MANDATED TESTS: REFORM OR QUALITY INDICATOR?

CSE Technical Report 283

Eva L. Baker

UCLA Center for Research on Evaluation,
Standards, and Student Testing

April, 1988

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This paper will as a chapter in *Test Policy and Performance: Education, Language, and Culture* (B.R. Gifford, editor), to be published in 1989 by Kluwer Nyhoff.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction

The thesis of this paper is that achievement tests have changed their primary function from serving as indicators of educational accomplishments. They have, in addition, become instruments of educational policy and have come to be regarded as effective means to alter educational achievement and productivity. I will explore this assertion by using examples of research and development from state and national testing activities. I will also consider how these alternative functions affect system behavior, legitimate policy inferences, technical requirements of tests, and ultimately our understanding of educational quality.

Educational testing has long been with us, but recently has demanded new levels of attention as states and the Federal government have increased their investment in and attention to the problem of measuring educational achievement. The function of tests used to be straightforward: to find out what some person knew or could do. Tests of this sort were given in schools to all of us. These tests were most often idiosyncratic in their design, made up, as they were, by one or more teachers. Such tests might have appeared to be very formal or even frightening to students, but their creation was informal, that is, the content they included and the standards used for their scoring were the decisions of teachers, people with close understanding of classroom instruction. Even bureaucratically entrenched and successful tests, such as the New York State Regents Examinations, were reasonably flexible, in that they were developed by teams of teachers and test writers and changed annually to reflect transitions and modifications in the curriculum.

The intellectual roots of the standardized test enterprise have been well documented (Coleman, Cronbach & Suppes, 1969). Driven by pressing national needs to make personnel decisions, during years of war, the mechanics of test design, administration, and analysis became more refined, more esoteric, and in turn, more credible to a technologically-oriented society.

For example, test based selection for admission to higher education, using the Scholastic Aptitude Examination (SAT), has been a regular part of students' experience for about the last forty years. Yet, the SAT became regarded as an end rather than as a tool in many conversations about education. When the scores on the SAT declined (See Harnischfeger & Wiley, 1975, and Wirtz et. al., 1977 for analyses), inferences were drawn about school effectiveness, even though that test was never designed to measure the goals of educational programs. The concurrent rise in the minimum competency test movement, where students needed to pass examinations on certain basic skills for graduation from high school or even for grade to grade promotion, further moved testing into the mainstream of American policy options. (Jaeger & Tittle, 1980). In minimum competency programs, in the 1970s in particular, the existence of the test triggered a variety of policies that dramatically changed the rules (Lazarus, 1981). Appropriate test performance became a goal, by almost any means necessary, and at first, at almost any cost as well, including long-term effects on children (Cohen & Haney, 1980; Kennedy, 1980). The most recent phase of testing involves a conceptual extension of the idea of minimum competency to content and skills purportedly demonstrating higher levels of subject matter competence (see the positions of Hirsch, 1987, and Finn & Ravitch, 1987), from minimal to optimal—but more about this topic later.

The Policy Attractiveness of Testing

Even though the testing of students and teachers remains controversial and occasionally the subject of litigation and judicial review, many policy makers in school districts, statehouses, and at the Federal level continue to see higher standards as the cornerstone of educational reform efforts and tests as their operational implementation.

Why? Even as the phrase, "There are no quick fixes," grows more popular in our rhetoric, we still continue our search for the chimera. Testing is assumed to be a relatively expedient remedy and magnetically continues to attract policy proponents, advocates who are in turn supported by a well-connected commercial testing industry.

What is it that testing seems to offer? I believe that testing suggests a wealth of metaphors, the most clear of which is based on the image of good management. Testing provides a "We mean business" orientation and functions as a lever on the efficiency and effectiveness of educational organizations. It provides a mechanism that promises to demonstrate how schools can be focused and be made more efficient. (See, for example, Kirst, 1981, p. 61.) In the most simple terms, testing sends the message that schools (and the expenditures that support them) can be managed. Thus, testing offers a convenient communication vehicle and one that is backed up by sanctions. The content of tests say "This is important! Pay attention!" Societal ascriptions of test importance are functionally derived from how tests are communicated to and interpreted by policy makers and the public. The stability of this perceived importance may turn out to be independent of the actual effectiveness of tests in improving educational quality.

Tests cost money, but their costs are relatively small compared to options such as adding teachers or investing dramatically in staff development to update teachers' content knowledge and pedagogical skills. Tests may not, in fact, be a quick fix; but they may be a cheaper option than grass-roots restructuring and reform. And they are tangible and palpable. Educational reform often deals in ideas, in words, and concepts whose distinctions are not well understood by the public. Recall, for instance, the public furor that cropped up when it was "New Math" time. Tests almost magically avoid such confusion. Everyone knows what a test is. Furthermore, tests may be one of the few options that can be imposed top-down (from the statehouse, for instance) and that appear to have any effect at all on our diverse educational settings. And as tests become policy instruments, with public political investments behind them, it is not too surprising that their findings take on more portent to those who require them.

Tests as Indicators

When tests are seen as one component of a system of outcome measures or indicators rather than the creators of effects, our need to attend to them differs dramatically. In some ways, the differences are paradoxical, and depend upon the conditions under which the test results are actually used.

For example, if a test is seen as a policy device ("Teach this because it is important"), then that which does not appear on the test loses credibility and currency in the school environment. If tests do not include science, then science (or art, music, history) may not be taught seriously. However, if tests are seen as one of many indicators of data that bear on educational quality but do not define it, then our teaching and curriculum do not hinge as precariously on what those tests indicate. The arguments for test-driven education (Popham, 1974) are persuasive, but I believe they ultimately can decrease the long range stability of an education system designed to improve performance for all students. For when tests are intended not only measure the effects of particular reforms but are the reforms themselves, the interpretation of positive and negative patterns of growth are extremely problematic. Interpretations of increases or decreases in performance are confusing. (Although increases in performance are not studied too diligently; they are usually accepted and attributed to the most recent set of reforms.) Did the test succeed or fail to communicate new standards? Were the programs and instruction that were put in place inferior? Are there any data conditions under which the policy of testing is itself questioned? Another difficulty results from the logical interest in looking at tests over time, to infer trends of various sorts for policy action. Such trend analyses place content constraints

and technical requirements on measurement that limit the real match between what are or could be important educational goals and what we, some years earlier, made a commitment to measure. Unless these concerns are explicitly accounted for, they can only ultimately impede our understanding of educational quality.

I believe that we are now in a phase where the transformation between tests as indicators and tests as policy instruments is under way. This transformation is caused in part by the staunch and apparently impervious belief in the validity or "hardness" of measurement data. It also is pushed by the insidious proposition submerged in the notion of testing as policy: to wit, if it isn't tested, it isn't important. When a system explicitly attempts to measure all the important areas of schooling—a task at which it can never succeed—the requirement for inclusiveness damages the entire educational enterprise and unbalances schooling. Part of the damage is caused because tests as policy instruments are almost always indirect. Teachers are tested because someone neither trusts the quality of their selection and the preparation they receive at colleges and professional schools of education nor knows how to influence them. Children are tested because we aren't sure teachers know how to teach them. Tests run downhill from the issue that we really wish to influence and often onto the people who are the recipients rather than the instigators of the suspect policies.

Maps of State Testing

In this section, I propose to shift from a general position statement on test functions to a relatively detailed description of the topography of state testing. I will report on state testing activities at a particular point in time, and attempt to provide a snapshot of some of the intense activities in testing at the state level. The purpose of this description is to show what is being tested where, what investments are being made, and to set the stage for a section which follows that demonstrates how our educational system responds to such mandates.

The impetus for state level testing was multiple, but can undoubtedly be attributed to the changes in funding for education derived from the Serrano decision, which was related to the state's responsibility to "equalize" educational expenditures and preempted responsibility on matters of accountability from local agencies. The retreat from programmatic Federal action in education that began with the Reagan Administration lodged additional power and initiative at the state level. Testing programs, already in place in California, Florida, and New York, became the focus of much new state activity.

How widespread was this activity? Let's start with a time period following the release of *A Nation At Risk* (National Commission on Excellence in Education, 1983), the U.S. Department of Education report that undoubtedly stimulated much state level reform. At the end of 1984, 39 states were operating at least one statewide testing program. Thirty-five states were conducting "Assessment Programs," programs that were to monitor the overall effects of educational services in the state in terms of student achievement. Thirty-six states were operating minimum competency testing (MCT) programs. Twenty-two states had both assessment and MCT programs. These data were developed as part of major study on the feasibility of using existing state achievement data to provide a picture of the national achievement of U.S. students (Burstein, Baker, Aschbacher, & Keesling, 1985). This study demonstrated clearly that a major investment had been made in testing. The rest of this section will draw upon this study as its primary source of information.

Who Gets Tested?

What were state testing programs like? Who were the students tested? Testing in states focused on eighth grade, with a total of 32 programs testing at this level. Other

frequently tested grades were grades 3, 4, 6, 10, and 11. Least frequent grades tested were grades 1, 2, 7, and 12. At the time of the report, 24 of the states were testing all students at the target grade level(s)—census testing—and as one would expect, all competency programs tested every eligible student. Most states with testing programs tested children in more than one grade. What other information was collected about the students? The most frequently obtained data were about students' sex and ethnicity, although about one-third of the reporting states did not require such information. Language status and program participation, e.g., Chapter 1, were information items collected by a relatively few states. Peculiarly, student age and years in school were of interest to only one or two states.

What's on the Tests?

What content was tested? In almost every case, the content areas tested were in reading and mathematics. Fewer than half the states giving tests conducted writing assessments, using student essays as the data. But more than half tested in at least one additional content area, such as social studies, science, or language arts. This research also reports a detailed set of analyses which focused test content. These analyses were developed by carefully categorizing the items on actual copies of these tests, or in some cases where test security was an issue, by inspecting the test specifications and sample test items. The research team first developed a model to guide our analyses consisting of a relatively flat hierarchy of major skills and subskills, shown in Figure 1. Based on this kind of analyses, major skill categories were developed for reading, mathematics, and writing; these are listed in Table 1.

Figure 1
Relations among Content Areas, Major Skill Areas, and Subskills

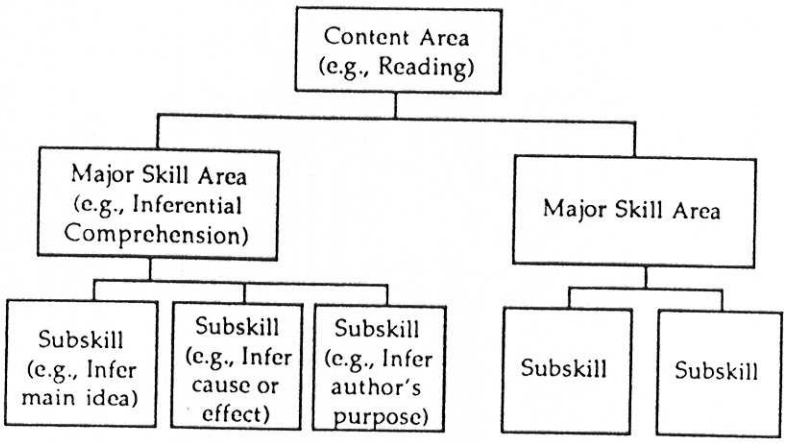
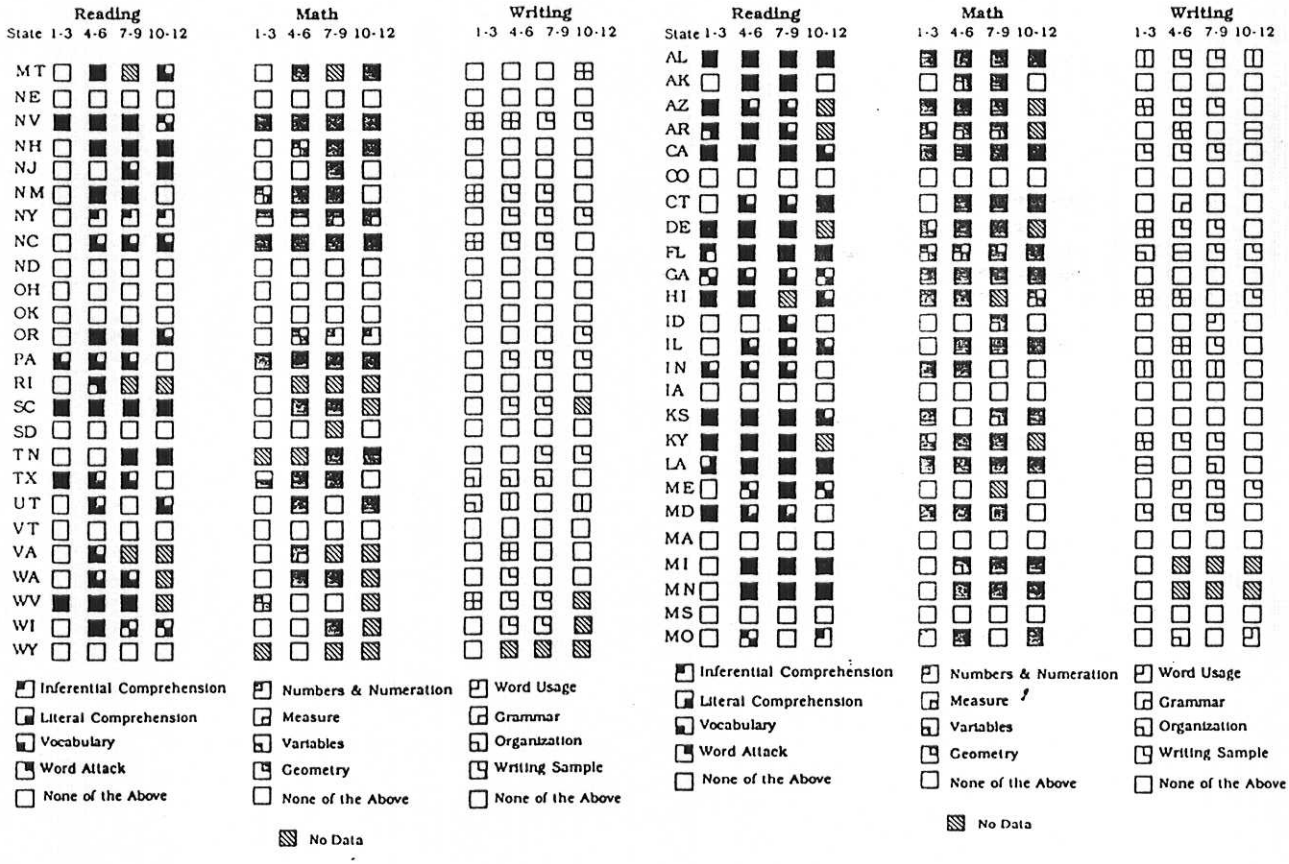


Table 1
Major Skill Areas Exhibited in State Testing Items

Reading	Mathematics	Writing
Inferential Comprehension	Numbers & Numeration	Grammar
Literal Comprehension	Measurement	Word Usage
Vocabulary	Variables	Organization
Word Attack Sample	Geometry	Writing

The graphic display presented in Figure 2 was created by Leigh Burstein. A quick review will give a good picture of the distribution of skills by content area and grade level tested.

Figure 2
Distribution of State-Tested Skills



The team's analysis was more intensive.[1] Analysis of the skill areas was decomposed an additional level into subskills, and examples of the type of items measuring such subskills were provided. In Table 2 an example of the inferential comprehension tasks in reading are presented.

Table 2
Decomposition into Subskills and Items for the Inferential Comprehension Skill Area in Reading

1.	DETAILS, SUPPORT STATEMENTS	(Given passage) Which statement best supports James Lee's claim that the late bus would benefit students? a. The school board should find a way to resume the services of the late bus b. Extracurricular activities provide students with valuable learning experiences c. Some students can get rides from their parents d. Some working parents cannot take their children home from school
2.	MAIN IDEA, SUMMARY, TITLE	(Given passage, infer best title, summary statement, title) The main idea of these rules is that: a. both adults and children enjoy the swimming pool b. there is a snack bar at the swimming pool c. safety is extremely important at the swimming pool d. the swimming pool is open every day
3.	MISSING/ IRRELEVANT INFORMATION	(Given passage, infer missing information or identify important information to include or exclude) Which of the following would be most important for the editors to include in this editorial? a. The school has never given the band any money for its uniforms b. Helmets and padding protect football players from injury c. Members of the marching band perform indoor concerts too d. The football team has longer practices than the marching band

[1]This work was primarily conducted by Pamela Aschbacher.

Table 2, continued

4.	MISSING WORDS	(Given reading passage with several words omitted, identify best word to fit in blank from context.) (Note: New York's entire reading test was like this.)
5.	SEQUENCE	(Given a passage, infers order of events or logic.) What indicates that Minnie was the first in her neighborhood to have a sewing machine? a. The neighbor women all came to see it b. She had to make everyone's clothes c. Fred bought it d. She didn't know how to operate it at first
6.	CAUSE/EFFECT	(Given passage, infer cause or effect) A major reason Paramount Studio moved to California was to: a. allow the Army to use the Astoria plant b. avoid the destruction of the studio by vandals c. enable the Astoria plant to become a museum d. be able to make movies less expensively
7.	CONCLUSIONS	(Given passage, chart, etc., draw conclusions) Based on the information in this chart, it may be concluded that: a. cross-ventilation helps to warm a room b. gas heat is more expensive than electric heat c. fans use very little electricity d. insulating walls conserve energy all year round
8.	PREDICTIONS	(Given a passage, predict probable outcome) What probably happened next in this story? a. The girl became angry and went home b. Marina and the girl told each other their names c. The girl made fun of Marina d. Marina became embarrassed and stopped talking

Table 2, continued

9.	FACT/OPINION	(Given passage or statement, distinguishes fact from opinion) Which of the following is an example of an opinion? a. "In 1860, a midwestern stagecoach company let people know about an exciting new plan." b. "The mail must go through." c. "The route cut directly across from Missouri to Sacramento." d. "Each rider rode nonstop for about 100 miles."
10.	PURPOSE, ATTITUDE	(Given passage, infer author's purpose or attitude) The author's attitude toward the Pony Express riders can best be described as one of a. confusion b. amusement c. worship d. admiration
11.	CHARACTER	(Given passage, identify character traits, identify motivations, draw conclusions about character's feelings) The beasts and birds can best be described as a. proud and closed-minded b. understanding and wise c. sleepy and lazy d. thrifty, hard-working
12.	FIGURATIVE LANGUAGE	(Given passage, identify meaning of metaphor, simile, idiom, or other image or figure of speech used) The author's choice of words "sets up business" and "cleaning station" are used to show that a. the wrasse's means of getting food is almost like a business service b. wrasse fishing is big business c. all fish set up stations d. the wrasse enjoys cleaning itself in the water
13.	TONE	(Given passage, recognize mood) At the beginning of the story, the mood is one of a. disappointment and sorrow b. curiosity and excitement c. fear and suspense d. thankfulness and joy

Table 2, continued

14 COMPARE, CONTRAST (Given passage, infer similarities, differences)

Compared to American managers, Japanese baseball managers are

a. better advisors
b. better paid
c. more knowledgeable
d. more powerful

15. ORGANIZATION (Given passage, select portion to complete outline or organizer based on organization of passage)

The following outline is based upon the last paragraph of the passage. Which topic below is needed to complete it?

A. Federalists
B. Republicans

a. Competing parties
b. Jefferson's rivals
c. Election pay-offs
d. Strong governments

16. SETTING, PLOT DIALOGUE (Given passage, identify and interpret time, place of story or event)

You can tell that his story took place

a. in a city park
b. at a zoo
c. in a forest
d. near a boot factory

17. LIT TYPE (Given passage, recognize example of fiction, nonfiction biography, autobiography, similes, metaphors, etc.)

The reading selection appears to be an example of

a. an autobiographical account
b. historical fiction
c. a biographical sketch
d. ancient mythology

Using this analytical framework, the distribution of state efforts was categorized in terms of the range of topics covered, the "spread" of items across subskills, the depth of coverage within subskills, or how many items, and the distribution of items on subskills classified as higher order skills, or skills with cognitive demands of inference, application, or problem solving as opposed to mere information retrieval by students. Eleven states were found to have relatively broad subskill coverage. In depth of coverage, only California had many items for each subskill area, a phenomenon directly

related to California's matrix sampling approach. (Many items for many skill areas on a census test would create a time and fatigue burden for students.) In other states, depth was a function of topic and grade level tested. About one third of the states included higher order test items in their testing programs. This analysis was used to help identify the commonality of tested skills and was, in fact, conducted, as part of a feasibility study to examine aggregation of state tests to serve as a national indicator of school achievement.

Recently, the Office of Technology Assessment (1987) provided an update on the grade, general content area, and ancillary data collected in state assessments. Because their data collection was within six months of the UCLA study, not surprisingly no major changes emerged. The Office of Technology Assessment (OTA) report did include some snapshots of state testing policy history and plans from a sample of eight states. What is striking is that in almost every case, the move is to more testing: in more grade levels, for more subject matters, for wider numbers of students. California cites plans for consolidation of local and state measures to meet this state goal (Bennett & Carlson, 1986). In addition, the institution of the Golden State Examinations are to provide individual incentives for students to achieve higher standards, standards which will be demonstrated by taking appropriate tests; successful students will receive special recognition on their diploma. California also uses a cash incentive program for schools to raise scores on their twelfth grade California Assessment Program (CAP) scores. The policy investment in this approach is high. Bennett and Carlson say:

> Standardized tests are expected to focus the attention of educators and policy makers at all levels on the knowledge, skills, concepts, and processes which are essential for success in the more demanding high-tech job market of the future, for responsible citizenship, and for personal fulfillment. The core of content and skills to be spotlighted represents a rigorous curriculum in the humanities, natural sciences, and math, and emphasizes higher-order skills such as those required to analyze complex relationships, draw inferences, and reason deductively. Although it is assumed that in practice, the scope and pace of the curriculum will reflect differences in aptitude and intelligence . . . it is also assumed that the majority of students are not working up to their potential, and that it is the responsibility of the schools to challenge them to do so—both for their own good and for the good of the society. (p. 169)

The Colorado state testing summary shows what happens when a test without a history of large scale assessment moves into it. (Martin, 1987).

Colorado has maintained an approach that is still at a very general level compared to recent efforts in California, Illinois, or in the Southeastern states. Nonetheless, during their deliberation, the tendency to use such tests as an omnibus solution to all problems arose. We specifically had conversations with policy makers on the costs and benefits of using such measures as indicators of teaching performance. We also have been observing other state efforts that attempt to generalize the use of student testing measures to teacher assessment, an approach that is clearly a bad idea on conceptual and technical grounds.

Systemic Responses to Tests as the Conveyors of Standards

In this section, I wish to recount briefly the findings of some of my colleagues at the Center for Research on Evaluation, Standards, and Student Testing (CRESST). One of our three research programs is studying the impact of testing on educational quality. Our question was simply whether having such standards and tests (as created by recent state and local reform efforts) helps or hurts the educational system. One question was

whether such tests as those required for high school graduation do in fact have positive or negative effects. James Catterall (1987) has reported on a study of those ten states with four or more years of a high school graduation test requirement. Interviews with educators were obtained that described their analysis of the function of such tests. Catterall went on to select two states with the highest and two with the lowest graduation rates. A survey was administered to 736 students sampled from within three representative districts in each state. On the basis of his survey, Catterall predicts that dropping out is significantly related to failure on such competency tests, and finds an interaction for Hispanic students. Expectedly, socioeconomic class is also a strong predictor as well as "track" in school. Furthermore, such relationships were "invisible" in his terms to the school personnel interviewed. If his findings are confirmed, then the role of tests as the conveyors of standards may need some review. Clearly, we can improve overall performance by driving out poorly performing students. But such a function is diametric to our intentions.

The second study, on the Texas Examination of Current Administrators and Teachers (TECAT), was conducted by Lorrie Shepard, a CRESST researcher from the University of Colorado. Shepard, Kreitzer, and Graue (1987) did an extensive set of work that looked at the TECAT from its policy inception to the results and remedies that resulted from its administration. The TECAT was designed to identify the teachers and administrators who were not qualified to serve in educational roles. Shepard and her team cast the impetus for the TECAT in the context of the need to revitalize the state economy and the goal to be a center of high tech involvement. Shepard traces the roles of the newly elected governor and the importance of H. Ross Perot, a successful technologist and head of the state task force on schooling. The timing of the Nation at Risk report was also critical. Shepard points out the TECAT was a literacy test measuring precollege and writing skills, with "harsh consequences" (Shepard et. al., p. 87), since failing the test twice resulted in loss of job. The process of preparing for this test was reported by Shepard and seemed to focus on succeeding at the particular test items included, and not reducing the kinds of grammatical errors ("he don't") that partly instigated public support in the first place. The TECAT had a passing rate of 99%. Because the standards were set so low, teachers could still make a few flagrant errors and pass. Moreover, the people the test "identified" may have been "real losses" to the system, such as those who worked with the institutionally mentally retarded, shop teachers who hadn't been certified through the usual means, and minority teachers. The test actually failed "1,199 teachers with some of the worst grammar skills. It may also have forced out another 1,000 to 2,000 teachers who considered themselves at risk on the test." (p.89) Shepard et. al. conclude that the TECAT harmed public opinion about education and involved a set of:

> ...unforeseen consequences: enormous cost, frenetic preparation and worrying about the test, demoralized teachers, and a public disillusioned by the high pass rate. Although these outcomes were not intended, they may be inevitable features of a reform that hangs so much importance on a test pitched to the lowest level of performance on the lowest of teaching skills. (p. 91)

The work of Shepard et. al. suggests that intentions are insufficient to assure positive use and interpretation of test results. Rudner (1987) reports on the status of teacher testing in 44 states. His analyses suggest that the "impact of such tests on minorities has been severe." (p.5) In a paper by Algina & Legg (1987), the validity and technical decision process is criticized. Given the demonstrable negative effects and potential validity problems, teacher tests as approaches to reform should be more carefully scrutinized.

Finally, Ellwein and Glass (1986), in a study conducted for CRESST, presented a series of case studies of standard setting using tests. In a shorter version of this study (Glass & Ellwein, 1986), the authors briefly summarize the six case studies, which

investigated four states and two local districts. By analyzing the intentions of such programs, in contrast to their operations and effects, the authors generate a devastating set of summary observations. They are:

1. When standards on tests are raised, safety nets are strung up (in the form of exemptions, repeated trials, softening cut-scores, tutoring for retests, and the like) to catch those who fail. If 100 incompetent persons enter the arena, 99 will ultimately survive. The one who doesn't was probably no less able than many, but lost heart and quit.

2. Both the courts and professional educators honor the principle that students should be warned of impending standards and remediated when they fail.

3. Even the most orthodox and doctrinaire justification of cut-scores in terms of skills and competence is moderated in the end by consideration of pass-fail rates. Norm referencing drives out criterion referencing. Pre-criterion referencing exists only in textbooks and scholarly journals; it is not found in the world of practice.

4. People focus on first-test failure rates and are less interested in ultimate failure rates.

5. In raising educational standards, the more technical looking approach packs more political muscle. The language of arbitrary authority is despised. The language of technical rationality is widely honored.

6. Cut-score determination methods require the added authority of political symbols for their credibility (titles, political composition of groups of judges, technical authority such as ETS)—these symbols are invoked to lend authority to what is actually a quite arbitrary procedure.

7. Managers of the educational system will act to soften the hard edges of technology and reclaim political discretion that has been appropriated by zealous technologists (in this case, technologists who would turn over the responsibility for determining who graduates from high-school or is licensed to teach to a test and a statistical standard).

8. Universities are raising standards, in part as an attempt to get out of the business of remedial instruction. But in a showdown, excellence comes in second to economics; competence loses out to enrollments.

9. In the end, standards are determined by consideration of politically and economically acceptable pass rates, symbolic messages and appearances, and scarcely at all by a behavioral analysis of necessary skills and competencies. The latter are relied on to the exclusion of the former to the extent that passing or failing the test has no lasting consequences in the lives of either students of teachers. (p. 4)

Glass and Ellwein distinguish between instrumental and symbolic acts and place testing and standard reform squarely in the symbolic category. As testing has shifted from something integral to instruction to a "policy" imposed from outside, it appears to have lost much of its assumed power. The authors conclude with a set of questions about standards and testing:

What purposes and political interests are served by raised standards? Whatever they are, we suspect they have little to do with the accomplishments and chances for "life success" of the pupils in whose name the reforms are undertaken.

What effect is the movement having on schools, teachers and the way pupils learn? Schools may be winning renewed public confidence. Teachers are bearing the brunt of both the blame for the crisis that brought about the reforms and the busy work that the reforms have engendered. Pupils take what is dished out and move on. (pp. 5-6)

The National Testing Scene

The questions above and the analyses by Catterall, Shepard, et. al., and Glass & Ellwein have growing salience in the light of a series of interesting policy deliberations related to the Federal role in testing. The National Assessment of Educational Policy (NAEP) was created as an indicator system (Tyler, 1965). However, the attempt by the Department of Education to create a national picture of educational performance with its infamous "wall chart" has begun to transform NAEP from an indicator to a reform instrument. The "wall chart," in summary, used college entrance examinations like the SAT to rank states on outcomes from best to worse without regard for socioeconomics, mobility or student ethnicity. Chief State School Officers attempted to argue for more valid measures (instead of against the entire enterprise). In fact, there was a short-term benefit to low rankings because it permitted arguing for greater resources from state legislatures, for reforms, and so forth. Subsequently, NAEP came under review by a broadly composed group of scholars and practitioners (Alexander & James, 1986). Part of their deliberations involved the redesign of NAEP to extend its sampling, reporting and interpretation to the fifty states. Thus, rankings or other measures of relative state performance would be possible. A series of discussions, planning activities, and now proposed legislation are moving this process along. While such a system would undoubtedly be an improvement over the SAT score base for state comparisons, the use of NAEP for such a function raises serious issues of the sort raised earlier in this paper and by my colleagues cited above.

State by state reporting would change NAEP from an indicator to a policy instrument since it would undoubtedly drive states to attempt to increase their relative standing. However, the existence of such a salient national measure has other implications. First, the content and skills tested would have to have widespread agreement among the fifty states. Clearly, to avoid making inappropriate inferences, states would want NAEP to conform as closely as possible to historical testing programs. If accurate, such a desire would drop content to the lowest common denominator. Secondly, the desire for trend data would tend to have a constrictive effect on the addition of new approaches to testing or to the addition or substitution of content areas. Third, the creation of such a test would result in a *de facto* national test and curriculum. While some claim that a national curriculum is already in place, created by the text publishing companies, the federalizing of standards through testing raises an alternative set of questions.

Further, a consequent of the adoption of NAEP with supplemental state funding as a proxy state educational outcome measure would be to reduce existing state assessment programs, since they will compete for some of the same funds. Thus, diversity will diminish. Under these conditions, a number of the concerns articulated by Shepard and Glass & Ellwein come into play.

Another set of concerns involve the national policy scene. A large investment in NAEP may reduce the actual support for other indicators of national achievement, such as the studies of comparative U.S. performance conducted through the

International Education Association (IEA) and supported by both government and private foundations. If NAEP becomes the megaoutcome measure of performance, then it will be used to assess a range of educational policy effects. Clearly, the danger of using a single measure is that it can produce anomalous results, as has been demonstrated in the recent difficulty in the 1986 NAEP reading scores (Rothman, 1988).

Without doubt, the functions of tests will continue to evolve. We can hope that continued, parallel research analysis of their actual functions becomes a regular part of the implementation or strong modification of any of the major testing programs. Tests as metaphors, signs, and symbols are important, but no less than their actual effects on educational quality and the people who participate within our educational system.

References

Alexander, L., James, H.T., & Glaser, R. (1987). *The nation's report card: Improving the assessment of student achievement.* Cambridge, MA: National Academy of Education.

Algina, J., & Legg, S.M. (1987). Technical issues. In L.M. Rudner, *What's happening in teacher testing: An analysis of state teacher testing practices.* Washington, DC: Office of Educational Research and Improvement.

Bennett, S.M., & Carlson, D. (1986). A brief history of state testing policies in California. In Office of Technology Assessment, *State Educational Testing Practices.* Washington, DC: Office of Technology Assessment.

Burstein, L. (1986). *Educational quality indicators in the United States: Latest developments* (Draft deliverable, OERI Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Burstein, L., Baker, E.L., Aschbacher, P., & Keesling, J.W. (1985). *Using state test data for national indicators of education quality: A feasibility study* (Final report, NIE Grant G-83-001). Los Angeles: UCLA Center for the Study of Evaluation.

Burstein, L., Baker, E.L., Linn, R., & Aschbacher, P. (1987). *Study group on pre-collegiate education quality indicators* (Final report, OERI grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Catterall, J. S. (1987). *Competency tests and school dropouts: Results of national study* (Draft deliverable, OERI grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation..

Cohen, D.K., & Haney, W. (1980). Minimums, competency testing, and social policy. In R.M. Jaeger & C.K. Tittle (Eds.) *Minimum competency achievement testing: Motives, models, measures, and consequences.* Berkeley, CA: McCutchan Publishing Corporation.

Coleman, S., Cronbach, L.J., & Suppes, P. (1987). *Research for tomorrow's schools: A disciplined inquiry for education.* New York: MacMillin.

Congress of the United States Congressional Budget Office (1987). *Educational achievement: Explanations and implications of recent trends.* Washington DC: Author.

Congress of the United States Congressional Budget Office (1986). *Trends in educational achievement.* Washington DC: Author.

Ellwein, M.C., & Glass, G.V. (1986). *Standards of competence: A multi-site case study of school reform* (CSE Report No. 263). Los Angeles: UCLA Center for the Study of Evaluation.

Glass, G.V., & Ellwein, M.C. (1986, December). Reform by raising test standards. *Evaluation Comment,* 1-6.

Harnischfeger, A., & Wiley, D.E. (1975). *Achievement test score decline: Do we need to worry?* Chicago: ML-Group for Policy Studies in Education.

Jaeger, R.M., & Tittle, C.K. (Eds.) (1980). *Minimum competency achievement testing: Motives, models, measures, and consequences.* Berkeley, CA: McCutchan Publishing Corporation.

Hirsch, E.D. (1987). *Cultural literacy: What every American needs to know.* Boston: Mifflin.

Kennedy, M.M. (1980). Test scores and individual rights. In R.M. Jaeger & C.K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences.* Berkeley, CA: McCutchan Publishing Corporation.

Kirst, M.W. (1981). *Loss of support for public secondary schools: Some causes and solutions.* Daedalus,110(3), 45-68.

Lazarus, M. (1981). *Goodby to excellence: A critical look at minimum competency testing.* Boulder, CO: Westview Press.

Martin, W. (1987). A brief history of state testing policies in Colorado. In Office of Technology Assessment, *State Educational Testing Practices.* Washington, DC: Office of Technology Assessment.

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: United States Department of Education.

Office of Technology Assessment (1987). *State educational testing practices.* Washington DC: Author.

Popham, W.J. (1974). *An approaching peril: Cloud-referenced tests.* Los Angeles: UCLA Graduate School of Education.

Ravitch, D., & Finn, C.R. (1987). *What do our 17-year-olds know? The first national assessment of what American students know about history and literature.* New York: Harper & Row.

Rothman, R. (1988, January 20). Drop in scores on reading test baffles experts. *Education Week*, pp. 1, 27.

Rudner, L.M. (1987). *What's happening in teacher testing: An analysis of state teacher testing practices.* Washington, DC: Office of Educational Research and Improvement.

Shepard, L.A., & Kreitzer, A.E., & Graue, M.E. (1987). *A case study of the Texas Teacher Test: Technical report* (Draft deliverable, OERI grant G008690003). Los Angeles: UCLA Center for the Study of Evaluation.

Tyler, R.W. (1966). Development of instruments for assessing educational progress. In *Proceedings of the 1965 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Wirtz, W. (1977). *On further examination: Report of the Advisory Panel on the Scholasitc Aptitude Test score decline.* New York: The College Board.