1

The Need for a New Science of Assessment

Harold Berlak

Introduction

The idea that schooling for all is essential for social progress and economic growth grew up alongside the development of industrial capitalism during the tail end of the nineteenth and early decades of the twentieth century. By the 1990s, the aspiration for universal schooling has come a long way toward realization, though many American youth still do not complete secondary school.[1] While universal provision of schooling is still widely seen as a noble, if unrealized goal, there is a growing consensus that the system of public education that has evolved over the course of this century in the United States is in serious trouble. Public officials, corporate leaders, and ordinary citizens are increasingly dissatisfied with the quality of the education provided by the nation's schools to the great majority of children. While the margins of the American political scene, left and right have long been critical of schools (albeit with quite different ideas of the problems and solutions), with the exception of racial desegregation, discussions of elementary and secondary schooling policy over the last 25 years were virtually absent in the national media, in the platforms of the national political parties, or in campaigns for national state or even local public office. For brief interludes, following the launching of Sputnik in the late 1950s and in the mid-1960s during Lyndon Johnson's "war on poverty," public attention focused on schools, but this interest was not sustained.

This changed in 1983 with publication of *A Nation at Risk*, a report of the National Commission on Excellence in Education (1983). It made national news with its assertion that American education was threatened by "a rising tide of mediocrity," and with its frequently cited lines: "If an

unfriendly foreign power attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. As it stands, we have allowed this to happen to ourselves. . . . We have, in effect been committing an act of unthinking unilateral disarmament."

Why this report received so much attention is a matter of some conjecture. Very serious problems, particularly in, but not restricted to, inner city and poor rural schools, had existed and been widely known for many years. In spite of the report's claims to the contrary, what had changed were not the problems[2]—though undoubtedly they had gotten worse—but the public's and elected officials' response. The reason for wide notice of *A Nation at Risk* had more to do with the particular historical moment it appeared than with the originality or profundity of its analysis. In the early eighties, the failures of the US economy had just begun to penetrate the nation's consciousness—dominating the news were the galloping US trade deficit; the failures of US industry; plant closings; and dramatic increases in unemployment, particularly in the older industrial cities. What this report offered was an explanation for these apparently inexplicable events, an explanation which was eagerly embraced by the mainstream press and corporate America, and widely repeated in the national media. The report told the American public that a major cause, if not the major cause, of America's fall from grace as the world's preeminent economic and industrial power was the failure of the nation's schools to educate a competent, dedicated work force. This was a palatable diagnosis of the nation's economic malaise that suited the times. It placed blame, not on the basic structural problems of the US economy, nor on the failures of corporate leaders and politicians to address the changing world economy, and to do something to relieve the accumulating social problems and the gross disparities between rich and poor; but on the politically impotent: the nation's elementary and secondary school teachers, nameless educational bureaucrats, and unskilled and/or unmotivated workers.

A Nation at Risk was not the work of right-wing ideologues. Terrell Bell, who initiated the report, and who was appointed by Ronald Reagan as his first secretary of education, was at the time widely regarded as a middle-of-the-road professional, and the eighteen-member National Commission on Excellence Bell appointed included, among others, the retired chairman of the board of Bell Laboratories, two professors from Harvard and University of California at Berkeley respectively, four university presidents (including Yale), a former governor of Minnesota, the immediate past-president of the National School Boards Association, two principals, two school board members, the superintendent of schools from Albuquerque, and the 1981–82 teacher-of-the year, a high school foreign language teacher from an affluent suburb of New York City.

Whatever its deficiencies, the *Nation at Risk* drew public attention to the schools, and this, attention contrary to the expectations of many, has continued to the present. The report and the wide attention if received stimulated responses from virtually every organization and group with an interest in educational policy. Since 1983 countless reports, articles, and books have been written or commissioned by every major foundation, dozens of minor ones, policy think-tanks across the political spectrum, associations of corporate executives and educational professionals, teachers' unions, children's and parents' advocacy groups, formal and *ad hoc* organizations of state and local educational officials, as well as by individual journalists and scholars. While there are major differences in the policy recommendations, very few reports contest the *Nation at Risk's* view of the economy, and none with dissenting views have received wide public notice.[3]

All this talk about education did, however, galvanize latent public discontent with the schools and create a political climate for change. Since 1983 virtually every governmental agency and administrative unit at the state, county, and school district levels that held some responsibility for elementary and secondary schools has initiated and implemented some reforms. State legislatures, governors, state and local education officers, the major foundations and think tanks, the two leading national teachers unions, and even the 1988 presidential candidates, Bush and Dukakis, felt the need to respond to the clamor for educational excellence.

Many of the responses can be passed off as media hype and political rhetoric. But there were also many concrete measures undertaken. I make no effort here to recount and analyze these efforts in any detail, a monumental undertaking far beyond the purview of this chapter. However, some effort to make sense of these intended reforms is essential if we are to understand the current movement for developing new forms of educational assessment and testing.

An Analysis of the Reform Movement: The Role of Testing

Two competing tendencies about how political decisions should be made and who should make them are represented by recent efforts to reform the nation's schools. One tendency is toward decentralization of authority and decision-making by those who are most immediately affected by those decisions. This view is often coupled with a distrust of centralized authority and a disdain for experts and intellectuals. From this perspective, "bottom-up" change is valorized along with direct, grassroots or participatory democracy.

The second tendency in this society is toward centralization of authority and decision-making, with responsibility for the difficult decisions left to the

man or woman at the top—the CEO, the chief of staff. In the case of schools, the superintendent or principal must be a tough-minded leader, able to shape up the troops, delegate responsibility and hold subordinates accountable for their performance. Efficiency and immediate, demonstrable results are valorized, and while democracy is not necessarily rejected, it is representative democracy and delegation of authority to those who know best which is endorsed—with little tolerance for participatory democracy, which is seen as chaotic and in the end as encouraging the lowest common denominator in terms of process and product.

The relative strength of these two tendencies and the ambivalence many Americans feel about how to reform schools are evident in the multiplicity of proposals advanced and policies instituted since 1983. The language that has dominated the discourse about school reform has been that of crisis, of disaster, of imminent threat to the very survival of the nation. I have already quoted *A Nation at Risk* with its military metaphors. Here are the words of *A Nation Prepared,* the second-most influential report, published by the Carnegie Forum on Education and the Economy (1986), created and supported by the Carnegie Corporation of New York:

> American's ability to compete in the world markets is eroding. The productivity growth of our competitors outdistances our own. As jobs requiring little skills are automated or go offshore and demand increases for the highly skilled, the pool of educated and skilled people grows smaller and the backwater of the unemployable rises. Large numbers of American children are in limbo—ignorant of the past and unprepared for the future. Many are dropping out—not just out of school but out of productive society.
>
> As in past economic and social crises, Americans turn to education. They rightly demand an improved supply of young people with the knowledge, the spirit, the stamina and the skills to make the nation once again fully competitive. (p.2)

In times of national crisis, it is no surprise that the strongest impulse by politicians most directly responsible for schools is to use their authority by employing the tools they understand and know best. In the United States, basic responsibility for schools resides with the states. Eight years after publication of *A Nation at Risk* virtually every state had instituted a combination of top-down measures intended to raise educational standards. These measures include requirements for academic courses, new or strengthened controls over textbook adoptions, mandated use of state curriculum guidelines which in some instances are closely aligned to required tests, and more pre-

scriptive regulations for certifying teachers. But, by far the most common measure is statewide testing programs throughout the grades that, in effect, increased the proportion of education dollars spent at the state level, and strengthened the control of the state's chief educational officer and/or state department of education.

While it is difficult to generalize about several thousand school districts, many, particularly the larger urban systems, responded much like state departments of education by tightening and centralizing bureaucratic control over curriculum, pedagogy, grading, student discipline, and personnel selection. In addition to the newly devised or revised state "basic skills" tests, and the standardized achievement tests which have been used for many years almost universally throughout the grades, some districts instituted their own district-wide tests, in some cases going so far as to specify textbooks for each grade level, and to link mandated tests to these texts.

The role of the federal government under Reagan-Bush is contradictory. On the one hand their administrations greatly reduced or eliminated programs supporting educational research and development, curriculum and staff development, as well as programs that aided particularly needy populations, using the justification that schools are primarily the responsibility of local and state governments. On the other hand, the Department of Education, whose elevation to cabinet-level status was bitterly opposed by Reagan and right-wing groups prior to 1980, in the ensuing years became an increasingly active instrument in efforts of right-wing forces within the federal government to shape local and state schooling policy through, for example, selective enforcement of and in some cases opposition to agreements reached by local and state school officials and the courts on civil rights issues, active advocacy of a national core curriculum, national assessment, and so-called "freedom of choice" plans which would, in effect, divert public funds to private schools. Among the more visible efforts by the federal government to shape schooling practice is the annual media event staged by the secretary of education upon publication of the "wall chart," which ranks the states' educational performance based on standardized test scores. In some instances a form of this annual ritual is repeated by states publicizing rankings of school districts, and by the central administrations of school districts releasing to the press rankings of individual schools within districts.

What explains the enormous emphasis on tests? I have suggested that a primary reason for this emphasis is that tests are a means of maintaining centralized control, providing those higher up in the educational bureaucracy (central office administrators, school board members, state education officials, legislators, etc.) with relative rankings of organizational units (classrooms, schools, districts, etc.) and/or students and teachers. This, however, is not an adequate explanation since it does not account for widespread popular

support for the use of tests. While there is increasingly vocal criticism of tests among professionals and by the national media, there is still remarkably little evidence of widespread discontent with current forms of testing. Indeed, many support increased testing, including African-American, and Latino-American parents who are convinced that their children, who consistently score lower on standardized and criterion-referenced tests, have been and continue to be victimized by low expectations on the part of teachers and school officials. For many within these communities, the only credible indicator of improved educational performance is improved performance on standardized tests. The irony in this is that, while the demand for more professional accountability is certainly justified, any gains on such tests are often temporary and local. The technology of these tests assumes there will be winners and losers, and in our society the winners are invariably the more affluent and the losers the poor and powerless.

Efforts to reform schools from the center continue, but a counter tendency toward more democratic school-level control has become more visible recently for several reasons, including organized opposition to centralized control by teachers unions, parent groups, and local school boards, and a growing conviction that mandating changes from above has not worked. What a few years ago was a fringe view that genuine changes in the end must occur in individual classrooms, which is not possible without active participation of teachers and without a large measure of autonomy within each school, has become increasingly accepted as the common wisdom by the public policy establishment and the mainstream press.[4]

Several states while tightening centralized control, have encouraged school–level decision–making by altering state regulations to permit principals and teachers more say about school expenditures, curriculum and staffing. Also several districts scattered across the country—New York City, Buffalo, and Dade County, Florida, are the most frequently mentioned in the press—not only tolerate but appear to foster school-level decision-making. However, although talk about, and arguments for, teacher empowerment and school-level governance are commonplace, it is the rare exception rather than the rule for central office bureaucracies to yield power.

This ambivalence over who should call the shots, the authorities at the center or the local school community, is probably nowhere more clearly exemplified than in the previously cited Carnegie report, *A Nation Prepared.* On the one hand, the report celebrates the role of the teacher and provides what it calls "a scenario," a hypothetical example of a high school run by the school staff in close collaboration with the local community. On the other hand, however, the report makes no recommendations as to how centralized administrative control by school districts or the state is to be relinquished. Its

key and sole concrete proposal is creating a new National Board for Professional Teaching Standards which would, in effect, centralize the certification of an elite cadre of master or lead teachers whom they assume would transform the schools.

If there is any consensus after almost eight years of intensive public discussion and activity, it is that tinkering with regulations and issuing more administrative mandates will not suffice, and that what is needed is *perestroika,* a basic restructuring of the entire system. *Restructuring* is one of those words like *democracy* and *accountability* that have an inexhaustible number of possible meanings, each aflame with ideological passion. At very least it implies an unfreezing of the central office bureaucracy and a shift in authority and the power of decision-making from existing to new formations.

In spite of the calls for *perestroika,* decentralizing authority, and empowering teachers and principals to institute changes from below, there has not been any wide-scale restructuring of the system. Except for some well-publicized exceptions, the evidence is that, overall, the system has become more and not less centralized over the past eight or so years. (Sarason, 1989) While there are several interconnected factors at work, one—if not *the*—single most significant in holding the current system in place, indeed in strengthening the current structures, is testing. Not any tests, but the *particular forms* of standardized and criterion-referenced testing which have become the main instruments of reform. Here we have the major paradox of the reform movement of the eighties: significant improvements in the quality of schooling are impossible without structural changes, but increased dependence on mass-administered tests at all levels has had the effect of strengthening existing structures and forms of control. The culprit is not educational assessment and testing *per se.* Rather, the argument I make here and in Chapter 8 is that the particular forms of testing in widest use for increasing accountability are rooted in a social science paradigm which takes as a given the necessity for centralized control.

Use of such tests are not the sole cause for the failures to restructure schools. *Re*-forming schools or any social institution is a complex business. It requires a commitment by national, state, and local, public officials, and professional educators to critically examine their own long standing practices and patterns of organizational control. It takes persistence and inordinate courage by leaders and governing bodies to dislodge entrenched, centralized bureaucratic power. If we know anything at all about politics and human behavior, it is that many endorse the need for change, but few risk challenging the many vested individual and institutional interests in maintaining business-as-usual. There are thousands of organizational entities, and tens of thousands of individuals within national and state governments, colleges and universities,

foundations, publishing companies, and central offices of local school districts whose power would be greatly diluted or lost if the current system of assessment were significantly altered.

The historically unparalleled growth in the use of mass testing as the chief instrument of school reform over the last several years has produced a counter-reaction as evidenced by increasing public criticism in mainstream journals and the popular national press questioning the credibility of these tests, and by a resurgence of interest in alternative forms of assessment. Two recent studies, the first conducted by the National Center for Fair and Open Testing (Medina & Neil, 1988) and the second by the National Commission on Testing and Public Policy (1990) document both the growth of and interest in the development of alternative forms of testing, and the resistance to use of current forms of testing by many mainline educators and citizen and professional groups. Skepticism of multiple choice tests, which for many years was largely confined to progressive critics and to academic traditionalists, is now voiced regularly in such places as the *Washington Post, New York Times, Wall Street Journal, Newsweek,* and even on prime time television documentaries.

The two reports cited above and a publication of the National Center on Effective Secondary Schools at the University of Wisconsin (1989) document in detail the deficiencies and problems with these tests. They show that the short-answer, closed-ended format precludes the assessment of higher-order thinking and mastery of complex material, that test items are frequently biased in subtle and not so subtle ways, and that dependence on these tests as the primary indicators of school quality and for making judgments about students abilities and achievements distorts schooling policies and practice in numerous ways.[5]

Though I (and all the writers included in this volume) would concur with most of these criticisms of the commonly used forms of educational tests, and that there is a need to develop alternatives, I do not focus here on critique nor on reviewing and examining proposed alternative forms of testing. Rather my purpose in this chapter is to raise questions about the theoretical foundations of the widely used forms of achievement testing, and to foreshadow the argument for a theory of testing and assessment, that is compatible with current interest in restructuring schools by dispersing power and shifting responsibility away from the center, towards local school districts and to the teachers and principals within individual schools.

Though there is critique in this volume, and discussions and exemplars of alternative forms of assessment, the book is primarily an effort to examine the theory and practice of educational assessment, and a modest step toward the development of a new paradigm. This book supports the view that fundamental changes in the way we think about education and the process of schooling must accompany the effort to rethink assessment theory and prac-

tices if we are to realize the aspiration of providing all the nation's children with schools which serve their best interests, the interests of the communities they live in, and the interests of the nation as a whole.

I must forewarn the reader that this book does not pretend to provide a fully articulated and coherent perspective on the theory and practice of educational assessment. The lack of unity and consistency of argument across chapters is, in part, a function of its history. Supported by a grant from the US Department of Education's Office of Educational Research and Improvement to the National Center on Secondary Education at the University of Wisconsin, I collected and edited a set of papers which were intended to provide some fresh perspectives on the testing and assessment question drawing upon work commissioned by the Center and from the existing assessment literature. This task was completed in 1988. In the course of this work, it became increasingly clear to me that some of the researchers whose writings I had collected and edited were pressing the limits of the familiar testing technology and moving in the direction of abandoning and replacing the measurement paradigm which has predominated for at least the last sixty years. Five chapters in this book are revised and edited versions of papers selected from that earlier collection, and three chapters (Chapters 1, 6, & 8) were written expressly for this volume. The first and last chapters are an effort to illuminate the arguments for a new assessment paradigm, arguments which I saw as largely submerged in the work of the writers of the other papers. In none of the chapters, except Chapter 5 by John Raven, and my two chapters, is there a self-conscious effort to articulate a case for a new science of testing and assessment. Although I make my case drawing freely from the work of others, from the writers of the other chapters, and from sources I cite in the endnotes of my two chapters, I alone must be held responsible for the way I have interpreted and used their work.

Foundational Assumptions of the Current Paradigm

I will state what I see as the four foundational assumptions of the paradigm which underlies virtually all standardized and most criterion-referenced tests. In so doing, I will also state four "counter assumptions" which are intended to foreshadow the argument for the development of a new testing and assessment paradigm.

Before proceeding I will clarify several commonly used terms:

Test Technology. Test technology refers to the structure of a test, the ground rules and conventions used for its construction, the procedures and protocols for scoring and summarizing results, and the matrix of practices required for everyday use.

The tests I refer to here are those generally composed of a relatively large array of short questions of "items." Each item includes a problem presentation—a sentence, paragraph, set of statements, a chart, graph, picture, or mathematical equation followed by a set of four or five possible responses, one of which is designated by the test-makers as the correct or best possible answer. The individual taking the test makes a selection and blackens a space provided, generally on an separate answer sheet which is subsequently machine scored. There is almost always a time limit for completing the test. Scores are usually computed by counting correct responses and subtracting this number from the number of incorrect responses. A variety of statistical operations is employed for summarizing test results so that they may be used for comparing scores of individual or groups. Some variations of this technology should be noted, which generally do not represent a significant change in a test's technology. A desktop computer or terminal may be used to present items to the test-taker and to tally responses in lieu of the printed test and answer sheet. Also, some tests may include open-ended test items, those which require a writing sample or solving a math problem. In scoring such items, responses are assigned a number by a person trained in the use of a set of scoring conventions. The scores are then treated in the same way as those derived from multiple choice items.

Standardized and Criterion-Referenced Tests. A distinction is commonly drawn between "standardized" (or norm-referenced) and "criterion-referenced" tests. Among the best known of the former are the California Achievement Tests, the Iowa Tests of Basic Skills, and the Standard Achievement Tests (or SAT). Criterion-referenced tests include virtually all National Assessment of Educational Progress (NAEP) tests and state-mandated "basic" or "essential" skills tests.

Standardized tests do not depend upon setting educational standards as is often assumed. The concept of standardization in this context refers to tests which are constructed in such a way that allows a standard score, grade equivalency, or percentile to be computed, thereby permitting comparison of an individual's score, percentile, or a group mean to those of another individual or group. Such comparisons are possible only if the test is "normed." What this requires is that during a test's development, it was administered to a sample of test-takers, and the distribution of their scores was compared statistically to a so-called "normal" distribution. The slope of such a distribution is bell-shaped, hence the commonly used term *bell curve*. A normal or bell curve does not appear naturally. To the contrary, test-makers attempt to compose test items so that there will be a suitable ratio of correct to incorrect responses. If too large a number of test-takers chooses the correct responses to sets of items, these items would be revised or abandoned even if there were unanimous consensus that the items tapped an educationally significant body

of knowledge or set of skills. The reason is that the items must "discriminate," that is, produce the proportion of correct to incorrect answers required by a "normal" distribution.[6] The technology of standardized tests, contrary to popular belief, do not warrant making *qualitative* statements about a person's (or group's) performance. The only claims which are warranted is how an individual's score or percentile (or group's mean or mean percentile) compares with others who have taken a version of the same test.

Though there are a number of recent efforts by the NAEP and several states to depart from the usual closed-ended format, the items in the vast majority of criterion-referenced tests are indistinguishable from those included in a standardized test. The major difference is that criterion-referenced tests are not normed. A panel of educators decides what percentage of correct responses constitutes passing or minimal competence. This score serves as the criterion for making judgments about an individual's or groups' competence or level of achievement. In practice, someone selects a score which sets the minimum number of items students at a particular grade level must answer correctly in order to be considered minimally competent in a given area—mathematics, reading, or whatever. Criterion-referenced tests (with some significant exceptions) also warrant only quantitative statements about how an individual's score or a group's mean (the group may be a single class, a school, a set of schools from a district or entire state or region) compares to the mean of another individual or group, or to an established criterion score.

It is important to note that in recent years, there have been efforts to develop so-called "performance-based" tests. The intent is to create assessments which avoid the multiple choice format and more closely approximate real tasks, such as conducting an experiment or writing a job application letter. While some of these efforts succeed in breaking the boundaries of the conventional testing paradigm, most do not depart significantly from the conventional standardized and criterion-referenced test technology. Rather than presenting four or five alternatives to choose from, a score is assigned to the test-takers' "free" responses (recorded on paper or computer) on the basis of previously-determined criteria. Aggregate scores are then treated in more or less the same way as those derived from multiple choice items. For all practical purposes most such assessments are rooted in the conventional psychometric paradigm.

Scientific Paradigms. Scientific endeavor in any area rests upon a set of *a priori* assumptions shared by persons who engage in that endeavor. With reference to testing, this means that those within the educational testing and evaluation community who design and construct educational tests, or who administer and interpret their meaning to others take for granted a set of beliefs, values, and practices (or "puzzle solutions"). It is the foundational assumptions and practices taken as normal within a particular community of

scientists which Thomas Kuhn, a well-known historian of science, calls a *paradigm* in a widely quoted book, *The Structure of Scientific Revolutions,* first published almost thirty years ago. A paradigm may be seen as what Michael Foucault calls a "regime of truth." A regime of truth in science is a set of practices and discourses taken as given in everyday scientific activity and which implicitly defines what are and are not considered legitimate scientific questions and methods.

What is significant to my argument here, and to the thesis of this entire volume is Kuhn's claim that paradigms or regimes of truth in science are transient and that the history of science is itself a history of paradigm breakdown and replacement. Paradigms are replaced because anomalies and problems appear that cannot be explained or be fruitfully addressed using the commonly accepted language, ground rules, or "puzzle solutions." Over time scientists develop new paradigms—that is different concepts, sets of "puzzle-solutions," and a constellation of beliefs and values[7] which appear to address the difficulties. It is these changes that constitute revolutions in scientific thinking and practice, and while they are infrequent, major transformations are to be expected sooner or later. In the meantime, normal science continues more or less undisturbed, as the old regime erodes and in time is replaced by a new one. Periods of transition and change, it should be added, are unsettling if not tumultuous because the new paradigm threatens existing interests and the institutional arrangements that hold the current regime of truth in place.

The scientific paradigm that undergirds standardized and virtually all criterion-referenced tests which has been in the process of breakdown for the last two decades has reached a critical stage. Standardized and criterion referenced tests, rooted in an anachronistic paradigm, are a major barrier to the renewal and restructuring of the nation's schools. As we enter the last decade of the twentieth century, it is becoming apparent, at least to those outside the testing and measurement establishment, that the assumptions which are intrinsic to the technology of standardized and most criterion-referenced tests are untenable. Out of the ashes of this paradigm, from the many varied and imperfect efforts underway to solve the practical problems of assessing educational achievement, is slowly emerging a new paradigm, one based on a set of foundational assumptions that are in sharp contrast to those that underlay the current paradigm.

The paradigm that is foundational to current forms of standardized and criterion-referenced tests I label the *psychometric* paradigm; the emerging one, a *contextual* paradigm. There is some risk in the use of these terms, as there is in any effort to classify and simplify complex ongoing human activities into categories. The distinction is helpful insofar as it helps to clarify the issues and distinguish significant differences in efforts to develop alternatives to the most commonly used forms of testing and assessment. The implication

that all tests and assessments may be classified in terms of two mutually exclusive categories, however, is potentially misleading and confusing because the distinction also may obscure significant differences within and similarities across categories. As Doug Archbald's and Fred Newmann's summaries of alternative forms of assessment show (see Chapter 7), some efforts appear to embody aspects of both paradigms.

It should also be underscored that the psychometric paradigm must not be considered as synonymous with *quantitative* methods, and the contextual paradigm with *qualitative* approaches. It is certainly true that psychometric assessments rely heavily on quantification and statistics, and contextual assessments more often than not employ qualitative methods. However, quantitative measurement and the use of statistics are not necessarily inconsistent with contextual approaches, and qualitative techniques are sometimes used in ways that ignore or bypass social context.

Assumption 1: Universality of Meaning. By universality I mean a view that there is or can be established a single consensual meaning about what standardized or criterion-referenced tests claim to measure which, in effect, transcends social context and history. For example, a standardized reading test purportedly indicates a person's ability to read in the real world, not just in the testing situation, and "ability to read," it is assumed, has a more or less universally understood and accepted meaning. Further, it is assumed that scores on a given reading test indicate individuals' level of reading ability—regardless of their or their families' history, culture, or race; regardless of gender, whether they live in Nome, Alaska or Newark, New Jersey; regardless of whether they have gone to a school with a first-class library or no library at all, and whether they reside in an affluent suburb or an area with high and chronic unemployment. The assumption of universality, in effect says, that a reading test score has essentially the same meaning for all individuals everywhere.

Within the discourse of psychometrics, postulated attributes or capacities of persons (their reading ability, or academic achievement in a particular area, for example) are called *constructs.* A standardized test of academic achievement presumably measures the *construct* of "academic achievement"; a criterion-referenced test of basic or essential skills measures the *construct* of "basic" or "essential skills." The term *construct* may sound strange and may perhaps be considered superfluous to non-specialists in the field of educational measurement. This term became commonplace in the field of mental measurement after its use in a seminal article by Lee Cronbach and P. E. Meehl titled "Construct Validity in Psychological Tests" (1955). According to Cronbach and Meehl a construct "is an intellectual device by means of which one construes events. . . . It is a means of organizing experience into

categories. . . . Construct validity, then, is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (pp. 281–82). The use of this term acknowledges an obvious but sometimes ignored fact that human attributes or capacities are not tangible, directly observable, or measurable. Thus, a reading test does not, indeed cannot measure reading directly. Rather, if the reading test does what it claims, it measures a construct the test-makers have labeled "reading" or "reading ability."

How do we know whether a test measures what it claims to measure, whether a test in fact measures authentic reading ability or genuine academic achievement? The response a traditional testing expert gives to the question of whether a standardized or criterion-referenced test measures what it purports to measure is that this determination depends upon the adequacy of the case a test-maker makes for the test's *construct validity.*

Establishing construct validity of a test requires getting things straight between (1) the world of human events and experience, (2) the construct label, and (3) the test. This entails establishing what Cronbach and Meehl refer to as a "nomological net," which is "a rigorous (though perhaps probabilistic) chain of inference" from an empirical body of knowledge and a logical analysis of the meaning of the construct. Almost thirty years later, Cronbach (1987) stressed that "the argument [for test validation] must link concepts, evidence, *social and political consequences and values"* [italics added]. Thus, in order to establish the validity of a test of academic achievement within the framework of the psychometric paradigm, for example, one would need to assume that the construct of "academic achievement" has a stable, universal meaning, or that unanimity on its meaning is both possible and desirable, *and* that it is possible to reach consensus on the desirability of the social and political consequences of the test's use.

The counter assumption: plural, and contradictory meanings. In Chapter 8, I examine in some detail the basic controversies and contradictions in contemporary America over education and the functions, purposes and practices of schooling. I demonstrate that the assumption that there can be a meaningful nationwide statewide, district-wide, or even schoolwide consensus on the goals of schooling and on what students should learn and how they should learn is untenable. I also argue that in a muticultural society which values difference, consensus is undesirable. While it is perhaps understandable that in the 1950s some would hold the view that consensus on basic educational beliefs and values is possible, from the vantage point of the 1990s this view is naive. The premise of what I call a contextual paradigm is that a plurality of meanings, and differences and contradiction in perspectives are inevitable in a multicultural world, where individuals, and groups have differing histo-

ries, divergent interests and concerns. There is no, nor can there be universal consensus on what constitutes "ability to read", the meaning of "academic competence" or "authentic achievement" in general terms or within specific academic fields. Experts and nonexperts alike hold plural and often fundamentally contradictory beliefs and values over the meaning of all educational terms. Validating educational tests based on psychometric canons represents a quest for certainty and consensus where certainty is impossible, and agreement is unlikely unless differences are suppressed and consensus is overtly or covertly imposed. Further, as I argue in Chapter 8, the entire concept of "construct validity" on which the scientific credibility of all such tests is rooted is itself internally contradictory and untenable. I also argue that it is possible to develop a system of educational assessment that takes plurality of perspectives and differences in values and beliefs as givens, and treats these differences as assets, rather than obstructions to be overcome.

Assumption 2: The Separability of Ends and Means, and the Moral Neutrality of Technique. Discourse and practice within psychometrics assume that tests, if constructed and interpreted according to accepted standards, are *scientific instruments,* which are value-neutral and capable of being judged solely on their scientific merits. The argument often made in defense of the technology of standardized and criterion-referenced tests is that their development represents an advance over prescientific and subjective forms of assessment, such as grades and teacher-made tests, which intermingle factual observations with the personal, subjective dispositions of the teacher. The basis of the argument for the moral neutrality of tests is that ends and means are separable. Questions such as what constitutes the good or just society, or what is the nature of a good or proper education, because they require moral choices, are not resolvable, and hence lie outside the domain of true science. The choice of means or the best route to a prescribed goal or end, however, is seen as an empirical matter, not a moral question, and hence may be decided scientifically. From this perspective, the job of the assessment expert parallels that of the engineer whose expertise is in the application of the science, not in making judgments about desirability or worth of the enterprise. The role testing expert is limited to dealing with technical or procedural questions within the moral framework set by society.

There are two closely connected assumptions here. First, facts and values, (or what is and what ought to be) are distinct and separable, or they are sufficiently distinct to make possible a non-normative science of educational measurement. What follows from this assumption is that testing experts can make technical decisions without making value judgments. Second, the assessment scientist is best equipped to make judgments about means, that is to develop the ways of assessing educational outcomes and how these are to be

properly used and interpreted. Just as it would be the height of irrationality to turn over to a non-engineer the responsibility for designing a bridge or a rocket's guidance system, so too would it be irrational to replace scientific techniques of measurement and the rules of evidence with the opinions and subjective preferences of the non-scientist.

Counterassumption: The Inseparability of Means and Ends. The impossibility of sustaining this fact-value distinction is argued in Chapter 8. In brief, the argument for whether a test measures what it claims to measure rests on the case made for its construct validity, which is considered a technical matter. However, establishing construct validity clearly is not merely a matter of empirics, getting the facts straight and interpreting them according to established rules of evidence. Judgments about an educational test's validity invariably require choices among contradictory values, beliefs and schooling practices (Cherryholmes, 1989; Messick, 1989). In the real world of schooling, separating means and ends is not possible. All assessment procedures have the power to directly or indirectly shape social relationships—how students, teachers, and administrators within a setting interact with one another, what they will or will not say or do in particular situations. Moral questions arise in all social relationships, which can either be resolved by the use of direct or indirect power where the values, beliefs and ideologies of those with the ability to impose their will prevail, or by a process wherein conflicts are acknowledged, and mediated recognizing both differences and commonalties in interests and values. If judgments about assessment procedures and testing are left to experts, then they assume the responsibility for resolving differences over basic moral questions which in a democratic society should be settled by ordinary citizens and/or their democratically elected representatives.

Assumption 3: The Separability of Cognitive from Affective Learning. The psychometric paradigm separates the assessment of learning outcomes and processes into distinct and mutually exclusive categories, separating cognition or academic learning from affect, interests, or attitudes. Sometimes a third category, psychomotor outcomes, is added to the set. Tests of academic achievement, and IQ tests fall into the first category; tests or inventories which solicit a person's beliefs, attitudes, or interests fall into the second; and tests of a person's capacity to perform a hands-on or vocational task (such as auto mechanics or typing) fall into the third. This three-way classification of human learning or capacities divides head, heart, and hand, that is, it separately assesses those areas of human learning and development related to the realm of the intellect, those related to the realm of feelings and values, and those which require manual or physical dexterity. A test of basic educational skills, for instance, purportedly will tell us how well a person knows a par-

ticular body of scientific facts or performs a particular set of math tasks. If we want to know the person's interest in math, or whether she is curious about science, we would need to administer a different instrument.

These distinctions are deeply ingrained and institutionalized within the psychometric sciences and are rarely given a second thought. They are legitimated by the Benjamin Bloom's (1956) *Taxonomy of Educational Objectives* which remains the most widely accepted system of classification in the field of education. The distinctions are treated as virtually self-evident and used widely in the everyday discourse of teachers and administrators.

The Counterassumption: The Inseparability of Cognitive, Affective and Conative Learning. As John Raven argues in Chapter 5 and elsewhere (1989), this classification distorts and obstructs efforts to assess significant educational achievements. Raven points out that not only are cognitive and affective outcomes treated as separate categories, but that what he calls the *conative* aspects of human behavior, those concerned with determination, persistence, and will, are inappropriately subsumed under "affective". A person, he points out, can enjoy doing something without being determined to see it through, and he or she can hate doing something, but still be determined to do it. He makes his argument focusing on the "ability to take initiative" which is generally acknowledged as a desirable educational outcome. He argues that taking initiative (which would be categorized as an "affective" outcome in the Bloom Taxonomy) is inseparable from intellectual or cognitive functioning, and from action:

> To take initiative successfully, people must be self-motivated. Self-starting people must be persistent and devote a great deal of time, thought, and effort to the activity. . . . The crucial point to be emphasized in attempting to clarify the nature of competence is that no one does any of these things unless he or she cares about the activity being undertaken. What a person values is therefore central. . . . What follows from this is that it is necessary to know an individual's values, interests, and preoccupations in order to assess his or her competencies. Important abilities demand time, energy, and effort. As a result, *people only display them when they are undertaking activities which are important to them* [Italics added] (Chapter 5, p. 89).

Raven goes on to argue that, if this analysis is correct, it does not make sense to attempt to assess separately cognitive, affective, and conative components of an activity. Affective and conative components are integral to the ability to cognize. "Not only do the three components interpenetrate if the

behavior in question—the taking of initiative—is to be successful, these components must be in balance. Determination exercised in the absence of understanding, and the converse, are unlikely to make for a competent performance."

The proposition that cognitive, affective, and conative aspects of human learning and development are inseparable is in sharp conflict with several accepted canons of traditional psychometry. It runs counter to the widespread practice of using one set of scales to assess values, attitudes, and beliefs and other independent scales to assess knowledge, skills, abilities, or competencies. Raven makes the intriguing suggestion that, if we are to assess such qualities as initiative, instead of trying to develop separate assessments which are difficult if not impossible to interpret, we need to develop indices which unify the cognitive, affective, and conative. He argues that development of all human capacities is highly contingent upon the opportunity structure (the social context), as well as on the learner's will, interest, and knowledge. In Chapter 5, Raven shows that it is technically possible to develop value-based indices that can do more justice both to the complexities of human qualities and capacities, and to how they are fostered and developed.

Assumption 4: The Need for Control from the Center. Testing and assessment procedures are forms of surveillance whose use is the superimposition of a power relationship. Criterion-referenced and standardized tests are sometimes criticized because they shape the school's curriculum and pedagogy. But the *raison d'etre* of *all* evaluative procedures in education, not only standardized and criterion-referenced tests, is to shape the educational process by exerting control over educational administrators, teachers, and/or students. Assessment procedures are inherently political, not only because whoever controls the assessment process shapes the curriculum pedagogy and ultimately the students' life chances, but also because particular forms of assessment promote *particular forms* of social control within the organization, while suppressing others.

My contention here is not that particular forms of organizational management and control inevitably follows from particular forms of assessment. Assessments are only one of many complex factors shaping how schools and school systems are governed. Rather, the claim is that the particular *form* of assessment is a key factor in producing particular forms of social control throughout the organization. In other words, the technology used in the assessment process, will encourage particular forms of management and human relationships within the organization while suppressing others.

The technology of mass administered standardized and criterion-referenced tests produces social relationships and management structures which are largely suited to exercising control from the center, that is, from the

central office by local or state educational authorities. Such tests provide virtually no information about what students are capable of doing or where they may need help. These tests produce relative rankings but little substantive information about what students know or can do which is useful to teachers, parents, prospective employers, or to students themselves for making programmatic or individual decisions. The psychometric technology only enables us to classify and rank order students (or teachers), and to constitute individuals as a "case," that is, as belonging to a class or category which possesses a particular set of objective characteristics (*e.g.* high, average, or low achievers, at risk students, etc.).

These tests are used primarily to facilitate what Michael Foucault (1979) calls *le regard,* (the gaze) or visibility to authority. Standardized tests and most criterion-referenced tests are particularly powerful forms of social control because they objectify the subject by reducing all human characteristics to a single number, thereby facilitating comparative rankings, and placing individuals into categories. These ranks and categories allow central office administrators to monitor and manage large numbers of students and teachers. Control exercised by such tests is not direct or overt, their effectiveness, rather, resides in the fact that those who are evaluated internalize or take into themselves the ranks and labels placed on them because these are presumably made by neutral, scientific instruments. Though individuals can and sometimes do resist these valuations of their capacities or achievements, the vast majority succumb because standardized and criterion-referenced test scores are the only educational currency accepted as scientific by the wider society.

Counter assumption: Assessment for Democratic Management Requires Dispersed Control. What should be emphasized for making a case for a contextual paradigm is that intrinsic to the use of standardized and criterion-referenced tests is a form of surveillance and exercise of power which is *unidirectional.* Central office administrators exert power over the everyday life and fate of students, teachers, and parents, who have no way of changing the system of assessment which controls them other than passive resistance or active subversion. While all forms of assessment, including any newer forms we might invent, represent a form of surveillance and constitute a means of control and an exercise of power, it is possible to alter the unidirectionality of control within the assessment system. That is to *re*-form the system of assessment in such a way that it disperses power, vesting it not only in administrative hands but also in the hands of teachers, students, parents and citizens of the community a particular school serves. If we are to have a system of public education supported by public funds, and governed by democratically elected bodies, then oversight by these bodies is essential. Some form of systematic assessment for holding educational institutions and the professionals who work in them accountable for their performance is necessary to monitor

expenditures, to insure that that professionals meet their responsibilities, do not exceed their authority, or violate the public trust or students' and parents' rights. But the exercise of power via the assessment process by central administrative authorities at the national, state, or district levels becomes coercive and oppressive without countervailing power over the assessment process exercised by teachers, parents, and students. From both experience and social scientific evidence, it is clear that good schools require a strong measure of autonomy by teachers, other school-level professionals, and participation by the local school-community. Without significant control over the assessment process at the school-level, teacher empowerment and school-based management is an illusion.

In Chapter 6 Elizabeth Adams and Tyrrell Burgess show that a system of assessment can be devised which vests significant power in the hands of central authorities, *and* in the hands of school-level professionals, parents, students, and the local school-community. Drawing upon their experience in the United Kingdom, they show how institutional arrangements and processes can be developed enabling the authorities at all levels to oversee the quality of schooling, to effect system-wide educational policies, and at the same time setting limits on the power of these authorities to trespass on the prerogatives of teachers, school heads, and students. In Chapter 8, I make an effort to extend their argument, and to show how such an effort could be adapted to fit the American experience.

Overview of the Book

A summary of the remaining chapters follows.

Chapter Two. Assessing Mathematics Competence and Achievement, by Thomas A. Romberg, defines and clarifies a conception of authentic achievement in mathematics and examines the validity of the commonly used instruments for assessing mathematics achievement. He concludes with a set of propositions to guide the development of new approaches to mathematics assessment and with an argument for the need to develop new approaches.

Chapter Three. The Assessment of Discourse in Social Studies, by Fred M. Newmann, suggests that a major aspect of social studies assessment should focus on the oral and written discourse that students produce on social topics. He addresses the questions of what discourse is and why it is an important indicator of student achievement in history and social studies. He concludes with his view of what experience and research suggest about the feasibility of this approach to assessment and with a discussion of how the assessment of discourse could provide meaningful and useful comparative indicators of student performance.