TESTING - HOW SHOULD EDUCATION BE EVALUATED?

Nelson, Carlson, Palonsky

The general theme of Part Four concerns the debate about the evaluation of education and the ways in which educational assessments ought to be conducted.

Evaluation is an assessment of value, a process for the determination of merit and worth, (Stufflebeam and Webster, 1988). Many aspects of education are subject to evaluation. Universities may test prospective teachers to find out whether they have the knowledge and skills necessary for success in the classroom. Schools test students to determine who merits special programs or scholarships, and to determine placement of students in one academic track or another. Schools also evaluate curriculum pro- grams in order to measure their effectiveness. Given their limited resources, schools are forced to evaluate their goals and decide whether it is better to support programs of excellence for the most talented students or general programs designed for all students. Arguments in Part Four also draw your attention to an evaluation of the public, a determination of whether or not public support for education matches the public's rhetoric of educational expectations.

The evaluation of education depends, in large measure, on the way in which schooling is viewed. For many years, the factory assembly line was among the most common metaphors used to depict education. Using the language of metaphoric comparison, schooling was described as a slow, thirteen-year crawl from raw material to finished product. Every year of schooling required that different parts be fitted to the product, each one enhancing its value. Along the way, measurements were made to ensure that prescribed growth was taking place' The product was regularly probed, poked, and tested and finally graduated. A few defective models were thrown out, but most were given the necessary correctives, and they made it through.

The metaphor is, in some ways, convenient and useful. Most people are familiar with factories, even if they never worked in one. Some factories, such as those run by Honda and Mercedes-Benz, turn out intelligently conceived, carefully crafted products. Other factories consistently produce lemons. The differences between the factories that turn out good products and the factories that manufacture bad products have never been entirely clear. It is suspected that the best products represent the most thoughtful designs and the highest assembly standards-Good factories evaluate their designs regularly to ensure that the-products manufactured are what the public will buy. Good factories also have exacting standards, and at every point along the assembly line they check meticulously for quality.

The earliest evaluation models were wedded to this factory model (see Smith and Tyler, 1942). Evaluation was narrowly conceived as the process of determining student merit. Learning objectives were established for students. Students were put in competition with one another and were graded and sorted according to the number of objectives they could achieve. The better students achieved more objectives; the best schools were those with the most students achieving the most objectives.

Despite its attractively straightforward reasoning, the production metaphor is no longer appropriate for education. The production model of education sacrificed individuality to accountability. The standardized outcomes prized by the manufacturing industry are not necessarily suitable for schools. No one assumes that a ton of steel, plastic, and rubber has the right to determine what it should be. Children are different. It is more than a bit disconcerting to think of them flowing along a conveyer belt, being fitted with identical skills, habits, and dispositions and denied any voice in how they should turn out or what is to become of them.

Today, educational evaluation means far more than determining whether objectives have been attained and ordering students in terms of who has achieved the most. Cur- rent evaluation practices raise questions about the appropriateness of objectives, the means by which objectives are established for students, and the ways in which students are examined. Evaluation is used not only to judge students but to help administrators evaluate the worth and merit of programs, and to inform the public about the level of attainment of education goals.

The three chapters in Part Four are designed to reflect some of the current debate about education evaluation. All the positions support evaluation; none argues that hi-h-quality education is likely to occur by chance. Schools must evaluate themselves and their students, both because they are accountable and in order to gather data required for good decision-making practices. However, reasonable questions can be asked about the gathering of evaluation data and the ways in which schools, teachers, and educational programs are asked to demonstrate their merit and worth.

Students planning for careers in education are aware that evaluation is playing a larger part in schooling than ever before, but they should know too that a great many questions have no generally agreed-upon answers. We have tried to include several of these questions in Part Four. For example, how should students be asked to demonstrate what they know? To what extent should society be held accountable for the outcomes of public education? Is it possible or desirable to use standardized tests that are designed to compare students? Is it fair to evaluate schools simultaneously for their excellence and their equity? Should schools be evaluated by the excellence of education they provide to the most able, or should they be judged by the manner in which they offer appropriate education to all students? As one educator puts it:

The fundamental question that any theory of evaluation must address is not what can be evaluated, or how. or whether or not objectives have been achieved. but how it is that humans come to know in the first place. And in the second, how it is that they represent what they know to others. (Eisner, 1985, p. 229)

The information in this introduction is designed to provide you with background information necessary to consider the competing perspectives found in Chapters 16 through 18.

STANDARDIZED TESTING

Educational and psychological testing represents one of the most important contributions of behavioral science to our society. It has provided fundamental and significant improvements over previous practices in industry, government and education. It has provided a tool for broader and more equitable access to education and employment. (American Educational Research

Association, 1985, Standards for Educational and Psychological Testing. Quoted in Mehrens and Lehman, 1987, p. 4)

Testing's effect on society extends far beyond the matter of who is admitted and who is rejected, of who is hired and who is not. Since what is tested directly influences what is taught, ETS's [Educational Testing Service] ubiquitous multiple-choice exams have an enormous impact on education, from kindergarten up through law school and beyond. And since what is taught influences how we live, the effect of these tests reverberates through society. (Owen, 1985, p. 261)

Too often, the public views evaluation narrowly, as a numerical indication of success or failure. Measures of achievement are confused with numbers, percentile ranks, reading levels, and the like. When asked about educational evaluation or the process of assessment, most people think of standardized tests, and for good reason. Standardized testing has become the most commonly used device to determine attainment and proficiency. As many as 300 million standardized educational and psychological tests are administered every year in the United States, most of them to public school students. A typical high school graduate will be subjected to six full batteries of standardized achievement tests in twelve years of schooling (Mehrens and Lehmann, 1987, p. 2). Test taking has become one of society's more widely shared phenomena, and it is unusual to run into anyone who has not taken at least one standardized multiple-choice exam.

Proponents of standardized testing argue that machine-scored multiple-choice testing instruments are the best available means for determining academic merit and for ensuring educational quality. Test advocates argue that well-designed tests, properly administered and interpreted, can provide schools with the information they need in order to make curricular decisions and judgments about the cognitive growth of students. Test programs can inform educators about the effectiveness of teaching, the power of certain courses to affect students, and the extent to which students of one generation compare with students of other generations or with students in other school districts. Standardized testing can also be the vehicle through which school districts demonstrate that the money spent on education is being used prudently; testing can justify expenditure on existing programs or indicate the need to increase or decrease spending levels in one area or another.

Opponents maintain that standardized tests are crude,-imprecise measures that re- ward superficiality, ignore creativity, and penalize those test takers who read too much into the questions. Instead of informing the public, it is argued, test makers confuse people with test results shrouded in an aura of mathematical precision. Rather than offering accountability, testing mistakenly applies the simpleminded methods of cost ac- counting to the complexities of the teaching-learning process. To meet demands for a large-scale testing program, testing and measurement experts have had to design instruments for a machine-scored multiple-choice format. Critics claim that few things of significance can be reduced to discrete bits and measured in a series of short- answer questions.

Standardized tests are one of the more controversial applications of social science findings to education, and despite the extent to which they have permeated every level of school experience, they represent a relatively recent development in education. Civil service examinations were first

administered centuries ago in China, but it was not until the nineteenth century that standardized exams became a common part of social and economic life. Nineteenth-century Britain, in the throes of an expanding domestic economy and an international empire, found that the demand for middle-class managers could not be satisfied by the traditional practice of patronage appointments. Large numbers of administrators were needed in the far reaches of the empire, and vacancies could not be filled only by tapping privileged males-the sons of civil servants, members of Parliament, or others with wealth and connections. Competitive examinations were introduced in Britain to open the civil service to a broader range of educated males.

The United States also viewed testing as a means to democratize the selection of government workers. Political abuse was rampant in the late nineteenth century. Those who worked for the government often secured their positions through pull rather than merit, and every change in congressional leadership was accompanied by wholesale shifts in officeholders, clerks, and cleaning staff. Civil service reform began with the Pendelton Act of 1883, which established competitive examinations for prospective public employees. Civil service tests were intended to provide a means of filling public offices on the basis of ability rather than party loyalty.

The original impulse for testing was meritocratic: to provided an objective measure of ability that allowed vacancies in public offices to be filled by the most qualified. Tests were to be used as means of demonstrating ability and securing entry to successful careers. Standardized tests were first used in Boston's public schools in 1845 to measure students' subject matter knowledge and to determine who was eligible for secondary education (Travers, 1983). In many ways, performance on standardized tests still controls access to education and power in society. Test results help determine a student's acceptance into selective school programs and admission to higher education and into prized vocations (Eggleston, 1984).

Testing is controversial. If standardized exams can deliver fair, unbiased access to the limited rewards of society, they are socially important and essential to democratic societies, and their use should be encouraged in assessment designs. But some argue that standardized exams restrict access to power, serve as biased agents of social control, are destructive to democratic ends, and should be abolished.

Standardized testing is used in screening teachers. Virtually every state now has in place or is considering testing teachers for licensure and certification; only seven states used such tests in 1981. Testing prospective teachers, like other forms of evaluation, is political (Darling-Hammond, 1990). Evaluation is always tied to larger policy issues and to judgments of comparative worth. By setting tests for teachers, test developers will determine which knowledge has the most worth, and test takers with less of that knowledge, or less ability to display it in ways called for by the testers, will be left out of the applicant pool for teaching jobs.

Chapter 16 presents two competing perspectives on standardized testing in assessment. Position I presents arguments against testing. It argues that little of real value can be measured by standardized testing. Position I regards standardized testing as a biased mechanism for controlling access to education and employment, stifling student creativity and motivation while perpetuating social injustice (Broadfoot, 1984; Crouse and Trusheim, 1988). Position 2, led by psychometricians and educational psychologists, argues in support of standardized testing,

saying that whatever is worth- while educationally can and should be measured through formal, objective evaluations. In order for public education to be accountable, decisions must be made about the ways in which resources are allocated. Supporters of the psychometric approach conclude that standardized tests are the best single means for gathering the data needed to make educational decisions and demonstrate responsibility.

PUBLIC SUPPORT FOR SCHOOLS

Applied to education, the term "evaluation" has typically come to refer to the assessment of student performance or the measurement of program effects. Chapter 17 directs attention to a different dimension of evaluation-the evaluation of the public- and it raises questions about whether or not the public has been adequately supporting its schools. The public has asked the schools to teach skills and values and solve social problems ranging from drug abuse to racial discrimination, but has the public been willing to give schools the financial support and academic authority necessary to do the job? Does the public's support for education match its rhetoric?

Although the questions raised in Chapter 17 may be difficult to answer, they are posed fairly. Educational evaluation is more than a technical process, a set of scientific skills with which to measure small, discrete bits of education (Scriven, 1983). Evaluators have been encouraged to see themselves in larger terms, as more than technicians. They have been urged to consider the social function of evaluation and to play a greater political role in the society (Cronbach et al., 1980; House, 1973). It is assumed that effective evaluation cannot be divorced from political and social ends. Evaluators must consider education as a political activity. The questions they ask and the ways in which they ask them rightfully link the schools to the people they serve.

Schools are part of the society that establishes them, and it is entirely appropriate to ask how well the public has been supporting the schools. If the society charges schools with a full portfolio of obligations, to what extent is the society responsible for providing the conditions that will allow the schools to discharge their duties? An evaluation of the public may be necessary in order to judge its intentions. Is the public charge to schools given honestly? Or has the public entrusted schools with a social and educational agenda which it knows schools cannot discharge, but which the public does not want to deal with squarely?

Position I presents the argument that the public gets from schools only what it has been willing to pay for, and that is not very much. Schools are publicly praised and privately patronized. The public may express great love for its schools, but it entrusts them with neither the money nor the power to bring about real educational or social change. Position 2 argues that the love affair between the public and the schools is the genuine article. In analyzing state and local statutes and data from public opinion polls, Position 2 finds much to praise in the public's record of support for its schools.

EDUCATIONAL EXCELLENCE

When schools are evaluated, questions naturally arise about the goals of schooling and the extent to which the pursuit of some goals may exclude the realization of others. For example, can the schools be held accountable for delivering academic excellence while maintaining educational equity? Are the goals of excellence and equity so in- compatible that in the reform of public education, society is forced to choose one or the other?

"Excellence" typically refers to rigorous educational programs and high academic standards. Excellent schools set lofty expectations for students and encourage all students to reach these goals (Adler, 1982; National Commission on Excellence in Education, 1983). "Equity" refers to the role played by schools in furthering social justice. Equity demands that schools provide appropriate educational opportunities and democratic advancement for all children regardless of their academic ability (Aronowitz and Giroux, 1985; Bastain et al., 1986).

At first glance, it might appear that no conflict exists between excellence and equity; education should be able to help the less fortunate while contributing to the welfare of the most able (Glazer, 1987; Strike, 1985). If "excellence" is defined in such a way that anyone can become excellent, then everyone can be treated the same way. All students can be given a similar education, and on the basis of merit the best will achieve excellence, the rest varying degrees of adequacy.

Some critics argue that such a formula serves only to reinforce social inequities. Education for excellence, they claim, too often leads to schooling that is socially repressive. Students do not enter school with similar advantages. In kindergarten, the children of the poor already perform at lower levels than the children of the wealthy, and the achievement gap between the two groups widens every year. Schools, it would seem, serve some children better than others. In the name of excellence, schools perpetuate social differences. These critics argue that schools should serve everyone, not just those who begin school with comparative advantages. Education for equity would put schools more squarely in the fight for social justice and an expanded democracy. Equity demands curricula that serve the career and personal needs of all students equally well (Aronowitz and Giroux, 1985; Freire, 1973).

Previous efforts to reform schools have been guided by the assumption that educational institutions cannot allocate sufficient human or fiscal resources to provide both Critics of the reports have been more skeptical, arguing-that they are part of a neo- conservative offensive in education. The endorsement of the reports by the political right (for example, the Heritage Foundation, Senators Orrin Hatch and Jesse, Helms, former Secretary of Education William Bennett, historian Diane Ravitch) has aroused the suspicions of political liberals who see the latest round of school reform as an elitist program designed to promote the success of the few at the expense of the many (Altbach, Kelly, and Weiss, 1985; Aronowitz and Giroux, 1985; Bastian et al., 1986; Pincus, 1984).

Chapter 18 presents a debate over the excellence and equity goals of education. Position I argues that the primary purpose of education is to improve the lives of people through knowledge, and that the quality of education ca n be measured only by its excellence. However, Position I finds

no inherent conflict between excellence and equity in planning for future school reform. Position 2 argues that schools cannot be excellent and equitable at the same time. The new call for excellence, it argues, is a sham, a smokescreen that promotes elitist and undemocratic goals in the name of re- form. Position 2 holds that excellence is an inappropriate goal for mass education.

REFERENCES

Academic Preparation for College. (1983). New York: College Entrance Examination Board.
Adler, M. J. (1982). The Paideia Proposal: An Educational Manifesto. New York: Macmillan.
Altbach, P. G., Kelly, G. P., and Weiss, L. (I 985). Excellence in Education: Perspectives on Policy and Practice. Buffalo, N.Y.: Prometheus Books.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (1985). Standards for Educational and Psychological Testing. Washington, D.C.: American Educational Research Association.

Aronowitz, S., and Giroux, H. A. (1985). Education Under Siege: The Consenative, Liberal and Radical Debate Over Schooling. South Hadley, Mass.: Bergin and Garvey.

Bastain, A., et al. (I 986). Choosing Equality: The Casefor Democratic Schooling. Philadelphia: Temple University Press.

Boyer, E. L. (1983). High School.- A Report on Secondary Education in America. New York: Harper & Row.

Broadfoot, P., ed. (1984). Selection, Certification and Control: Social Issues in Educational Assessment. London: Falmer.

Cronbach, L. J., et al. (1980). Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

Crouse, J., and Trusheim, D. (1988). The Case Against the SAT Chicago: University of Chicago Press.

Darling-Hammond, L. (1990). "Teacher Evaluation in Transition: Emerging Roles and Evolving Methods." In The New Handbook of Teacher Evaluation, edited by J. Millman and L. Dar- ling-Hammond. Newbury Park, Calif.: Sage.

Educating Americans for the 2]st Century: A Plan of Action for Improving Mathematics, Science and Technology Education for all American Elementary and Secondarv Students so that Their Achievement is the Best in the World by 1995: A Report to the American People and the National Science Board (I 983). Washington, D.C.: National Science Foundation.

Eaaleston, J. (1984). "School Exarninations--Some Sociological Issues." In Selection, Certif cation and Control, edited by P. Broadfoot. London: Faimer.

STANDARDIZED TESTING: RESTRICT OR EXPAND

Testing is virtually universal in schools. Should the use of standardized tests be increased or decreased?

Position 1

Restrict Testing

VEXED TESTS

In a witty attack on standardized testing, Banesh Hoffmann (1962) recounted a debate that was played out on the pages of the Times of London. A letter to the newspaper's editor asked for help in solving a multiple-choice problem from a battery of school tests taken by the letter writer's son. At first glance the question seemed to be straight-forward and not surprising to anyone who has attended public schools. It asked,

"Which is the odd one out among cricket, football, billiards, and hockey?"

The letter writer believed that the answer must be billiards because it is the only one of the four games played indoors. He admitted to being less than sure of his answer, and he reported that there was no agreement among his acquaintances. One of his neighbors argued that the correct choice was cricket because in all of the other games the object is to put a ball in a net. The letter writer asked readers of the Times for help.

Ensuing letters and arguments succeeded only in muddying the waters, since the logic supporting one choice was no more compelling than the logic supporting any other. For example, billiards could be considered the odd one out because it is the only one of the four games listed that is not a team game. It is the only one in which the color of the ball matters. It is the only one in which more than one ball is in play, and it is the only one played on a green cloth rather than a grass field. Unfortunately, equally convincing briefs could be submitted in behalf of the other choices.

Hoffmann fumed about the inherent cultural bias in the question. He assumed that the test was designed to measure reasoning ability and not knowledge of sports, but he argued that the test taker may be disadvantaged by having too little or too much experience with athletics. For example, not all students with good reasoning skills know how cricket is played. Test takers who know too much about sports might also be disadvantaged. They might choose hockey as the odd one out because it is really two different names that share the same name: in England and several other counties, hockey is typically played on grass by players who receive no salary; elsewhere it is played on ice by professional athletes.

The language of this test item may trip up students, preventing it from measuring reasoning ability. For example, many working-class students 'may not be familiar with either cricket or billiards. This item, not unlike many standardized test items, tends to reflect the language and culture of the middle and upper middle classes. Low scores may reflect measures of class, race, and ethnicity more than achievement or ability (Neill and Medina, 1989). Americans could also be at a disadvantage, confused by the language of test directions, which asks test takers to select the "odd one out." The question stem more commonly found in the United States asks, "Which of the following does not belong?"

Test questions of this sort seem silly. There is no readily apparent "right answer," and test takers have no opportunity to demonstrate the thought processes that led to their decisions. Multiple-choice questions are an unnatural problem-solving format that is discontinuous with the way in which real-life problems present themselves. Rarely are life's dilemmas accompanied by four answers, one of which is guaranteed to be correct. Good problem solvers in the real world are seldom locked away, deprived of books, computers, and human contact, and told to respond to a set of timed multiple-choice questions that have no meaning for them or anyone else. Hoffmann asked, "What sense is there in giving tests in which the candidate just picks answers, and is not allowed to give the reasons for his choice?" (Hoffmann, 1962, p. 20).

If multiple-choice questions, such as the one that vexed readers of the Times, were nothing more than parlor games, a form of Trivial Pursuit to amuse guests after dinner, there might be nothing wrong with them. However, as everyone knows, standardized testing, is serious business. On the basis of scores on standardized multiple-choice exams, decisions are made about placement in reading groups, about who should be admitted to the college-track programs in public high schools, who should go to elite colleges, who should be awarded scholarships, who should be admitted to medical and law schools, and who should be allowed to practice a profession or trade.

IF TESTING IS THE ANSWER, WHAT WAS THE QUESTION?

Standardized testing has an unsavory history. In the early twentieth century, defining "native intelligence" and attempting to measure it through the use of standardized examinations instigated one of the most controversial legacies of the testing movement (Gould, 198 1). Although some twentieth-century Europeans, such as Galton and Binet, attempted to measure mental capacities through individual and standardized tests (Cremin. 196 1), widespread testing was first used by psychologists working for the United States government during World War 1. The army was interested in classifying all new recruits, with special attention given to two groups: those of exceptional ability and those unfit for military service. Binet and others used individual IQ tests that were not well suited to large-scale testing; under the direction of American psychologists, the army developed the first mass testing program in history (Gumbert and Spring, 1974, pp. 87-112).

The army used the tests to answer questions about the placement of soldiers: Who would best fit where? How could the army best use the varied talents and abilities recruits brought with them?

After the war, colleges and universities bought surplus exams. The language of the army tests required only slight modification for use in the schools. The original instructions given to soldiers read:

Attention! The purpose of this examination is to see how well you can remember, think and carry out what you are told to do in the army… Now in the army a man often has to listen to commands and carry them out exactly. I am going to give you these commands to see how well you carry them out.

In schools, these instructions were changed:

Part of being a good student is your ability to follow directions. . When I call "Attention," stop instantly what you are doing and hold your pencil up--so. Don't put your pencil down on the paper until I say "Go." . . . Listen carefully to what I say. Do just as you are told to do. As soon as you are through, pencils up. Remember, wait for the word "Go." (Gumbert and Spring, 1974, p. 94)

The army used IQ tests to predict the ability of recruits to do well in the military. Schools looked at intelligence testing as a scientific means to group children according to the education they should receive. Terinan argued that intelligence tests could be used to sort children into differentiated curricula designed to prepare them for their appropriate lot in life:

Preliminary investigations indicate that an I.Q. below 70 rarely permits anything better than unskilled labor; the range of 70-80 is pre-eminently that of semi-skilled labor-, from 80-100 that of skilled or ordinary clerical labor; from I 00- I I 0 or I 1 5 that of semi-professional pursuits-, and that above all of these are grades of intelligence which permit one to enter the professions or other large fields of business. (Tennan, 1922, in Wolf et al., 199 1)

For many years, schools followed Terrnan's advice and used IQ tests to track children on. the basis of their test performance. Students performing at the lowest levels would receive an education designed to prepare them to be tractable unskilled laborers. Complex skills would be introduced into the curriculum of only the highest- achieving students. Intellect was viewed as a biological trait much like eye color or height: it was thought to be inherited, measurable, and fixed. To make education appear more rational and efficient, IQ tests were used to sort students into appropriate curricula (Callahan, 1962).

IQ TESTS ARE BIASED AGAINST THE POOR

Standardized IQ tests have been shown to discriminate against the poor. In the United States, Americans of European descent score more than 15 points higher on average than African-Americans. There is also a significant gap between the standardized test scores of European-Americans and Mexican-Americans and native Americans. Some educational psychologists believe that most differences in IQ scores are attributable to genetic endowment (Jensen, 1969). However, most anthropologists and educational sociologists argue that IQ is more reflective of

the child's socioeconomic status than his or her native ability (Ogbu, 1978). For example, when children are grouped ac- cording to family background and academic experiences, the differences in achievement scores between white and minority children tend to disappear.'

Test makers argue that the lower test scores of racial and ethnic minorities reflect differences in the quality of schools and bias in the greater society. However, re- searchers claim that the bias may well be in the tests themselves. They argue that the tests "reflect the language, culture, or learning style of middle-to-upper-class whites. Thus scores on these 'tests are as much measures of race or ethnicity and income as they are measures of achievement, ability, or skill" (Neill and Medina, 1989, p. 69 1).

The ability of IQ tests to predict later life success has been challenged. Many in the education community have come to view intelligence as more than a biological trait that can be measured by paper-and-pencil tests, and IQ tests are often challenged because of their bias as well as their inability to capture the test taker's range of abilities (Gardner, 1982; Gould, 1981). Although school administrators admit to making fewer curricular decisions on the basis of IQ scores today than in the past (NASSP, 1988), testing programs similar to those suggested by Terrnan are still used to determine entrance into most gifted and talented programs and to decide who takes business math and who takes algebra (Wolf et al., 1991). Standardized examinations are still used to measure aptitude, achievement, performance, interests, personality, and attitude.

MISLEADING THE PUBLIC

Validation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle combining the attraction of chess and mud wrestling. (Cronbach, 1988, p. 3)

Until the last few years, despite questions about the validity of individual test items on standardized tests (Crouse and Trusheim, 1988; Hoffmann, 1962; Nairn, 1980; Owen, 1985), test takers were never able to see a list of the "right" answers after they had taken the exams. The Educational Testing Service (ETS) of Princeton, New Jersey, and other test developers published only a few sample questions, claiming that full disclosure would compromise the tests. In order to make the tests reliable, they argued, many items had to be repeated from year to year, and the answers therefore had to be held back from public scrutiny. The ETS admitted that it was possible to construct new equivalent exams every year; however, it would be an expensive process whose costs would ultimately be borne by the test takers.

Recognizing the power standardized exams have on the lives of individual test takers, and not persuaded by the ETS's arguments, New York and California enacted legislation that allowed test takers to see the answers after they had taken the exams. These truth-in-testing laws revealed ambiguity in test items. In some instances, there were two or more correct answers. The ETS and other test makers took the issue to court, and in early 1990'a federal district court judge in-New York set aside the test disclosure law because it interfered with copyright laws. The matter is being considered on appeal. Whatever the outcome, the truth-in-testing laws have cast doubt on

the ability of tests to measure what it is they claim to measure and have opened up the is- sue of validity to public examination.

There is good reason for public suspicion. In some cases, the results of standardized tests have intentionally been used to mislead the public. Take the case of the 44magic mean," uncovered by a physician in West Virginia. According to newspaper accounts, the students in the state were performing above the national average on standardized tests. This was intriguing, considering that West Virginia has one of the highest rates of illiteracy in the nation. Further checking by the physician revealed that no state using this test was reported to be below the mean. The test results made even the worst test taker (and the school systems that bought the tests) appear to be above average. It was concluded that "standardized, nationally normed achievement tests give children, parents, school systems, legislatures, and the press inflated and misleading reports on achievement levels" (Cannell, 1987, p. 3). Unfortunately, this is not an isolated example (see Measuring Student Learning, 1988).

Indeed, by the late 1980s, it was hard to find any school districts or states that scored below the mean on nationally normed standardized tests. These data have contributed to what has been termed the Lake Wobegon effect, after the mythical Minnesota town created by Garrison Keillor, in which "the women are strong, the men are good-looking, and all the children are above average" (Fiske, 1988; Phillips, 1990).

The point is simply that for over fifty years psychometricians and companies that market tests have convinced the public that short-answer tests are objective, scientific measures deserving of public confidence and faith, when in fact these tests suffer from vagueness, ambiguity, imprecision, and bias. There is nothing scientific or objective about these exams; they are written, tested, compiled, and interpreted by highly subjective human beings (Owen, 1985).

BIAS AGAINST WOMEN

The results of standardized testing too often are marred by their bias against women and the poor. Take the Scholastic Aptitude Exam (SAT), a test commonly taken by college-bound high school students. The ETS has encouraged colleges and universities to consider the SAT a scientific predictor of students' first-year grades in college. Consequently, SAT scores are often used by colleges in making acceptance decisions. According to the ETS, students with higher SAT scores should earn higher grades during their first year in college. One study, however, indicates that the SAT might be a gender-biased exam (Rosser, 1987). The gap between men's and women's scores on the test is 61 points. Female test takers scored 50 points lower on the math section and 11 points lower on the verbal section of the exam. If the SAT accurately predicted grade point average, males would have higher first-year grade point averages than female students. But this is not the case. Despite lower scores on the SAT, women earned higher grades than men. The SAT, does not predict what it is supposed to predict: success in college. The scores students get on the SAT have less meaning than the ETS has promised. Rosser concluded that because of sex bias in the SAT, women have a diminished chance of receiving financial aid, being accepted to college, and being invited to join programs for the gifted.

Because of an invalid exam, women are likely to earn less money and lose out in appointments to positions of leadership.

TESTS DRIVE SCHOOL CURRICULUM

Standardized tests are terribly flawed, but despite their problems they continue to exert tremendous influence. Every teacher knows that testing drives the curriculum. What is tested is taught. No teacher wants his or her students to perform poorly on standardized achievement tests, and no school administrator wants his or her school to be ranked below others in the state or district. Everyone in education knows that too of- ten, newspapers report the results of statewide testing in much the same way they re- port basketball standings. "We're Number One" or "County Schools Lowest in State" are not uncommon headlines in many local newspapers. To avoid invidious comparisons, instruction is geared to the test. Over time, material not tested tends not to be taught. Teachers and administrators fall victim to test makers' promises and the public's misplaced faith in testing. In truth, IQ tests are of little value in making decisions about children's education. Nationally normed achievement tests are often no better, and there is no compelling reason to subject students to large-scale multiple-choice exams. Why should students in any given school learn the same content as students in any other school? And why should all students be asked to demonstrate their level of academic achievement in the same way?

National testing has become a national obsession. Encouraged by President Bush's call for the development of "New World Standards" in each of "five core subjects" (U.S. Department of Education, 199 1, p. I 1), test makers began failing over each other in a mad scramble to rush testing plans to the market. Monty Neill of Fair Test argued that all these proposals are based on premises not supported by an examination of re- cent history:

During, the 1980s, U.S. schoolchildren became probably the most overtested students in the world-but the desired educational improvement did not occur. Fair Test research indicates that our schools now give more than 200 million standardized exams each year. The typical student must take several dozen before graduating. Adding more testing will no more improve education than taking the temperature of a patient more often will reduce his fever. (Neill, 1991, p. 36)

There is an antidote to standardized testing that does not sacrifice accountability. In every community, teachers, parents, and administrators should select appropriate con- tent based on the students' interests, experiences, goals, and needs. Teachers should teach the content with all the skill at their command and evaluate the extent of student learning with a wide variety of instruments. Students should be encouraged to demonstrate their ability to think, through written exercises, verbal expression, and informal papers, and they should be given ample opportunity to describe or demonstrate their reasoning. The assessment of student learning requires that educators develop a broader, richer array of measures. Student achievement cannot be reduced to a single numerical score. Multiple choice tests cannot tell the story of academic success. Standardized testing is deceitful and biased; standardized tests should be abolished. A student's record of school achievement should include a rich portfolio of papers, essays, videotapes, poems, photographs, drawings, and tape-recordings, not a series of test scores.

TESTING TEACHERS

The 1990s may see fundamentally new forms of teacher tests, implemented for new purposes and reflecting new views of the teaching profession and of teaching and learning processes. The prospects are exciting, the promises as yet unfulfilled. (Haertel, 199 1, p. 3)

Testing teachers is the latest example of social science sorcery promising to improve education. The current enthusiasm for testing teachers rests on several interrelated assumptions: (1) there is a clearly defined body of academic content knowledge shared by good teachers; (2) there is a clearly defined body of professional knowledge shared by all good teachers; and (3) prospective teachers should demonstrate their command of this knowledge before being permitted to teach.

Testing teachers is not a new business. Until the 1920s, tests were commonly used to screen teacher candidates. Teachers typically had less than four years of education beyond the eighth grade (Haney and Madaus, 1990). Local and county exams largely determined who would be hired by the schools. Local testing was abandoned at the urging of education reformers who argued that rigorous college-based programs for the preparation of teachers offered better assurance of quality than did testing (Gifford, 1986). The current enthusiasm for teacher testing comes, not coincidentally, during a period of general concern about the quality of college teaching and learning. There is doubtless a relationship between the qualities and knowledge possessed by prospective teachers and the likelihood that students in their classes will learn some- thin- of value, but the knowledge and skills needed by teachers are unlikely to be captured by standardized tests.

Consider the three assumptions that undergird teacher testing. The first assumption is that good teachers share a common body of academic content knowledge. At first blush, it seems entirely reasonable to ask prospective teachers to demonstrate a command of the subject they intend to teach. However, at present there is no agreement about the content a teacher needs to master in order to be effective in the public schools, and without such agreement there can be no rational and valid tests (Hilliard, 1986).

This problem is more difficult in some subject areas. Math educators, for example, were among, the earliest to define the necessary content knowledge in their field. How- ever, they had a certain historical advantage. As one math educator noted, most schools "teach eight years of 18th-century arithmetic, followed by a year of 17th-century algebra, followed by a year of 3rd-century-B.C. geometry" (Rothman, 199 1). The problem is greater in English and social studies, subject areas that do not reflect long histories of agreement concerning their content.

Identifying and testing professional knowledge is more difficult. Simply put, "a common knowledge base in professional education has yet to be identified or supported by the majority of professional educators" (Hilliard, 1986, p. 307). It has been argued that the National Teacher Exam has so little relevance to what good teachers do in the classroom that it should be abandoned (Owen, 1985). Tests cannot be relied upon to identify what good teachers know about teaching and how they use that knowledge to encourage students' intellectual and emotional

growth. The link between pedagogical knowledge and learning outcomes is too tenuous to justify restricting access to teaching to those who do well on tests of pedagogy (Sykes, 1990). Standardized tests that try to predict who will be successful in the classroom show no promise of value.

In fact, relying on standardized tests to screen teachers can do harm by excluding potentially good teachers while not guaranteeing that those who pass will be successful in the classroom. Requiring tests for teacher certification has had a devastating effect on African-Americans and other minorities. With the highest failure rates of any group of test takers, African-Americans are being turned away from teaching at a time when the percentage of African-American families with school-age children is increasing. Blacks are already victimized by poor public school education and the failure of the state and federal government to adequately fund teacher education in historically black colleges. The new wave of teacher tests has been described as "an academic electric chair" for African-Americans (Dupre, 1986).

Standardized tests are quick and dirty solutions to the problem of selecting teachers, and the tests should be abandoned. If you want to know who will be a good teacher, don't waste your time with paper-and-pencil tests. They are unable to predict future classroom effectiveness (Millman and Darling-Hammond, 1990). Selecting future teachers on the basis of tests is like licensing automobile drivers on the strength of only written exams. Every state requires a road test. Before teachers are selected, they must be given a "road test," observed by experienced teachers as they teach real students in real classes. Anything less is a fraud.

Position 2

EXPANDING TESTING

Because we need a reliable way to measure our progress toward the national goals set by President Bush and the governors. Because employers need a lot more than the high-school diploma to tell them what their applicants have learned. And because even as most indicators tell us that our schools are failing, America continues to spend hundreds of billions of dollars annually on an enterprise that has little or no means of accounting for results. It's time to develop a national achievement exam, required for all students. (Kean, 199 1, p. 36)

Education was "rediscovered" in the 1980s and carefully examined. Researchers, critics, and government officials raised questions about the quality of teaching, student learning, and school leadership. Public education was rescued from years of neglect, dusted off, and reassessed. After a long period of inattention, it was no surprise that problems were discovered everywhere, from the head to the tail of the academic pro- cession. The schools, it was generally concluded, were again in need of reform.

Previous generations of education reformers had concerned themselves with making education available to the children of all classes and races, and to a large extent they were successful. By the 1990s, a higher percentage of students were completing high school than ever before. Instead of availability, the current generation of reformers is now forced to consider the quality of those school experiences. As Mortimer Adler (1982) argues, the legal mandates for education cannot be satisfied simply by guaranteeing all children access to education. In order to satisfy the educational responsibilities of a democratic society, public education must demonstrate that each student acquires requisite skills and knowledge. Educational outcomes can no longer be measured only in terms of quantity-years of schooling and number of high school diplomas granted. Schools must guarantee that education has a demonstrably positive effect on students. Schools must show that students benefit from their years of attendance; that increased investment in schooling can be measured in greater ability to read, write, and do mathematics; and that moving up the academic ladder from grade to grade is based on merit rather than social promotion.

The issue of educational quality raises a broad range of questions:

How good is the education provided students in grades K through 12?

How do the students of today compare with former students?

How do students in School A or District A or State A compare with others?

How can prospective employers be assured that students who graduate from high school possess a minimum level of skills, knowledge, and ability?

How can the public know that the teachers who work in public schools are qualified to teach the subjects they are hired to teach?

How can taxpayers know that the dollars given over to public education are being well spent?

If changes are made in public education, how can it be determined that they have contributed positively to leaning outcomes?

Answers to these questions must be based on hard data. Schools need quantifiable measures of student performance and teacher effectiveness if they hope to maintain public support. Intelligent policy decisions should be based on objective information, and although no single means of data collection is sufficient, the data generated by well-designed standardized tests are crucial to an understanding of school outcomes. Good tests and good testing programs permit schools to gather information about curricula, students, and teaching personnel that are not available to them by other means. Without these data, schools cannot make good decisions about the quality of the curriculum, the ability of the teachers, or the power of specific programs to produce academic learning by students.

Testing is the scientific base that supports decisions about the art of teaching. It is also the yardstick against which society charts the progress and shortcomings of education, and it is the

form in which schools report the status of education to public officials and parents. Test and measurement experts are often at odds with others in education, and they have suffered abuse from' critics who are skeptical about the power of testing and testing agencies to influence public policy. The purpose here is not to answer the critics or submit a brief in support of the Educational Testing Service or the National Assessment of Educational Progress. Instead, we argue that (1) standardized testing is an essential tool for examining the measurable dimensions of education; (2) education has entered an era of accountability in which school officials must demonstrate that the money being spent for education is paying dividends in quality; and (3) new forms of performance testing available in the 1990s provide hands-on assessment techniques.

TESTING FOR THE GOOD OF SCHOOLS

Standardized testing is an essential element of rational curriculum work. The data generated by testing programs help curriculum planners determine whether the measured outcomes of a given set of instructional inputs match the intended goals. In other words, tests can help educators find out whether a specific program is working the way it was designed to work. When taxpayers are asked to foot the bill for a new science program in the high school or a new math program in the elementary school, they should be informed of the likely effects of these programs. They should also have hard data by which to judge how well these programs have worked elsewhere. It is a simple question of cost accounting and fiscal responsibility.

Effective change does not occur by chance. Educational decisions must be made about the progress of the students, the rate of achievement of proximate goals, and the best choice among the competing paths to the next objective. Education planners need to choose appropriate measures of student attainment. Impressionistic data are not sufficient; anecdotal evidence is not scientific. It is not enough that a program "seems to be working" or the teachers "claim to like" this method or that approach. Schools need to have better answers to direct curriculum questions. At what grade level are the students performing? What do diagnostic and prescriptive tests tell us about a child's performance in academic skill areas? How much of the required curriculum have students mastered?

In order for schools to make rational decisions, they must have hard, sound, objective data. Standardized testing should not be viewed as a report card but as part of an assessment system that permits schools to make decisions about curriculum and instruction. Standardized achievement tests are objective measures of performance. They are not designed to provide apologies for ineffective programs, nor are they arbitrary standards of excellence. Standardized tests are designed to measure the goals of education and determine the extent to which the nation is meeting its responsibilities to provide a quality education to all children.

SHOOTING THE MESSENGER

Determining educational quality across state boundary lines is especially difficult. The United States has no national curriculum, and although education is essentially an enterprise run by the individual states, Americans have a right to know how well their children are being educated

when compared With the children of other states and regions. Since 1969, the federal government has funded an assessment program known as the National Assessment of Educational Progress (NAEP). The NA.EP has been gathering data about the knowledge, skills, and attitudes of students across ten subject areas: art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies, and writing. Tests are given to four age groups (ages 9, 13, 17, and young adults). They have been administered since 1983 by the Educational Testing Service of Princeton, New Jersey. Education planners need to have the data generated by this program in order to reform schools.

Unfortunately, much of the test data have been negative; school children appear to know less today than in previous periods in our history. Although these findings grab headlines and cause a great deal of collective handwringing, they are not an end in themselves. The NAEP is designed to facilitate reconsideration of the quality of teaching and learning in public schools. Too often, the response to negative findings has been to blame the test makers instead of addressing the cause of poor scores. More energy has been expended attacking the validity of standardized testing than in examining the conditions revealed by the tests. It seems that it is easier to shoot the messenger than to consider an unpopular message.

In 1985, a project funded by the National Endowment for the Humanities and ad- ministered by the staff of the NAEP assessed students' knowledge of history and literature in a test called the National Assessment of History and Literature (NAHL) (Ravitch and Finn, 1987). The results were unequivocal: the 8,000 17-year-olds who took this exam were, in the words of the authors, "ignorant of much of what they should know." The assessment group included an equal number of boys and girls comprising a representative sample of the national population by geography and ethnicity. Among all test takers, only 20 percent could identify Joyce, Dostoevsky, Ellison, Conrad, o

Ibsen; fewer than 25 percent were able to identify Henry James or Thomas Hardy; only one in three knew that Chaucer was the author of The Canterbury Tales; 65 per- cent did not know what 1984 or Lord of the Flies was about. Three-quarters of the students did not know when Lincoln was president; one-third were unfamiliar with the Brown decision, 70 percent could not identify the Magna Charta.

Critics screamed that the test was not valid; it did not measure knowledge of the history and literature students learn in school. This criticism cuts to the heart of testing Wainer and Braun, 1987). The goal of psychometric testing is to provide policymakers with valid, reliable data on which to base decisions. Too often, criticisms of standardized tests come from people who are uninformed about the field of measurement. (Admittedly, this technical area seems to defy understanding by most of the general public and many educators.)

The NAHL was certainly a valid exam. It was written in cooperation with public school teachers, and the major portion of the questions were drawn directly from the most important material covered in textbooks and school curricula'. Most of the questions covered fundamental material that students of this age might reasonably be expected to know. Citing a handful of the literature questions-such as biblical references-that covered content not typically taught in school, critics raged that certain students were put at a disadvantage.

The detractors of the NAHL were apparently unmindful of the goal of the exam. The NAHL was not designed to grade students in the hope that many could be failed. It was designed to determine what students knew so that the curriculum and the nature of instruction could be improved. The test did not try to identify individual or typical 17-year-olds. The sample was stratified for sex, race, ethnicity, Geography, and private school attendance in order to reflect a national population. The test results were to be used as one body of objective data for considering what is learned in schools. The NAHL was not intended to replace teacher tests or substitute the judgment of the test makers for the individual judgments of state legislators or curriculum workers.

Standardized test results cannot be ignored. One of the goals of the NAHL was to provide baseline data for future assessments in history and literature. Relatively little is known about these fields of instruction other than enrollment statistics. It is, frankly, shocking that the test has been attacked so viciously. Although testing is far from a perfect science, at present no other measures can compete with standardized tests for gathering economical, valid, and reliable data about what children have learned in school.

The NAHL reported "large differences in achievement on both assessments [knowledge of history and knowledge of literature] among racial and ethnic groups" (Ravitch and Finn, 1987, p. 132). Asian and white students performed significantly better than black, Hispanic, and native American students. Rather than claiming that these differences are a function of biology, the authors suggest that geography and in- come are correlates of performance. For example, African-Americans in certain areas of the country raged behind other African-Americans. More of the highest-achieving African-American students came from families with a history of college graduation, and their families had home computers. These are important data that reflect the power of social and school forces to influence academic performance independent of race.

Admittedly, standardized tests, as a measure of educational achievement, are not without problems. Many of the goals of education are difficult to measure. The ability to communicate verbally and the possession of a healthy self-concept are hard to determine with paper-and-pencil tests. It is also well known among test developers that minimums tend to become maximums and teachers tend to teach for the test. Not everything that is taught can be included on an exam, and material not tested has a tendency to disappear from the curriculum. Test writers do not want to dictate what should be taught, but schools and the public must realize that tests of minimum competencies cannot cover everything that is taught in schools. Despite the risk of skipping some learning outcomes that are significant and emphasizing others that may have lesser significance, standardized achievement tests are a cost-effective means of ensuring educational quality. Instead of attacking the tests, those who do not like standardized testing should develop better measures of school achievement. So far, not one exists.

Opponents of standardized tests argue that too often the use of these instruments has resulted in discrimination against minority groups. Indeed, standardized testing is designed to discriminate, to make distinctions about what is known and by whom. If there were no differences in test scores-that is, if they did not discriminate among categories of test takers-the tests would be

worthless. It is not a question of whether or not tests discriminate. Rather, do the tests discriminate unfairly, and are the results of even the most fair tests used for unfair purposes? If a racial, ethnic, or gender group is found to outperform others on a standardized test, we cannot assume that the test is biased (Haertel, 199 1). However, such performance differences should set in motion a review to determine whether there was any bias in the construction of the exam, its administration, or the areas it was designed to assess.

PERFORMANCE ASSESSMENT

Much of the criticism directed at today's assessment may well be an attack on older test designs. Most of us are familiar with tests that indirectly measured what we knew. For example, a test maker who wanted to determine a student's woodworking ability might devise a test that was little more than a series of multiple-choice items. The student might be asked:

Which of the following tools would be needed to make a wooden bowl? (a) A ball peen hammer; (b) A lathe chisel; (c) A screwdriver; (d) A wrench.

Other questions might probe the student's knowledge of various types of wood, appropriate procedures for using power tools, types of finishing materials, and safety procedures. These items taken together might indicate a student's knowledge of bowl making, but the student's score on the test would tell the test maker very little about the student's actual ability to fashion a wooden bowl. A better measure---of ability would entail taking the student into a fully equipped woodworking shop and observing him or her set about making a bowl from a block of wood. This direct measure of performance would allow the test taker to demonstrate actual ability in a real-life situation, and it would allow the test giver to ask why certain procedures were followed or others omitted (Cizek, 1991).

Performance assessments of this sort have become common not only in vocational education but also in foreign language proficiency exams, as measures of writing ability, and in the sciences and mathematics. The NAEP first used performance components on tests in 1990 and plans further expansion of performance items. Performance assessment is one of the more exciting new developments in evaluation design. It promises to give educators at the local level answers to questions previously addressable only through high-inference measures. Performance testing can tell teachers, parents, and students whether or not a child can write an essay, conduct a science experiment, or make a wooden bowl.

Test critics argue that standardized testing leads to unsavory practices. For example, they argue that commercial materials are widely advertised, often fraudulently, as a means to raise scores on college admissions tests (Mehrens and Kaminski, 1989). Well, that may or may not be true, but performance testing would eliminate proxy measures of ability. Test makers and psychometricians are eager to design assessment instruments that allow test takers to demonstrate what they know in authentic situations. Critics of standardized testing seem to discount new developments in the field. Performance testing is already here. By 1994, the verbal section of the SAT will place greater emphasis on reading and reasoning; the mathematics

section will place more emphasis on data interpretation and real-life problems, and students will be asked to demonstrate how they got correct answers. "Multiple guess" will no longer be a valid synonym for multiple-choice assessment. Critics should consider the new advances in the field instead of attacking test designs that psychometricians have abandoned.

TESTING TEACHERS

Test makers cannot solve all the problems of education, but by developing good tests, they can bring scientific rationality to the licensure and certification of prospective teachers (Anrig, Goertz, and McNeil, 1986; Nitko, 1990). Testing teachers is not a trivial matter, or one that psychometricians enter frivolously or in simple pursuit of profit. Licensing exams are designed to identify the minimal level of content appropriate to the field being licensed (Pybum, 1990). They are used as a means of assuring, the public that impartial oversight has been exercised and that schools are staffed by those individuals who demonstrate an array of skills and knowledge deemed to be essential for the classroom.

When tests for prospective teachers were introduced in the 1980s, the antitesting lobby cried foul. They said that the tests were not fair to minority groups. A test may be unfair if members of minority groups (by gender, race, or ethnicity) score lower than others, and the scores are unrelated to what the test purports to predict. For example, say that a school district wants to hire a kindergarten teacher. If the school uses a test that measures knowledge of football trivia from the 1950s, you might guess that as a group, young, women would score lower on the test than middle-aged men. A hiring decision based on this test would be unfairly discriminatory because the knowledge being tapped would be unrelated to what is being predicted-success as a kindergarten teacher-and the test would tend to favor one group of test takers over others.

On the other hand, the use of scores on the National Teacher Examination (NTE) in making hiring decisions could be an example of discrimination that is fair. The NTE is typically administered to students completing teacher education programs. The distribution of scores shows regional, racial, and ethnic variations. However, this differentiation alone does not make the test unfair. The NTE is constructed by experts in education who argue that the content being, tested is most often possessed by the more able teachers and those who have not mastered this knowledge -&e disadvantaged by their ignorance. The NTE measures knowledge that is central to good teaching, so it is related to job success. It may be useful in hiring decisions because it accurately predicts who is likely to succeed in the classroom.

Albert Shanker, president of the American Federation of Teachers, argues that post-baccalaureate examinations are given to lawyers, engineers, and other professionals, and they should be welcomed by teachers. Critics have claimed that these tests are not designed to tap the ability of teachers to engage children in worthwhile activities, pique their curiosity, or encourage their intellectual growth-and they are right (see U.S. Department of Education, 1987). The NTE, which has been in use since 1939, is not a perfect exam, but it is of value because it will weed out those candidates who have not minimally mastered the subjects they plan to teach. The ETS has also announced that a new form of the NTE will be on the market. Recognizing that paper-

and-pencil tests have drawn sharp criticism, the ETS has invested $2 million in the development of a multitiered test for prospective teachers. In addition to the traditional test of knowledge and reasoning, the new design calls for live exercises, computer simulations, teaching situations with real students, and other practical exercises in which teachers can demonstrate their ability to make decisions and exercise professional judgment.

No test is designed to stand alone as the sole criterion for hiring. To a certain ex- tent, all standardized tests reflect environmental factors, including the level of education of the test taker's parents, family income, geographic origin, and the like. Intelligent use of the NTE in conjunction with subjective data from interviews, recommendations from universities, and so on has provided schools in thirty states with hiring, data for the past fifty years. The new NTE will be even better. It will give more powerful reassurance to parents that the new teachers who meet their children at the classroom door in September scored well on a nationally normed test.

Minimal investment in standardized testing programs to assist in educational decision making cannot help but pay dividends in better student evaluation and the wiser selection of teachers. The intelligent use of testing provides educators with scientifically generated data that are not available by other means (Smith and Hambleton, 1990). Licensure exams are not frivolous; they are necessary to protect children from potential harm inflicted by incompetent or underprepared practitioners (Shimberg, 1990). Without standardized testing, what kind of evaluation system would we have? How would we know whether schools were doing as well as we expect or whether teachers have the knowledge and skills demanded of them?