DOCUMENT RESUME

ED 363 630 TM 020 632

AUTHOR Latting, John
TITLE Assessment in Education: A Search for Clarity in the Growing Debate.
INSTITUTION National Center for Research in Vocational Education, Berkeley, CA.
SPONS AGENCY Office of Vocational and Adult Education (ED), Washington, DC.
PUB DATE Dec 92
CONTRACT V051A80004-92A
NOTE 43p.
AVAILABLE FROM National Center for Research in Vocational Education, Materials Distribution Service, Western Illinois University, 46 Horrabin Hall, Macomb, IL 61455 (MDS-254, $2.75).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Educational Assessment; Educational Change; Educational Objectives; Educational Policy; Educational Trends; Elementary Secondary Education; Policy Formation; *Psychometrics; Scoring; *Student Evaluation; Test Use; *Vocational Education
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

Current assessments in education are described, with particular attention to assessments intended for vocational education in order to help teachers, administrators, and policy makers make sense of the variety of efforts now taking place. At the moment, there is a great deal of controversy surrounding educational assessment. Central questions are whether we should test for achievement in education and whether or not tests work. There are many criticisms of traditional testing, and many calls for new approaches. Broadly speaking, assessment falls into three camps today, that of the psychometric tradition, performance assessment, and alternative assessment. Performance assessment represents a way to deal with the shortcomings of traditional assessment without necessarily changing the goals of testing and the variety of possible uses. Performance assessments are meant to be direct assessments of behaviors, and they are subjectively scored. Alternative assessment is linked to assumptions about learning and based on the idea that assessment systems should be multidimensional and should facilitate learning. Advantages and drawbacks of each approach are explored. Four figures illustrate the discussion. (Contains 36 references.) (SLD)

NCRVE

National Center for Research in Vocational Education

University of California, Berkeley

ASSESSMENT IN EDUCATION: A SEARCH FOR CLARITY IN THE GROWING DEBATE

Supported by
the Office of Vocational and Adult Education,
U.S. Department of Education

This publication is available from the:

National Center for Research in Vocational Education
Materials Distribution Service
Western Illinois University
46 Horrabin Hall
Macomb, IL 61455

800-637-7652 (Toll Free)

ASSESSMENT IN EDUCATION: A SEARCH FOR CLARITY IN THE GROWING DEBATE

John Latting

Graduate School of Education
University of California at Berkeley

National Center for Research in Vocational Education
University of California at Berkeley
1995 University Avenue, Suite 375
Berkeley, CA 94704

Supported by
The Office of Vocational and Adult Education,
U.S. Department of Education

December, 1992

MDS-254

FUNDING INFORMATION

Project Title:	National Center for Research in Vocational Education
Grant Number:	V051A80004-92A
Act under which Funds Administered:	Carl D. Perkins Vocational Education Act P. L. 98-524
Source of Grant:	Office of Vocational and Adult Education U.S. Department of Education Washington, DC 20202
Grantee:	The Regents of the University of California National Center for Research in Vocational Education 1995 University Avenue, Suite 375 Berkeley, CA 94704
Director:	Charles S. Benson
Percent of Total Grant Financed by Federal Money:	100%
Dollar Amount of Federal Funds for Grant:	$5,775,376

ACKNOWLEDGMENTS

The author's thanks go in particular to Professor W. Norton Grubb, whose advice and directions for sources of information were valuable throughout the process of preparing this document. In addition, Professor Grubb generously offered his expertise by preparing the following preface, which places this work on assessment in context.

PREFACE

This monograph began, in part, as an effort to understand the developments in educational assessment. As school reforms have proliferated in bewildering variety over the past decade, assessment has gone through its own search for alternatives. The realization that new forms of teaching and learning could all too easily be stifled by conventional testing has been part of the movement for different forms of assessment, though a much longer history of criticism has played its part, too. The result has been enormous energy devoted to revising conventional approaches to educational assessment—wonderful for the diversity of efforts but also confusing to those trying to understand what forms of assessment might be most appropriate for their classrooms, their schools, or their reforms. This monograph provides guidance to those trying to understand recent developments.

Within vocational education, too, this has been a period of experimentation and reform, driven in part by federal legislation and in part by the pressures to come up with new forms of vocational education as economic and demographic conditions have changed. The "movements" to integrate vocational and academic education and to develop Tech Prep programs linking secondary schools and postsecondary programs are representative of these reforms. Especially in an era of accountability, innovators are constantly under pressure to justify the success of their reforms, to show improved outcomes, and to prove to legislators that their funds are well spent. But—as in reforms within the rest of education—it has been unclear how to measure outcomes. Certainly conventional academic tests—the Scholastic Aptitude Test (SAT), for example, or the subject tests of the National Assessment of Educational Progress (NAEP)—are inappropriate if the goal is to develop hybrid programs, combining both "occupational" and "academic" content, developing more active teaching methods, and using occupational issues to contextualize learning. But it is not yet clear what the alternatives might be.

At the same time, the 1990 Amendments to the Carl D. Perkins Vocational Education Act requires that vocational programs develop performance measures, including those measuring achievement. This requirement gives additional urgency to the need to develop appropriate assessments because of the possibility that performance measures and standards will be used in the future to allocate funds, identify poorly performing programs, or reward outstanding programs. Similarly, discussion of the need to certify the skills of

individuals in vocational programs raises the possibility that vocational assessments will have powerful consequences in the future as they become screening mechanisms that determine which students find the best employment. In several ways, then, these developments portend the development of "high-stakes" assessment in vocational education, with powerful consequences for both programs and students.

In this setting, there has been little guidance for those searching for alternatives to conventional academic tests. In part, then, this monograph describes current developments in assessment generally and in assessments intended for vocational education in order to help teachers, administrators, and policymakers make sense of the variety of efforts now taking place. Unfortunately—as is often the case in periods of reform—there are not yet clear results. The new assessments described in this monograph are for the most part still far from completion. None of them is going to be able to serve all the purposes that tests now serve. This makes it all the more important, however, for educators to understand the nature of current developments so that they can distinguish those approaches to assessment that are likely to be useful for their specific purposes from others that are unlikely to be of much help.

W. Norton Grubb

TABLE OF CONTENTS

INTRODUCTION

There is at the moment a great deal of controversy surrounding educational assessment. Administrators, teachers, and even policymakers increasingly have opinions about how we should measure the outcomes of education—an arena that has traditionally been left to specialists in educational measurement. As one expects in a mass debate, there is a broad spectrum of opinion on the issues surrounding assessment. For some, testing "has been distorting what is taught into pellets which are the intellectual equivalent of rabbit food" (Mitchell, 1992b, p. vii); or similarly, we have managed over the years to subject students to "testing for the TV generation," which is "superficial and passive."[1] Others comment that it is "*impossible* to be impressed by the lack of objectivity and lack of scientific rigor" of many expressing opinions in the debate (Mehrens, 1992, p. 1) and that there seems to be "a disdain for professional standards" among some of the more vociferous participants (Williams, Phillips, & Yen, 1991, p. 1). It is inevitable that as discussion on an important matter moves from one group of like-minded professionals to many groups with varying backgrounds, interests, and expertise, disagreement and confusion emerges. The issue of assessment is in just such a state of transition, and the effects are quite plain for even the casual observer.

To define the terms around which the debate centers, an *assessment* is any attempt to determine the importance, size, or value of the outcomes of some process—in this case, of education (i.e., schooling). *Tests* are *objective* and *standardized* methods for estimating the nature of educational outcomes based on a sample of those outcomes (Office of Technology Assessment, 1992, p. 115). In other words, tests are a subset of assessments.

One of the questions in the assessment debate that, while quite controversial, will not concern us here has to do with the object of testing: Exactly what cognitive domain should we assess? The two most likely possibilities are educational achievement and aptitude. In the latter case an assessment would seek to determine the ability of students to learn or their general level of intelligence. Questions about the appropriateness of aptitude testing have been hotly contested for almost seventy years now, as Cronbach (1975) points out. The relative importance of environment and heredity with respect to educational success, as well as other forms of behavior, is an illustration of the controversy that has

[1] Statement attributed to Linda Darling-Hammond.

persisted for so long. This debate, however, is separate from the debate over the assessment of student learning—that is, of educational achievement. Therefore, to simplify what is already a complicated debate, I will restrict discussion to assessments of student achievement and ignore assessments that attempt to determine the capacity for such achievement to take place.

The questions that *will* concern us can be stated as follows: First, should we test at all for achievement in education? Do tests work? That is, are tests adequate assessments of educational outcomes? Answers to these questions depend, of course, on how we go about testing, so the question of just what kinds of tests should be used is also important. Second, if we are satisfied with the "test-like" assessments that have been or will be developed, what should be done with the results? What decisions about children should be made based on the information provided by educational assessments? These questions, I will argue, have driven the current debate on assessment.

THE CONTEXT OF THE CONTEMPORARY SCENE

Achievement tests, in a general sense, have existed for many years. Probably for as long as the teacher and student relationship has existed, there has been a need to assess the capabilities of students—perhaps in the form of a written or oral recitation. The most common form of testing continues to be that undertaken by the classroom teacher as part of his or her responsibility of assigning grades to individual students. As we shall see, however, the kind of assessment on which the debate centers goes well beyond the schoolroom: The controversy surrounds tests used for certification, placement, and selection as well as for systemwide accountability. These sorts of assessments have existed on a large scale in America for roughly sixty years.

The earliest paradigm in American testing treated the measurement of both aptitude and achievement as an exact science. The aim of "psychometricians" was to assign quantities to a whole range of mental properties (Mitchell, 1992b, p. 22). Ranking students on a ratio scale was the point of testing; and the multiple-choice format, which

became the hallmark of tests constructed in the psychometric tradition, allowed this ranking to take place cheaply and efficiently.[2]

This virtue of efficiency, as well as the prestige of the community of highly trained measurement specialists working in assessment, certainly has much to do with the hegemony that has been achieved by testmakers in the psychometric tradition. Today, such tests are far and away the most common given to schoolchildren in the United States: Last year, it was estimated that 127,000,000 multiple-choice tests were taken by the roughly 40,000,000 American primary and secondary students (Mitchell, 1992b, p. 4). At this rate, each student will have taken, on average, over thirty such exams by the end of the secondary school years. It has recently been estimated based on survey research that over seventy-five percent of public school districts use commercially available tests annually to assess students in grades K-9 ("Testing," 1992).

While alternatives to this tradition of testing have been discussed for some time,[3] not until the late 1980s did any self-conscious movement exist to promote alternatives. It has been argued that it was at a 1988 conference sponsored by California's Department of Education that the debate seen so clearly now began (Mitchell, 1992b, p. 175). The conference labeled itself "Beyond the Bubble" and seemed to ignite the storm of activity, and the rush of articles on assessment, that continues.

Although split, the reaction to the psychometric tradition has been united in the desire to move "beyond the bubble" of multiple-choice testing. The criticisms of the traditional methods have been varied: Proponents of alternatives to multiple-choice tests have complained, for example, that such tests are culturally biased, do not test what really interests us as educators, and have deleterious instructional and learning effects (Mehrens, 1992, pp. 2-4). There has been a call, therefore, for new approaches to assessment. On the one hand, critics have emphasized the supposed failure of tests to measure accurately the kind of knowledge that is of most use to society. The camp that has emerged following

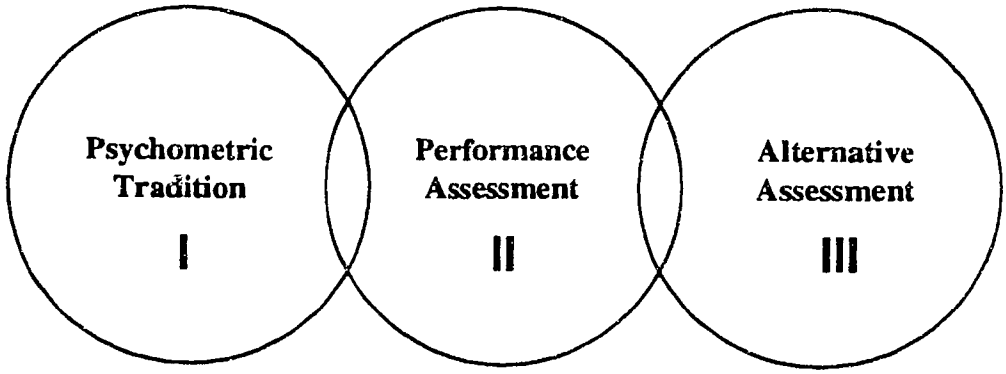
---

[2] See Nunnally and Wilson (1975) for a discussion of measurement scales. Ratio scales are the most elaborate in measurement: the rank order of objects is known, as is the "distance" between objects, all of which are positioned relative to a rational zero. Ordinal scales place objects in order only, and nominal scales simply label objects without reference to magnitude. Only a test measuring students on a ratio scale, therefore, allows the full range of mathematical and statistical operations with test scores.

[3] Note Quinto and McKenna's (1977) book *Alternatives to Standardized Testing*. In addition, Fitzpatrick and Morrison spoke of "performance tests" in 1971.

this line of argument I will refer to as advocates of "performance assessment"—an overly broad and possibly unsatisfactory label. The second form of collective criticism has emphasized the instructional and learning effects of testing. Such advocates will be known in this discussion as advocates of "alternative assessment." The map of the current debate in assessment, therefore, at least as I have conceived it, consists of three fairly self-conscious camps (see Figure 1).

Figure 1
The Three Camps of Educational Assessment



By explaining the rhetoric of each camp, as well as the assessments envisioned by each, I will clarify what has become a confusing debate.

THE CHARACTERISTICS OF ASSESSMENT

Before proceeding to a discussion of the positions taken and assessments envisioned by each camp, I need to make explicit some of the basic assessment vocabulary. Specifically, I will address (1) the possible uses of achievement tests, (2) the features of assessments in general, and (3) some of the terms often used to evaluate particular tests.

The Uses of Educational Assessment

There are three basic categories of the functions of educational assessments. The first and most common function of assessment is to contribute to classroom instruction. An assessment used for this purpose should provide accurate feedback to teachers about individual students and indicate how teachers can better educate each student. Informal assessments such as any particular teacher's weekly spelling test should be included in this category. Second, assessments can serve a system-monitoring function. As funds for public education have shrunk over the past several years, it has become increasingly desirable to monitor systemwide educational outcomes and attempt to hold programs—and even individuals—accountable for net loss or gain. And third, assessment can serve selection, placement, or certification functions. This kind of application of assessment results is probably the first of the three to come to mind: All Americans go through the ritual of testing in the process of allocating educational or vocational resources to selected individuals. I have included Figure 2 to illustrate the specific examples of each of the three functions of assessment. Given this wide range of specific functions for which assessments are increasingly used, it should come as no surprise that the design and selection of assessment instruments is hardly a trivial matter. Later in the discussion, I will turn to how the assessment community is attempting to meet this list of desired applications.

Figure 2

The Functions of Educational Assessment

Functions	Examples
1. Classroom instructional guidance Used to monitor and provide feedback about the progress of each student and to inform teaching decisions about *individuals* on a day-to-day basis	• Diagnose each student's strengths and weaknesses • Monitor the effects of a lesson or unit of study • Monitor mastery and understanding of new material • Motivate and organize students' study time • Adapt curriculum to progress as indicated by tests • Monitor progress toward curricular goals • Plan lessons that build on students' level of current understanding • Assign students to learning groups (e.g., reading groups)
2. System monitoring Used for monitoring and making administrative decisions about aggregated *groups* of students (e.g., a school, instructional programs, curricula, district)	• Report to parents and school board about a school or district's performance • Make decisions about instructional programs and curriculum changes • Evaluate Chapter I programs • Evaluate experimental or innovative programs • Allocate funds • Evaluate teacher performance or school effectiveness • Provide general information about performance of the overall educational system
3. Selection, placement, and certification of students ("gatekeeping") Used to allocate educational resources and opportunities among *individuals*	Selection • Admission to college or private schools Placement • Place students in remedial programs (e.g., Chapter I) • Place students in gifted and talented programs Certification • Certify minimum competency for receipt of high school diploma • Certify mastery of a course of study (e.g., Advanced Placement examinations) • Make decisions about grade promotion

Source: Office of Technology Assessment (1992)

The Features of Educational Assessments

An assessment can be thought of, for heuristic purposes, as a complicated organism. What makes any particular test distinctive depends on its constituent parts, and there are enough layers of features that a great many combinations are possible. There are, therefore, hundreds of possible kinds of assessments—each type different from all others in at least one important respect.

We can also think of an assessment in more formal terms—for example, as an equation containing several variables. As each variable takes one of at least two possible values, the assessment becomes more distinctive. In Figure 3, I have illustrated some significant variables that comprise any single example of an assessment of student achievement. Educational assessments are, of course, a small subset of the enormous set of possible psychological assessments; because we are interested here in educational assessment, however, I have included some variables peculiar to educational assessment and might have sacrificed detail with respect to other kinds of psychological measurement.

In what follows, I will briefly clarify the terms used in Figure 3, concentrating on the more obscure features. Note that the organizational and technical features of assessments are not obvious based on a review of the test protocol alone. In other words, assessments that appear similar might in fact be radically different in several crucial details.

Figure 3
The Features of Educational Assessment

Organizational Features

Purpose	Type	Unit	Population	Term
aptitude	formal	individual	individual	administered
achievement	informal	group	classroom	cumulative
interest		self	school	
			district	
			state	
			nation	

Figure 3 (cont.)

Physical Features

Instrument	Product
item response[1]	paper
other response[2]	project
checklist	presentation
interview	portfolio

Technical Features

Mode	Link	Tactic	Reference	Criterion	Scoring
power test	integrated	direct	criterion	subjective	external
timed test	independent	indirect	norm	objective	internal
					self

[1] For example, multiple-choice, matching, or true/false.

[2] For example, essay, short answer, completion, experiment, or simulation protocol.

- *Purpose*

  This refers, of course, to the function the assessment is intended to serve. As has been stated earlier, this paper is concerned only with assessments of educational achievement, so I need not go into other purposes. The reader should simply note that several other applications for assessments exist, even within the field of education.

- *Type*

  Divides assessments into two fundamental categories. A "standardized" assessment is one in which subjects have identical instructions, materials, practice items, and time with which to work (Office of Technology Assessment, 1992, p. 173). Standardized or formal assessments are properly referred to as "tests." Most, but certainly not all, educational achievement assessments are standardized and are therefore achievement tests. There is, however, a growing movement to develop informal or nonstandardized assessments.

- *Unit*

  Indicates the basic unit of measurement. Nearly all assessments in education use individuals as subjects; here, too, however, a movement exists to assess groups of students as an alternative.

- *Population*

  The population to be measured. In some cases (e.g., state and national assessments) only a sample of the population is measured.

- *Term*

  The duration of the assessment. Most assessments are administered on a single occasion, but newer assessments measure students over time. Portfolios are the best example of an instrument used to assess cumulatively.

- *Instrument*

  The type of stimulus used to illicit student response. Psychometric tests rely on response from a relatively large number of items—multiple-choice questions in particular. Alternative forms of assessment favor other, more "authentic" kinds of instruments (see the third table in Figure 3, for examples).

- *Product*

  The student-generated material that is assessed. These can range from circles around the "most correct response" to portfolios and presentations.

- *Mode*

  Assessments with a fixed and presumably limiting time constraint are, obviously, "timed." Tests intended to measure the achievement of students without regard to the speed with which answers can be generated are "power" tests.

- *Link*

  Refers to the relationship of the assessment to the curriculum. Integrated assessments measure material explicitly covered "in class," so to speak. Independent assessments measure information only indirectly related to any particular school curriculum. The SAT is a fine example of an "independent" test. Surprisingly, some achievement tests are not integrated with course curricula.

- *Tactic*

  The manner in which information about an achievement domain is measured. Direct assessments use a realistic sample of the domain to make inferences about the "true" domain. A direct assessment of writing, therefore, would use a writing sample as an indicator of general writing achievement or ability. Indirect assessments use related skills or subskills as indicators of true achievement. Tests that use student multiple-choice responses regarding the quality of preset sentences or paragraphs indirectly assess writing achievement.

- *Reference*

  The standard against which individual student scores are calculated. "Norm-referenced" tests compute scores relative to the mean and standard deviation of raw scores of other students in the sample. Because the scores of norm-referenced tests represent primarily a ranking of student performance, they do not generally offer good information about student content knowledge. However, these scores are correlated with content knowledge and allow test constructors to avoid the difficulties involved with clearly specifying content domain. "Criterion-referenced" tests compute scores relative to specified educational goals or objectives. These assessments focus on "what test takers can do and what they know, not how they compare to others" (Office of Technology Assessment, 1992, pp. 169-170).

- *Criterion*

  Indicates the nature of the scoring decisions made. Objective scoring is not possible for all types of assessment instruments and is characterized by an "excluded middle" in scoring decisions—that is, responses are either correct or incorrect, not partially correct. Such tests have the advantage of allowing machine scoring. Subjective scoring requires deliberation and might involve a complicated set of criteria in order to arrive at a final score. The assessment of student essays, for example, seems to be a candidate for subjective scoring, given the wide range of acceptable responses to any particular prompt.

- *Scoring*

  Indicates where and by whom the actual work of scoring assessments is done. Internal scoring is undertaken by the teacher administering the test; external scoring, either subjective or objective, takes place apart from the teacher involved.

Reliability and Validity

A final word by way of introduction to the assessment debate has to do with the two primary technical criteria for evaluating the worth of an assessment: (1) reliability and (2) validity. These two terms are thrown around to a surprising extent in the debate, both in defense of particular assessment programs and in criticism of competing programs. It is worthwhile to understand, even if at a superficial level, these often misunderstood concepts.

Reliability is a rather straightforward property of an assessment system: It refers to the consistency of test scores. There are three dimensions of consistency, hence three types of reliability (National Research Council, 1991, p. 119). First, there is *test-retest* reliability or "stability." In this case, a student taking a test on two separate occasions should achieve similar, if not identical, scores. Second, there is *parallel-form* reliability or consistency despite variation in test items. A test that is reliable in this sense should assign similar scores to a student regardless of which particular questions were asked of him or her. And third, there is *inter-rater* reliability. Scores given to a student based on performance on a single exam should not depend on which person scores or grades the test. Reliability, in all three of its meanings, is a clear prerequisite for test results that can be trusted. Unreliable tests make it difficult for any of the three functions of assessments mentioned earlier to be properly executed.

Validity is a related test property. It refers to the extent to which inferences from a test score are appropriate. It is not the test itself, by the way, that is "valid," but instead it is the inference about the meaning of a test score (Williams et al., 1991, p. 9). There are three ways in which a test score can be valid or provide useful decision-oriented information. First, a test score should be generalizable from the sample measured to the entire domain of knowledge. A test, in other words, needs to adequately represent the whole range of information it is intended to measure. This is called *content* validity. Note that the scores of achievement tests, in particular, require content validity (Nunnally & Wilson, 1975). An achievement test that either failed to reliably translate student knowledge of a subject area into a test score or failed to assess all parts a subject could not be relied upon as an indicator of student achievement. Second, some tests are intended to provide information that can be generalized to other kinds of tasks. A test of vocational competency, for example, is intended to predict how well one will do on the job. How

well a test accomplishes this function has to do with *predictive* or *criterion-related* validity. And third, there should be evidence that a test actually measures the traits or skills that it attempts to measure. It is charged, for example, that some tests of mathematics assess test-taking skills rather than mathematical knowledge. The kind of evidence described in this third case is *construct-related* validity (Office of Technology Assessment, 1992, pp. 177-178). Note that reliability is a prerequisite for all three types of validity. An unreliable score clearly makes inferences based on that score questionable.

"Reliability" and "validity" are terms that surface continually in the assessment debate. These qualities become particularly significant when important decisions are to be made based on test scores. Whether test scores will be used to allocate resources to individuals or to groups of students (e.g., schools and their instructional programs), tests need to serve as trustworthy and appropriate measures of whatever it is they claim to measure.

At this point, the reader should be familiar with the three most important functions of assessment, the various features that assessments possess, and the two major criteria for evaluating the technical quality of an assessment. I now turn to the three camps that exist in the assessment debate and the characteristics of the assessments favored in each.

THE THREE CAMPS OF THE ASSESSMENT DEBATE

I argued earlier that there have been two reactions to the psychometric tradition in educational assessment. One reaction seems to emphasize the effects of this kind of assessment on instruction and learning. The group of advocates making these kinds of criticisms approach assessment, for reasons I will soon make clear, in such a way that I have labeled it "alternative assessment." The second reaction seems to emphasize the appropriateness of traditional assessment for some purposes, and I have labeled it "performance assessment." In what follows, I will discuss each of these camps in some detail.

While these camps are in fact relatively self-conscious, the world of assessment does not necessarily fit neatly into these three possibilities. There is, to be sure, significant overlap in the camps. Thinking of the current debate over educational assessment as a

debate among three camps, however, is useful as a means for understanding the confusion and disagreement that clearly exists. I will discuss each camp in the order that it would be placed on a continuum, from traditional to alternative.

The Psychometric Tradition

As has earlier been mentioned, assessments in the psychometric tradition are used to an astonishing extent in the United States. These tests are developed by state school systems but much more commonly by commercial test publishers. Interestingly, this is an industry that exists in no other country on the scale that it does in America.

Given the ubiquity of this kind of test, it is not surprising that its characteristics have become almost synonymous with assessment in this country. In organizational terms (refer to Figure 3), tests in the psychometric tradition are standardized, and measures are administered on one occasion. Indeed, these assessments are often, erroneously, referred to as "standardized" tests. All assessments, however, can be standardized, and care should be taken not to misspecify a very general assessment characteristic.

Psychometric tests are quite recognizable for their physical features. They rely overwhelmingly on item response instruments such as multiple-choice, matching, and true-false questions. It is in the technical arena, however, that these assessments are most clearly distinguished from the other two approaches to assessment. A common property of these tests is that they are timed; in fact, time limits are precisely chosen in order to achieve maximal standard deviation of scores (Nunnally & Wilson, 1975). Too much time allowed to complete a test would limit variation as students working relatively slowly would join their fast-thinking peers in the ranks of high-scorers. Too little time allows no one to answer correctly a large number of items. In addition, these tests tend to be independent, indirect, norm-referenced, objective, and externally scored.

Tests in the psychometric tradition are widely used for all three functions of assessment: (1) classroom instructional practice; (2) system monitoring; and (3) individual student selection, placement, and certification. Indeed, in rather questionable practice, some tests are used for all three functions. Some of the more notable examples of tests out of the psychometric mold would be the Achievement Tests and Scholastic Aptitude Test

(SAT), available through the College Board; the Test of Adult Basic Education (TABE), published by CTB/McGraw Hill; and the Iowa Silent Reading Test, published by Harcourt Brace Jovanovich.

Given the characteristics of tests of this kind, the psychometric test maker's ideal is clear: Tests should be easy to administer and score, versatile and broad in measurement, relatively inexpensive, and understood with respect to reliability. Valid scores are taken to result from these objective and reliable conditions. Leaving the issue of validity aside, it is difficult to claim that what I have defined as psychometric tests fail in general to achieve these ideals. The tests are administered through standardized procedures that require no special training in assessment on the part of the examiner. The tests are, in addition, often machine scored. Versatility is clearly a property of such tests, one need only note the extraordinary number of applications that have been devised. And a broad sample of any topic is more achievable using perhaps hundreds of multiple-choice questions rather than two or three essay questions. The low cost of these assessments is proven, normally around $6.00 per student—although this figure does not include indirect costs stemming from time to administer the test and lost instructional time preparing for a test which might not be directly related to curriculum. The total cost of such assessments, therefore, is likely to be much higher (Office of Technology Assessment, 1992, p. 27).[4] Finally, the reliability of such tests is widely known to be quite high, at least in the case of "respectable" psychometric tests. Typical test-retest reliability for multiple-choice tests of writing, for example, lies between .8 and .9; whereas, reliability for an essay test of writing is normally not higher than .7 (Nunnally & Wilson, 1975). It is the larger number of items in a multiple-choice test that contributes to reliability: The inevitable random error in scoring items is mitigated by their large numbers, a benefit not enjoyed by two- or three-question essay assessments.

There has been considerable criticism of tests in the psychometric tradition, however. These criticisms of standardized, norm-referenced, multiple-choice (i.e., psychometric) tests have been based on three general test properties: (1) the underlying theory of learning assumed by testmakers, (2) the content measured, and (3) the

[4] Here, it was estimated that the *total* cost of a commercial standardized test would be roughly $110 per pupil.

instructional effects of tests (Baker, Freeman, Clayton, 1990, p. 1). I will consider each of these three sets of criticisms in turn.

It has been asserted with increasing forcefulness that tests in the psychometric tradition rely on an outdated—or, rather, inappropriate—theory of learning. Essentially, test constructors have assumed an ordered hierarchy of knowledge, which can be broken down into independent skills that are mastered through rote learning (Grubb, Kalman, Castellano, Brown, & Bradby, 1990). Psychometricians assume, in other words, that learning is linear and sequential and that complex understanding can only occur after elemental learning has taken place (Shepard, 1990, p. 8). Designers of such tests are rather locked into this position on learning, however, because of the constraints on test construction. Content domain must be measured by individual items with fixed answers. Even complex tasks such as writing and reading need to be broken down into their supposed elemental skills such as spelling or decoding.

The problems associated with this approach to learning have been thought to be severe. Learning becomes decontextualized and therefore meaningless. The tests can stunt innovation on the part of teachers insofar as they must concentrate on low-level skill components in their teaching rather than higher-level processes. In addition, students, too, are given an incentive to focus on low-level skills—and test-taking strategies—in order to succeed on the high-stakes versions of psychometric tests. Finally, these tests limit the range of rewarded and successful learning styles to the single depersonalized type already mentioned (Frederiksen & Collins, 1989; Grubb et al., 1990).

This limitation on learning styles contributes to a related problem often charged against standardized, multiple-choice, norm-referenced tests: They are biased against language and ethnic minorities and, in some cases, females (Neill & Medina, 1989, pp. 691-692). Such groups, it is theorized, employ different styles of learning than majority, or male, students (a claim that is contested by test publishers). There is ample evidence, however, that some minority groups do less well on traditional tests of achievement and aptitude.

Critics of psychometric-based tests, not surprisingly, have a much different view of learning in mind and, therefore, different kinds of assessment. They speak of the need for a "wholesale" transition from a "testing culture" to an "assessment culture"—that is, a need

to regard thinking as performance rather than as just decoding or calculation (Wolf, Bixby, Glenn, & Gardner, 1990). This "constructivist" view treats learning not as a progression from facts to comprehension to analysis but as a progression from simpler to more complex models (Shepard, 1990, p. 11). Assessment must look to the level of complexity of student understanding, therefore, and not simply the number of facts recognized, which tends to be the approach of traditional testing methods (Wilson, 1991, pp. 3-4). The criticism of the learning theory assumed by psychometricians, therefore, constitutes one of the attacks on traditional assessment that has taken place in the ongoing debate.

Critics also point to the content measured by traditional tests as a reason to try new assessment techniques. There is a sense that tests in the past have not measured the kind of learning that educators (and anyone else, for that matter) judge to be important. Constrained by test characteristics as well as by the "skills and drills" theory of learning, it is argued that tests have too often focused on what is easily measurable. Testmakers have dealt in numbers of facts rather than in levels of understanding (Wilson, 1991, p. 6). An analogy for this type of measurement would be to assess a basketball team by its average vertical leap rather than by its success in basketball games. What has been thought of as a benefit in multiple-choice testing, therefore—the efficient measuring of students' accumulated declarative knowledge—has come to be one of its primary liabilities. The misdirected instruments of traditional assessment, in spite of their cost efficiency and reliability, are thought not to be valid measures of student learning. They measure "irrelevant," or at least tangential, content (Mehrens, 1992, pp. 3-4).

The final type of criticism of the psychometric tradition and its tests has to do with a commonly held notion about the relationship between testing and instruction. Tests, particularly those involving high stakes (e.g., teacher evaluation or student selection to college), have a remarkably large influence on instructional practice: Teachers teach to tests. Yet multiple-choice tests are not intended to support or enhance instruction. They are intended to be neutral and objective measures of student achievement, much like a thermometer measures air temperature without changing the temperature of a room.

Testing effects might not be so benign as this, however. As already mentioned in connection with theories of learning, multiple-choice test constructors do make assumptions about how students learn. Teachers desiring to achieve success on such tests, therefore, have an incentive to emphasize elemental skills in instruction. As one critic has

put it, educational evaluation has been sending the message to students, teachers, administrators, and legislators that "the system values rote memorization and passive recognition of single correct answers" (Mitchell, 1992b, p. vii). Traditional testing has apparently corrupted the relationship between student and teacher insofar as teachers are unable to freely direct instruction to its highest ends.

These criticisms of the psychometric tradition have been dealt with recently in two ways. That is, two additional camps, or collective points of view, now exist in assessment circles.

Performance Assessment

The first of the two ways to deal with the supposed shortcomings of tests in the psychometric tradition has been to improve the design of the assessment instruments used. One might think of performance assessment as an effort to build a better mousetrap. For advocates of performance assessment, the goals of testing and the variety of possible uses do not need to change. Indeed, the same ground rules in the evaluation of assessment systems are used as in the traditional models: reliability and validity. As has earlier been discussed, performance assessment is an outgrowth of the view that assessment has failed to meet the need for valid measures of student achievement.

Advocates of performance assessment emphasize one way in particular that traditional assessment cannot be relied upon to consistently provide appropriate information about student achievement. Tests, especially tests involving high stakes for individuals or school systems, play a role in the educational process: They exert pressure to modify behavior. High-stakes tests set an agenda for educators by establishing the kinds of behavior (i.e., content and instructional practices) that are valuable. Testmakers, it is argued, should keep this dynamic system in mind by designing assessments that take test effects into account (Frederiksen & Collins, 1989).

Frederiksen and Collins refer to tests that do take the effects of instructional changes brought about by their own introduction as "systemically valid." The principles for the design of such tests establish a kind of paradigm for performance assessment: First, a set of desired tasks, more than basic skills, should be identified. Examples might be essay writing, auto repair, or scientific investigation. Second, the "primary traits" for

each task should be established. Such are the learnable subprocesses that enable complicated thinking to be assessed. Examples of primary traits in writing assessment might be clarity and persuasiveness. Third, test designers should establish a kind of library of exemplars in order to make clear to all the desired level of achievement. And finally, a training system is necessary for those who will be scoring the assessment.

The two most obvious characteristics of performance assessment, therefore, are that they are direct assessments of behaviors and that they are subjectively scored (Frederiksen & Collins, 1989). Other characteristics of performance assessment are that they are standardized, consisting of a wide range of instruments (though *not* of items tapping elemental skills, e.g., multiple-choice questions), integrated, criterion-referenced, and externally scored. Particularly important distinctions between performance assessment and tests of the psychometric tradition are that the former require students to construct their own responses to items, are largely syllabus-driven, and employ a greater variety of instruments (e.g., projects, experiments, simulations, and portfolios) (Office of Technology Assessment, 1992, pp. 18-20).

Yet performance assessments are tests in every sense that traditional assessments have been, and performance assessments place considerable value on the reliability and validity of assessment programs. These technical qualities are particularly important when test results are to be used for high-stakes applications such as postsecondary selection and system accountability. Indeed, it is just the high-stakes effects of which an important constituency within the performance camp wishes to take advantage. Proponents of national, high-stakes assessment assert that assessment can be "a cost-effective lever for changing the system." Educational assessment can serve a policy function by setting "well-defined and demanding standards" in order to break the cycle of gravitation to "*de facto* national minimum expectations" (National Council on Education Standards and Testing, 1992, p. 2). The position of the National Council on Education Standards and Testing, as well as like-minded organizations, is to view national assessment as a means for serving several national interests such as promoting educational equity, preserving democracy, and improving economic competitiveness by holding students to "clear standard[s] of achievement that matter to them" (Resnick, 1991, pp. 3, 6). Reward for effort is a "fundamental American value," and students are presently not enjoying it under the auspices of the mental testing movement (Resnick, 1990, p. 1). The council envisions that the new assessments could "eventually be used for such high-stakes purposes as high

school graduation, college admission, continuing education, and certification for employment," as well as system accountability (National Council on Education Standards and Testing, 1992, p. 5).

Not surprisingly, the call for national standards and assessments in America is based largely on existing assessment programs in Europe and Asia. British and German systems, for example, both in vocational and academic subjects, realize to a remarkable extent the assessment ideals expressed by advocates of performance assessment—advocates of *national* performance assessment in particular.

Performance assessment is a movement much larger than one going hand-in-hand with national standards, of course. In fact, thirty-six states are currently using this model in the assessment of writing for system monitoring purposes (Office of Technology Assessment, 1992, p. 201). Alaska, Arizona, Arkansas, Connecticut, Kentucky, New York, and Vermont are particularly committed to using performance assessment in multiple disciplines to evaluate student achievement (Brewer, 1992, p. 28). Also, the National Assessment of Educational Progress (NAEP), a purely system-monitoring assessment, will conduct a large-scale assessment of writing later this year using student portfolios. While these state and national efforts do not represent a complete departure from the earlier methods of the psychometric community, they do signify a reduced reliance on key features of traditional assessment such as multiple-choice items.

These and other examples of the movement from psychometric to performance assessment are very much in their infancy. Initial efforts at performance assessment have uncovered technical issues that will require considerably more research before such assessments will be able to accurately and fairly measure achievement for high-stakes applications. Indeed, curricular goals and standards of achievement have not been agreed upon, yet they are prerequisites for developing appropriate assessment methods (Office of Technology Assessment, 1992, pp. 18, 26). There has already been in-fighting among professional organizations attempting to set criteria for discipline-based performance assessments (Mitchell, 1992b, p. xvi). In the case of national standards and assessments, a national effort is needed to "facilitate and coordinate" test development (National Council on Education Standards and Testing, 1992, p. 28)—surely this will not come quickly or easily.

Technical issues such as reliability and validity also pose a significant but probably not insurmountable challenge. The subjective scoring required in performance assessment has been dealt with successfully in other applications: diving, figure skating, and gymnastics come to mind (Maeroff, 1991, p. 275). Already, inter-rater reliability in CAP writing assessments, for example, has recently been as high as .9 (Mitchell, 1992b, p. 188). The important thing at this point is to continue to address the criteria of reliability and validity in a systematic way, despite complications and high costs, and not to move too quickly. Performance assessment used to make decisions about curriculum changes is a relatively benign application, but high stakes applications require much more complete technical knowledge about performance assessment than is now available. Nevertheless, there are examples of disagreement on this point: Already the state of Kentucky has moved to reward improving schools, based on performance assessments, with cash awards and to penalize failing schools with a "school in crisis" label—with faculty and administration subject to transfer or dismissal (Foster, 1991). Hopefully, however, greater caution with performance assessment will prevail.

Alternative Assessment

As we have seen, advocates of performance assessment as a rule do not dispute the goals of traditional assessment instruments which are to measure student achievement for a wide range of purposes, including high-stakes applications, and to do so reliably and validly. Instead, performance assessment is a movement to *more effectively* accomplish the goals of testing. A second reaction to the psychometric tradition, however, here referred to as "alternative assessment," is comprised of advocates who do not share even a common view of the goals of assessment with those in the camp of performance assessment—and therefore also the psychometric tradition. The "build-a-better-mousetrap" mentality implicit in the examples of performance assessment, which links it to the psychometric tradition, and the wariness felt by many about the common uses of educational testing as found in both the psychometric and performance testing camps make alternative assessment perhaps the most distinct of the three approaches to assessment (Haney, 1984, p. 685).

There are three criticisms of the psychometric and performance traditions that unite advocates of alternative forms of assessment. First, both norm- and criterion-referenced

measurement and therefore both traditional and performance assessment can reflect the "sequential mastery" learning theory discussed earlier. In order to measure student achievement rigorously against some standard, learning needs to be broken down into disciplines and competencies, becoming decontextualized and possibly meaningless to students. Even in the case of performance assessment, there is an incentive to emphasize the rote learning of skills which are taken to be prerequisite for more complex tasks. Many tests of vocational curricula, for example, attempt to assess students based on extensive lists of individual competencies thought necessary to accomplish involved job tasks. In addition, the study and assessment of Spanish, for example, independent from the study of English might be "confusing for kids" insofar as disciplinary boundaries divide students' attention (Theodore Sizer quoted in Rothman, 1992a, p. 15). The approach within alternative assessment is to avoid the dissection of complex skills by treating occasions of learning as context-specific, indivisible moments.

Second, performance assessment, like psychometric assessment, keeps assessment and the individual classroom at a distance. The national standards and assessment movement, in particular, causes problems associated with reduced autonomy and professionalism among teachers. Competency-based assessment "dictate[s] the curriculum, a policy that is anathema to proponents of alternative assessment" (Maeroff, 1991, p. 276). Assessment should be used and even developed as a tool by individual teachers for the sole purpose of furthering instruction. There should be complete flexibility in materials, activities, and definitions of progress so that teachers can ensure that instruction meets the needs of the wide range of children they teach. Instruction that is tied to assessment ("test-driven" instruction), which tends to be the case with most of the competency-based models in performance assessment, defeats these instructional goals (Wolfe, 1989, p. 4). Proponents of alternative assessment see the national standards and assessment movement, for example, as "the tip of the iceberg" in which the central government commands strict authority over children "in the name of high standards and international competition" (Theodore Sizer quoted in Rothman, 1992b, p. 8).

Finally, taken from the student point of view, both psychometric and performance models of assessment have deleterious effects. "High-stakes" assessment, both of the psychometric and performance variety, has—and will continue to have—a positive effect on student dropout and failure rates and will contribute to whatever feelings of overwork and stress students might already have. Advocates of performance assessment, for

example, look to European assessment systems as models for America—despite the fact that testing in these countries is used to limit participation in postsecondary education by the use of rigorous exams with high failure rates. In the shared view of those in the alternative assessment camp, testing—whether of the traditional or more recently fashionable brand—has too often been used against children and against the better judgment of teachers as well. "Alternative assessment" therefore constitutes a move away from a "testing culture" to a more benign "assessment culture" (Wolf et al., 1990).

This new assessment culture has its foundation in the belief that assessment's primary service is to individual classroom instruction *and not student selection and certification or system monitoring*. Alternative assessment is therefore linked to a set of assumptions about how students learn, what motivates them to learn, and what teachers should be doing to facilitate learning.

Perhaps most distinctive in the alternative camp is the view that both high-stakes and standardized tests do not facilitate learning. Such tests, it is argued, inevitably seek to measure what is inexpensive and easy to measure rather than the skills associated with more genuine learning. For advocates of alternative assessment, the most important properties of an assessment system are that it be multidimensional—that is (recalling the terms from Figure 3), that it be cumulative, using a variety of types of test instruments, integrated with the curriculum, and subjectively evaluated. In addition, members of this camp tend to favor nonstandardized, group assessment. Only with assessments of this type can assessment support instruction by engaging teachers in a debate about learning and involving students in a process of reflection on their work (Wolf et al., 1990).

Examples of alternative assessment instruments, therefore, tend to have a "home-made" feel to them—particularly when compared to the sophisticated test systems of traditional American assessment. Folders of student work, or portfolios, for example, are favorite assessment instruments, as are interviews, demonstrations, and projects. Assessment of groups, with its focus on realistic problem-solving situations, is also encouraged (Hill, 1989, p. 7; Kobrin, Albin, & Rudman, 1989, pp. 3-4).

Given the sole legitimate purpose of assessment in the opinion of many in the alternative camp—aiding instruction—it is not a problem that assessment instruments turn out to be unstandardized and therefore utterly unreliable and not generalizable. Students

need not be compared against one another, particularly when only the number of facts recalled is assessed. Instead, assessment should help to instill a value in students and teachers that assessment is a personal responsibility, that "first draft" work is never good enough, and that personal development is as important as individual achievement (Office of Technology Assessment, 1992, p. 230).

A Response from the Psychometric Tradition

I argued earlier that both performance assessment and alternative assessment can be understood as reactions to the psychometric tradition in American educational assessment. As I have attempted to show, these two relatively recent camps have formed out of dissatisfaction with assessment as traditionally conceived: Performance assessment has been carried by the belief that assessment had been systematically *failing to measure* the skills and knowledge that educators care most about. Alternative assessment, on the other hand, has been primarily concerned with the *uses of assessment* (i.e., that tests too often have been used against schoolchildren by failing to support—and even by impairing—classroom instruction).

Not surprisingly, the psychometric establishment has not been silent in the face of criticism over roughly the past half decade. Recall the commentator who recently wrote that "It is *impossible* to be impressed by the lack of objectivity and scientific rigor of many of those" in the assessment debate (Mehrens, 1992, p. 1). Another group noted the "disdain for professional standards of technical quality among some of the more committed advocates" of the new assessments (Williams et al., 1991, p. 1).

This backlash has been most severe against the positions taken by those in alternative assessment. Some reformers, for example, have spoken of alternative assessment as stronger than traditional assessment with respect to validity while conceding the superior reliability of traditional tests. Naturally, psychometricians have responded that because reliability is a prerequisite for validity such claims are ridiculous and indicate the lack of technical sophistication of many in the debate. Indeed, rarely among advocates of alternative assessment (and to a certain extent performance assessment) have the criteria for evaluating assessments been addressed. The new assessments are simply assumed to be valid because they measure an entire performance rather than a single skill (Linn, Baker, & Dunbar, 1991, p. 16).

Apparently advocates of alternative assessment feel that a serious discussion of technical issues is unnecessary given the camp's relative lack of interest in two of the three functions of assessment: (1) system accountability and (2) student selection, certification, and placement. The insistence of advocates of alternative assessment on collaboration and constant feedback, for example, makes the assessment of individual students problematic, if not impossible (Williams et al., 1991, p. 14).

Performance assessment fares much better in the eyes of most psychometricians because of the similar stance these two camps take on the preceding issues. In general, advocates of performance assessment do take seriously the selection and the accountability functions of assessment and recognize the importance of technical issues such as reliability and validity in the pursuit of these functions. Indeed, there has been quite extensive collaboration between psychometricians and advocates of performance assessment. The Educational Testing Service, for example, has consulted with several states in the development of performance assessments. ETS has also, of course, for many years designed the Advanced Placement examinations published by the College Board, most of which are textbook examples of the performance assessment ideal.

There are concerns with respect to performance assessment, however, having generally to do with cost and reliability. Scoring discipline-based examinations of sixteen-year-olds in the United Kingdom—tests which look much like those envisioned by performance assessment advocates—costs over $100 per student. Similarly, scoring a recent NAEP fourth-grade mathematics assessment required about $150 per pupil (Shepard, 1991, p. 238). Imagine what costs like these would do to the assessment budget of, say, Massachusetts, which currently spends but $1.2 million to test students in three subjects and in three grade levels. Indeed, it would cost almost $7 million to assess only the 65,000 or so sixteen-year-olds in Massachusetts in one subject alone using the performance assessment model (Office of Technology Assessment, 1992, p. 141). Performance assessment might well turn out to be well over ten times more expensive than psychometric assessments have been. In addition, due to the small number of items normally used in performance assessments, their lack of internal consistency, and the subjective scoring processes they require, unreliability might just turn out to be the "Achilles' heel" of performance assessment (Mehrens, 1992, pp. 16-17; Williams et al., 1991, p. 12).

Those sympathetic to traditional assessment have also pointed out that many of the criticisms by both the performance and alternative assessment camps have been plausible not because the instruments of the psychometric tradition (i.e., norm-referenced, multiple-choice tests) are bankrupt but instead because these assessments have been misused. Norm-referenced tests were "pressed into service" to meet the expanding demand for educational assessment in America and they might not have been designed for many of these applications. Multiple-choice formats continue to have unique properties for educational assessment, it is argued, and test developers are exploring ways to improve these formats in the measurement of complex thinking skills (Office of Technology Assessment, 1992, pp. 165, 195).

The position of those in the psychometric tradition, despite the wave of criticism and re-evaluation I have discussed, continues to be that assessments in education should be easy to administer and score, versatile and broad in measurement, relatively inexpensive, and understood with respect to reliability and validity. If these are to be the criteria for assessment, then the position is, furthermore, that educational measurement as traditionally conceived in America continues to be a necessary component of any new assessment system. Relying only on the methods proposed by alternative and performance assessment advocates, at least at the moment, would leave the educational establishment floating in "uncharted" waters (Office of Technology Assessment, 1992, p. 249).

Indeed, in order to deal with such uncertainty, an interesting development has recently emerged among those designing assessment instruments (some of whom are simply observers with respect to the debate I have described). Particularly in the vocational assessment world, assessment constructors are beginning to offer hybrid instruments—that is, assessments that incorporate more than one of the three general approaches to assessment. For example, the V-TECS assessment of Tractor Trailer/Truck Driver[5] uses both multiple-choice and matching items to assess factual knowledge and uses performance items with pre-established criteria for evaluation to assess practical expertise. Until a greater technical understanding is reached of the alternatives to traditional assessment, we can expect to see continued reliance on traditional assessment practices and more

[5] For more information, contact V-TECS (the Vocational-Technical Education Consortium of States), 1866 Southern Lane, Decatur, GA 30033-4097.

combinations of old and new approaches. In times of uncertainty, a diversified portfolio, if you will, is the safest option.

WHY SO MUCH DISAGREEMENT? A CONCLUSION

That schoolchildren in America should be assessed in some way is a view, naturally, that unites each of the three assessment camps that have been outlined in this paper. We have seen that there is considerable disagreement, however, over how to proceed from this common point of departure. Should we test for achievement in education? Are tests adequate measures of educational outcomes? What kinds of assessments should we construct? What should be done with their results? Each of these questions has led to a variety of answers. Given the importance of such questions, it seems worthwhile to find where consensus exists and where it does not to attempt to explain the disagreement. This paper will conclude, therefore, with a brief discussion of several points which will help explain some of the disagreement illustrated earlier.

Charlatanism

After reviewing the assessment debate, it would be difficult to deny that many in the debate are expressing untenable positions. Perhaps because of the number of individuals now involved in the discussion or the widespread sense of importance regarding its outcome, there is an abundance of uninformed opinion and misuse of terms which has clouded the issues. As earlier discussed, that "as validity goes up, reliability and objectivity go down" (Burstein, 1991, p. 4) is simply not the case. And while claiming that "standardized testing" is "harmful to educational health" (Neill & Medina, 1989, p. 688) might succeed in conveying a sense that assessment has been misused in the past, it also badly misuses the term "standardized"—a common feature of the assessment debate.

Theories of Learning

There are, on the other hand, disagreements in the assessment debate that are quite fundamental. Much of the reaction to the psychometric tradition is based on disagreements

on how students learn—and therefore on how they should be taught and assessed. Test constructors have assumed in the past that learning is sequential and hierarchical and can be broken down into independent skills mastered through rote learning. Others argue increasingly that learning is subjective, context-specific, and complex—even at early cognitive stages. This quite theoretical side of the assessment debate is responsible for a certain amount of the disagreement and ensures that the debate will continue well into the future.

Exaggerated Test Capabilities

Apart from the inaccuracies and mistakes just mentioned, there is also a fairly widespread phenomenon of what might be thought of as wishful thinking. Many seem to be arguing that one brand of assessment is capable of serving all three functions of educational assessment or that assessment alone is a lever powerful enough to solve America's educational problems. As Williams et al. (1991) have pointed out, it is unrealistic to expect assessment, however "authentic," to "make all students effective problem-solvers and critical-thinkers" (p. 3).

Figure 4 is a useful illustration of the properties needed to satisfy the three main functions of educational assessment. Not surprisingly, given the logistical and instrumental differences between an assessment of, for example, an entire state's sixth-graders and a single teacher's sixth-grade class, there are differences in the requirements that should be placed on assessments depending upon their function. An assessment of several hundred thousand students, therefore, should be relatively cost-efficient and widely comparable—both secondary criteria for an assessment of Mr. Jones's math class. Despite this range in assessment functions, exaggeration continues to thrive. One influential advocate of alternative assessment, for example, has recommended that "for accountability, it is only necessary to sample portfolios" (Mitchell, 1992a, p. 36). Portfolios of student work are indeed a promising innovation for some functions of assessment; but for system monitoring, they introduce dramatic problems of cost and reliability of scores both across students and over time. Precision about the uses to which assessment is put and awareness of the technical requirements applicable would help to reduce some of the confusion in the debate.

Figure 4

Properties Needed for the Three Functions of Assessment

	Classroom Instructional Guidance	System Monitoring	Selection, Placement, and Certification
Who needs to be described	Individuals	Groups of students	Individuals
"Stakes" or consequences attached	Low	High or low	High
Characteristics of the test needed			
Comparability of information	Low	High	High
Impartial scoring (not teachers)	No	Yes	Yes
Standardized administration	No	Yes	Yes
Type of information needed			
Detailed vs. general	Detailed	General	General
Frequency	Frequently during a single school year	Once a year or less	Once a year or less
Results needed quickly	Yes	No	No
Technical requirements			
Need for high test reliability (internal consistency and stability)	Can vary	Depends on group size	Very high
Type of validity evidence	Content	If low stakes, content; if high stakes, content and construct	Content; additional validity evidence must be demonstrated for the specific purpose (e.g., certification = criterion validity, selection = predictive validity

Source: Office of Technology Assessment (1992); adapted from Resnick and Resnick (forthcoming).

Interests and Ideology

One of the most fascinating—and debilitating—aspects of the assessment debate is that it has become, for many, ideological. The three camps that were earlier discussed are increasingly self-conscious groups of scholars, teachers, and others, each group with an approach to assessment and an interest in seeing successful implementation of that approach. Emotionally charged terms such as "authentic," "direct," and "intelligent" have been applied to the alternative camp by its members (Office of Technology Assessment, 1992, p. 202), while the psychometric community has seen itself as part of the larger scientific community and too often above evaluating its own approaches and assumptions. Indeed, many advocates of the new assessments seem to see themselves as revolutionaries standing up to a traditional assessment regime whose sophisticated mathematical and statistical techniques they do not always understand.

The most deleterious effect of this ideologically charged debate has been the devaluing of assessment functions for which a particular approach is not successful. The psychometric community, for example, has traditionally seen instructional guidance as irrelevant to serious assessment: Like thermometers, telescopes, and other instruments, educational tests should accurately measure while remaining a neutral influence on the environment. Furthermore, citing the effects of "high-stakes" assessment on students, most advocates of alternative assessment ignore the selection, placement, and certification function of assessment—a rather unrealistic position.

Overdoing Testing

Finally, the existence of a widespread debate on assessment was probably unavoidable given the history of assessment practice in this country. For a variety of reasons—such as increasing numbers of students, equity concerns, concerns about economic competitiveness, and the growth of the federal and state government role in school funding (Office of Technology Assessment, 1992, pp. 54-56)—America has experienced ever-increasing pressure to rely on formal assessment in education. That kindergartners in Minneapolis will no longer face a twenty-minute test to assess their readiness for the first grade (Cohen, 1992) will have little impact on a nation that uses testing far more than any other nation (Office of Technology Assessment, 1992, p. 135). In addition, the "stakes" associated with assessment have grown over the years as

important decisions about students have increasingly been made based upon the result of a single test score.

It seems unlikely that the debate in educational assessment will lead to a reduction in the reliance on assessment in American education. Now, more than ever, testing is seen as the principal catalyst for educational reform.[6] Hopefully, however, in the end, this debate will cause those involved in education to re-evaluate the assumptions that have been made about student assessment. If not, then the answers to the questions with which this paper began will continue to be difficult questions, but not as difficult as the recent assessment debate would lead us to believe.

[6] Commission reports from *A Nation at Risk* (National Council on Excellence in Education, 1983) to *Raising Standards for American Education* (National Council on Education Standards and Testing, 1992) have advocated student achievement examinations as key to educational improvement.

REFERENCES

Baker, E. L., Freeman, M., & Clayton, S. (1990). *Cognitive assessment of subject matter: Understanding the marriage of psychological theory and educational policy in achievement testing*. Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Brewer, W. R. (1992, April 15). Can performance assessment survive success? *Education Week*, *11*(30), 28.

Burstein, L. (1991). *Performance assessment for accountability purposes: Taking the plunge and assessing the consequences*. Unpublished manuscript. Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Cohen, D. L. (1992, April 29). Minneapolis schools suspend high-stakes kindergarten testing. *Education Week*, *11*(32), 1.

Cronbach, L. J. (1975). Five decades of controversy over mental testing. *American Psychologist*, *30*(1), 1-14.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 237-270). Washington, DC: American Council on Education.

Foster, J. D. (1991, February). The role of accountability in Kentucky's Education Reform Act of 1990. *Educational Leadership*, *48*(5), 34-36.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Grubb, W. N., Kalman, J., Castellano, M., Brown, C., & Bradby, D. (1990). *Coordination, effectiveness, pedagogy, and purpose: The role of remediation in vocational education and job training programs*. Unpublished manuscript.

Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research*, *54*(4), 597-654.

Hill, S. (1989). Alternative assessment strategies: Some suggestions for teachers. *Information Update* (special issue on alternative assessment), *6*(1), 7, 9.

Kobrin, D., Moser, A., & Rudman, B. (1989). Is a new U.S. history curriculum possible? A progress report on the pilot project at Hope Essential. *Radius* (newsletter of American Federation of Teachers Center for Restructuring), *2*(2), 2-9.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Maeroff, G. I. (1991). Assessing alternative assessment. *Phi Delta Kappan*, *73*(4), 272-282.

Mehrens, W. A. (1992, Spring). Using performance assessment for accountability purposes: Some problems. *Educational Measurement: Issues and Practice*, *11*(1), 3-9.

Mitchell, R. (1992a, April 29). Beyond the verbal confusion over "tests." *Education Week*, *11*(32), 36.

Mitchell, R. (1992b). *Testing for learning: How new approaches to evaluation can improve American schools.* New York, NY: Free Press.

National Council on Excellence in Education. (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: U.S. Government Printing Office.

National Research Council. (1991). *Performance assessment for the workplace* (Vols. I & II). Washington, DC: National Academy Press.

Neill, D. M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan*, *70*(9), 688-697.

Nunnally, J. C., & Wilson, W. H. (1975). Method and theory for developing measures in evaluation research. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (pp. 227-288). Beverly Hills, CA: Sage Publications.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington, DC: U.S. Government Printing Office.

Quinto, F., & McKenna, B. (1977). *Alternatives to standardized testing*. Washington, DC: National Education Association.

Resnick, L. B. (1990). *An examination system for the nation.* Press release. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Resnick, L. B. (1991, March 7). Testimony before the Senate Committee on Labor and Human Resources, Subcommittee on Education, Arts, and Humanities.

Resnick, L. B., & Resnick, D. P. (forthcoming). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction.* Boston, MA: Kluwer Academic Publishers.

Rothman, R. (1992a, May 6). Program to offer courses, exams for high schools. *Education Week, 11*(33), 15

Rothman, R. (1992b, February 5). Standards and testing report is hailed, criticized. *Education Week, 11*(20), 8

Shepard, L. A. (1990). *Psychometricians' beliefs about learning*. Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*(3), 232-238.

Testing. (1992, March 18). *Education Week, 11*(26), 10.

Williams, P. L., Phillips, G. W., & Yen, W. M. (1991, April). *Measurement issues in high stakes performance assessment.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wilson, M. (1991). *Implications of new perspectives on student assessment for Chapter I and its evaluation: Educational leverage from a political necessity.* Unpublished manuscript.

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1990). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Wolfe, M. (1989). Rethinking assessment: Issues to consider. *Information Update* (special issue on alternative assessment), *6*(1), 3-4.