to be exempt from assessment? Earlier I pointed out that achievement contrasts over cohorts could confound shifts in demographics with shifts in instructional effectiveness. But within a cohort, there are similar considerations. In schools with high absentee rates, must assessment ensure a representative sample of all in attendance on

the day of the assessment as well as all who are not? And what about students with special needs? Are the norms to include students with various types of disabilities? Detailed answers to this question are required, and the assessment practices must remain constant over time.

International Comparisons

The president and governors' goals don't stop with implications for national assessment. They call for international assessment, as well. In the goal statement document (U.S. Department of Education, 1990), four of the eight introductory paragraphs have one or more references to international comparisons. This enthusiasm for international competitiveness is not carried through systematically in the goals, but it is not totally absent either. Our nation is to be first in the world in mathematics and science achievement by the year 2000.

This desire for international comparisons creates some very real problems. Our history with international assessment is checkered.

The primary instrument for international assessment has been the International Association for the Evaluation of Educational Achievement (IEA). Incorporated in Belgium and operating out of Sweden, the IEA has been forced to operate on a shoestring budget for cross-national activities and at the mercy of each separate country's government for within-country work. The results have been interesting but flawed. In mathematics, IEA assessments have occurred only in 1964 and 1982, nearly a 20-year period of time between assessments. In each case, serious criticisms have been made of the sampling procedures, not only in terms of the representativeness of the sample taken in the United States, but especially in differences among countries in sampling procedures followed.

Decisions about what and who to test are driven by the need for international consensus. Procedures that are fair in the sense that they achieve international consensus are not necessarily procedures that are valid for assessing the accomplishments of a particular country. If, for example, one country places a great deal of emphasis on concepts and applications in the area of probability and statistics, but other countries do not, that domain is unlikely to find its way into an IEA assessment. If one country emphasizes that all students are to master challenging subject matter, while most other countries are heavily tracked with only a select few students *expected* to master challenging subject matter, then

international consensus might result in a sampling plan that compares all students in one country to only academically elite in another.

The dilemma is this: Must the United States have one assessment procedure for its own internal use and another assessment procedure that it uses cooperatively with other countries for purposes of international comparison? So far, dual assessment has been true. If this practice of separate assessments continues, can the tests be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can only be done when tests measure the same thing. The primary motive for having a separate test for international comparisons has been due to disagreements over what should be tested.

What we need are international assessments that serve U.S. purposes. Our intended curriculum must be validly represented on the international assessment. Sampling procedures must provide a representative sample of all students of a given age or, alternatively, the sampling procedures must use exclusion rules that are identical across countries. Issues of response rate, representation of students in private schools, and the like must all be resolved in ways that eliminate confounding in international comparisons. Since the U.S. goal is hard content for all students, international assessments should provide descriptions of within-country variability in accomplishment, for example, through contrasts on gender, socioeconomic status, and race.

Student motivation to do well on the assessment exercises must also be controlled. If assessments are done for descriptive purposes only, with no implications for accountability, then mechanisms must be put in place to control for differences in student motivation. These controls seem especially important in international assessment where different cultures could lead to quite different response strategies. I don't have a good solution to offer, but one option may illustrate the point. Students could be paid for participation, with the amount dependent on their performance.

There is the question of the ages or grades that should be used to define international comparisons. Our goals say grades four, eight, and twelve, but grade levels do not have the same meanings across countries. Perhaps international assessment should use age instead of grade, but even using age presents a problem. Deficits or advantages in achievement of one country over another can change over age cohorts. The United States might fare relatively poorly in grades four, eight, and

twelve, but look competitive on international comparisons of achievement for persons of age 21. We have a relatively large percentage of students who participate in higher education, and our system of higher education is generally held to be among the best in the world.

It is easier to articulate the properties that the U.S. would like an international assessment to have than it is to put them in place. Each country participating in an assessment will have its own demands, and there are sure to be serious uniquenesses across countries. Fortunately, a panel of the National Academy of Sciences was recently created and placed in oversight of U.S. participation in international assessments. This panel provides considerable assurance that from now on when we become involved in international assessments, we will do so with full understanding of what will and will not be provided by way of useful information.

The National Academy's oversight may also help to solve another problem that has plagued U.S. participation in international assessments. For past international assessments, inadequate early funding has resulted in compromises to the quality of conceptual work. This has in turn seriously diminished the utility of entire assessments. Hopefully, the National Academy's approval will help secure timely and adequate funding.

Assessing Inputs and Procedures

A fifth dilemma is whether we can get away with assessing only student performances, or must we assess inputs and procedures as well? Here, again, the goals and objectives provide a partial answer. Only six of the 21 objectives are clearly about outcomes. The majority of the objectives are about inputs and procedures:

- Students will have access to high quality and developmentally appropriate preschool programs
- Students will be involved in activities that promote and demonstrate good citizenship
- The number of teachers with substantive background in mathematics and science will increase by 50 percent
- The number of quality programs . . . to serve more effectively the needs of the growing number of part-time and mid-career students will increase substantially

- Every school will implement a firm and fair policy on use, possession, and distribution of drugs and alcohol

If these objectives are to have the same stature as those stated in terms of performance, then indicators of the nation's position relative to them will need to be put in place.

More generally, there are at least three good reasons for creating a system of indicators reflecting inputs and procedures as well as indicators of performance (Porter, in press). First, indicators of procedures and inputs are needed for descriptive purposes. When investments are made in education, taxpayers have the right to know what their investments purchase. Information must be made available to describe the probability of a particular type of individual receiving the opportunities intended. Questions about the amount, quality, and distribution of benefits must be answered. Second, reforms call for changes in current practice. They seek to alter the types of inputs and procedures available. The fact that the bulk of the 26 objectives elaborating on the goals concern inputs and procedures, not performance, makes this point. Monitoring reform requires a system of indicators that goes beyond looking at outcomes to assess whether or not intended changes in inputs and procedures are being made. Third, indicators of inputs and procedures are needed for diagnostic purposes. As our nation strives to reach the goals, there will almost certainly be instances of success and instances of failure. Explanations for these differences in success may lie in differences in the types of students served, differences in support from business and industry, differences in the nature of education programs provided. Without diagnostic information that only indicators of inputs and procedures can provide, appropriate corrective measures will be difficult to determine.

The challenge to assessment is enormous. A focus on outcomes alone would be a challenging enough task, but the number of potentially relevant inputs and procedures to assess is virtually infinite. Without a strong theoretical model to guide decisions about what to monitor and how, the entire effort will surely collapse under its own weight. The dilemma is that ignoring indicators for inputs and procedures seems an equally dangerous course of action.

National Goals

The president of the United States and our state governors have given us six goals and 26 objectives, but whose goals are these? Are they really

national goals, goals we all believe are worth striving for and to which we are willing to invest personally? I am afraid the answer is no, at least not yet.

Senate Bill S2034, introduced by Senator Bingaman, calls for the creation of a national council on educational goals that would, within its first year, recommend a "comprehensive set of national educational goals to be achieved before the year 2000." The council is to "consider the goals already set forth or recommended by the President, the National Governors Association, and other governmental and nongovernmental organizations." But clearly the bill allows for the possibility that this new council's goals would be different from the six we are considering now.

Nor is there agreement about who would control assessment of the goals. Both Senate Bill S2034 and House Bill HR5115 call for a panel appointed by the president of the United States and comprised of individuals with "experience in, knowledge of, and commitment to education and educational excellence." In contrast, the National Governors Association is creating a 14-member panel of governors and federal officials.

While the Congress and the governors argue over panels, other organizations are moving directly to considerations of specific assessments of student achievement. As mentioned earlier, the president's Educational Policy Advisory Committee believes a new test is needed. Independently, the National Center for Education and the Economy has concluded the same. Both organizations appear committed to moving forward on their ideas. In contrast, Secretary of Education Cavazos speaks against having a national test, although perhaps he means no federal test. There is the National Assessment of Educational Progress, with a long history of providing information about student achievement at a national level. But for various reasons, NAEP is not seen as sufficient. States, too, are hard at work on assessment. Connecticut educators are in the process of putting together what they hope will be state-of-the-art performance assessments in a variety of subject matter areas. One can't help but wonder whether real national goals and assessment of those goals will emerge out of these disparate efforts and political bickerings.

A Gallup Poll sponsored by Phi Delta Kappa shows that the general public is not yet on board either. Respondents gave high priority to each of the six goals but, at the same time, expressed great skepticism about the goals being reached ("National Goals," August 23, 1990).

The public and the politicians, important as they may be, are not the most important groups to worry about when answering the question, whose goals are these? What do teachers and education administrators

think? Do they endorse the president and governors' goals as their own? Or is this another instance of what Larry Cuban labels "remote control teaching"? (Cuban, 1984). Teachers and administrators had little to no participation in the development of the president and governors' goals, an approach that is clearly at odds with current thinking about the need to empower educators and deregulate schools. Nor is it consistent with effective research that points to the importance of clear and shared goals at the school level.

My colleagues and I in the Center for Policy Research in Education (funded by the Office of Educational Research and Improvement) have recent findings that are suggestive of what might happen through efforts to reach national goals that teachers haven't adopted (Porter, 1989b). We are conducting investigations of the influence of state and district curriculum policies on classroom practices in social studies and mathematics. Our work is in states with curriculum frameworks calling for substantial change in classroom practice, in particular shifting the curriculum away from emphasis on facts and skills and toward emphasis on conceptual understanding and application. In each state, districts are selected that contrast in the extent to which they support and extend state initiatives in curriculum reform. Schools are carefully matched on student body composition. Our preliminary findings are as follows. Districts that reiterate and reinforce state curriculum frameworks through their own testing practices, textbook adoptions, and staff development efforts are successful in moving classroom practice toward a greater emphasis on conceptual understanding and problem solving. This result is just as might be hoped. At the same time, however, we find teachers in the curriculum-reform-active districts less willing to accept responsibility for student success or failure, less willing to hold students to high standards of academic achievement, and less pleased with their professional role in their schools. One interpretation is that teachers are willing to change the content they emphasize in their instruction, even if they are not convinced that such shifts in practice are appropriate for students. The down side is that they resent being told what students can and should learn, and their commitment is negatively affected. It may be that as teachers grow familiar with the new curriculum and gain confidence, these results will change in a positive direction. Hopefully, this will be the case

For goals to be effective, they must possess at least two characteristics. First, they must be agreed to and held as important; and second, they must be seen as within reach if great effort is made. At the moment,

current goals appear to lack sufficient amounts of either characteristic. If teachers are not somehow brought on board, the net effect could be even worse than what we have now.

Dilemma Management

The problems and dilemmas presented by the president and governors' six goals and 26 objectives are many. They are political, educational, and measurement in character. I have focused on the measurement aspects, but my analysis has repeatedly spilled over into the educational and political arenas. Clear national performance goals for education in the United States hold potential for doing a great deal of good, but that potential is dependent on a number of factors not yet in place.

My greatest concern is with the goals and objectives themselves. I won't argue with their focus, though I could quibble here and there. My real concern is with their development. To be useful, the goals and objectives must be widely shared and widely seen as achievable. For the goals to be useful, a great deal of effort must go into seeing that this happens, and happens quickly. Procedures to assess progress toward the goals could help.

First, the goals should be interpreted as consistent with the current curriculum reform, which I have labelled as hard content for all students. This will give the goals credibility through endorsement by professional organizations and a good many individual experts. National assessment procedures that are criterion referenced to the curriculum-reform vision could make this happen.

National assessment procedures should also possess the following characteristics. They should be norm referenced so that they report progress over time and within time on contrasts among types of students. The assessments should be comprehensive across subject matter areas, grades, and levels of student accomplishment. Thus, influences on practice will be applied evenly in all areas of need. The assessments must have the weight necessary to influence practice. Among other things, the assessment should serve accountability purposes. The assessments should be standardized and unbiased so that they provide valid contrasts over time and across subgroups. The assessments should be repeated frequently enough that data are current and progress toward goals can be monitored. Ideally, the assessments should allow for international comparisons. For diagnostic purposes, the assessments should be coupled with indicators of inputs and educational procedures.

Assessments .hat satisfy this long list of characteristics will be expensive. Their development and implementation will require money, expertise, and time. Some may argue that the costs are more than we can afford. This presents the greatest measurement dilemma of all. Can we afford to take shortcuts now, or will they compromise our future?

References

American Association for the Advancement of Science. (1989). *Science for all Americans* (A Project 2061 report on Literacy Goals in Science, Mathematics, and Technology). Washington, DC: Author.

Cuban, L. (1984, November). School reform by remote control: S.B. 813 in California. *Phi Delta Kappan*, 213-215.

Feature: Alternative scopes and sequences. (1989). *Social Education*, 53 (6), 375-403.

Floden, R. E., Porter, A. C., Schmidt, W. H., Freeman, D. J., & Schwille, J. R. (1981). Responses to curriculum pressures: A policy capturing study of teacher decisions about content. *Journal of Educational Psychology*, 73, 129-141.

Freeman, D. T., & Porter, A. C. (1989). Do textbooks dictate the content of mathematics instruction in elementary schools? *American Educational Research Journal*, 26 (3), 403-421.

Gagnon, P., & the Bradley Commission on History in Schools (Eds.). (1989). *Historical literacy: The case for history in American education*. New York: Macmillan.

Mullis, I. V.S., Owen, E. H., & Phillips, G. W. (1990). *America's challenge: Accelerating academic achievement. A summary of findings from 20 years of NAEP*. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Commission on Social Studies in the Schools. (1989). *Charting a course: Social studies for the 21st century*. Washington, DC: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National goals important but unattainable, poll says. (1990, August 23). *Education Daily*, pp. 1, 3, 4.

National Governors Association. (1990, February 25). *National education goals*. Washington, DC: Author.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

Newmann, F. M. (in press). Linking restructuring to authentic student achievement. *Phi Delta Kappan*.

Porter, A. C. (1989a). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher, 18* (5), 9-15.

Porter, A. C. (1989b). *Impact of the new curriculum policies on course content teachers' sense of teaching effectiveness, and teacher morale*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Porter, A. C., Archbald, D. A., & Tyree, A. K., Jr. (in press). Reforming the curriculum: Will empowerment policies replace control? In S. Fuhrman (Ed.), *The politics of curriculum and testing*. London: Taylor & Francis Ltd.

Porter, A., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 96-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, W. H., Porter, A. C., Floden, R. E., Freeman, D. T., & Schwille, J. R. (1987). Four patterns of teacher content decision making. *Journal of Curriculum Studies, 19* (5), 439-455.

Schwille, J. R., Porter, A. C., Belli, G., Floden, R. E., Freeman, D. J., Knappen, L. B., Kuhs, T. M., & Schmidt, W. H. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. S. Shulman & G. Sykes (Eds.), *Handbook on teaching and policy*. New York: Longman.

Tyler, R. W., & White, S. H. (Eds.). (1979). *Testing teaching and learning: Report of a Conference on Research on Testing*. August 17-19 1979. Washington, DC: National Institute of Education.

U.S. Department of Education. (July 1990). *National goals for education*. Washington, DC: Author.