DOCUMENT RESUME

ED 353 445 CE 062 881

AUTHOR	Hill, Clifford; Larsen, Eric
TITLE	Testing and Assessment in Secondary Education: A Critical Review of Emerging Practices.
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	108p.
AVAILABLE FROM	NCRVE Materials Distribution Service, 46 Horrabin Hall, Western Illinois University, Macomb, IL 61455 (order no. MDS-237: $5.25).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Achievement; *Achievement Tests; *Educational Testing; Grades (Scholastic); Grading; *Holistic Evaluation; Job Skills; Outcomes of Education; Portfolios (Background Materials); Secondary Education; Self Evaluation (Individuals); *Student Evaluation; Vocational Schools; Work Sample Tests
IDENTIFIERS	*Authentic Assessment

ABSTRACT

A study of assessment practices focused on vocational schools, especially those concerned with developing generic workplace skills, as well as comprehensive secondary schools that are developing closer relations with the workplace. The major reasons for educational testing in secondary schools are managing student learning, monitoring educational systems, and evaluating students for institutional purposes. Principles that are frequently used in discussions of testing and assessment policy are excellence, equity, and efficiency. These three qualities are better accomplished with authentic assessment, which does the following: (1) requires students to construct responses rather than select among preexisting options; (2) elicits higher order thinking in addition to basic skills; (3) uses direct assessment of holistic projects; (4) is integrated with classroom instruction; (5) uses samples of student work collected over an extended period of time; (6) is based on clear criteria of which students are made aware; (7) allows for the possibility of multiple human judgments; and (8) is more closely related to classroom learning. Three fundamental goals of education that authentic assessment can help to achieve are reforming curriculum and instruction, improving teacher morale and performance, and strengthening student commitment and capacity for self-monitoring. Compared to conventional testing, authentic assessment that teaches students how to monitor their own work makes far greater demands on both students and teachers. However, given the demands of the future workplace, internal evaluation and self-monitoring are increasingly critical skills. (Contains 126 references.) (CML)

NCRVE

National Center for Research in Vocational Education

University of California, Berkeley

TESTING AND ASSESSMENT IN SECONDARY EDUCATION: A CRITICAL REVIEW OF EMERGING PRACTICES

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)

TESTING AND ASSESSMENT IN SECONDARY EDUCATION: A CRITICAL REVIEW OF EMERGING PRACTICES

Clifford Hill

Eric Larsen

Teachers College, Columbia University

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-237

FUNDING INFORMATION

Project Title:	National Center for Research in Vocational Educatic
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
Disclaimer:	This publication was prepared pursuant to a grant with the Office of Vocational and Adult Education, U.S. Department of Education. Grantees undertaking such projects under government sponsorship are encouraged to express freely their judgement in professional and technical matters. Points of view of opinions do not, therefore, necessarily represent official U.S. Department of Education position or policy.
Discrimination:	Title VI of the Civil Rights Act of 1964 states: "No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance." Title IX of the Education Amendments of 1972 states: "No person in the United States shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving federal financial assistance." Therefore, the National Center for Research in Vocational Education project, like every program or activity receiving financial assistance from the U.S. Department of Education, must be operated in compliance with these laws.

ACKNOWLEDGMENTS

We wish to thank Sue Berryman and staff members of the Institute on Education and the Economy at Teachers College, Columbia University. They have provided effective support throughout the course of this research. In particular, Sue has offered thoughtful guidance about the delicate relations between secondary education and the workplace.

We would also like to thank Linda Darling-Hammond, who initially conceived this research. She has provided a wealth of information about authentic assessment in American high schools and, more importantly, her vision of its role in fostering humane and rigorous education.

TABLE OF CONTENTS

SITUATING AUTHENTIC ASSESSMENT

Major Purposes of Testing and Assessment

In its recent report, *Testing in American Schools: Asking the Right Questions,* the Office of Technology Assessment (OTA) (1992) provides a useful summary of major reasons for educational testing:

- to aid teachers and students in the conduct of classroom learning;
- to monitor system-wide educational outcomes; and
- to inform decisions about the selection, placement, and credentialing of individual students (p. 10).

These reasons, which the report characterizes as "functions," can be conveniently summarized as *managing student learning, monitoring educational systems,* and *evaluating students for institutional purposes.* There is a further convenience in describing the first of these functions as *classroom-oriented* and the second and third as *system-oriented*: in this way, we can distinguish between testing that is internal to the classroom and testing that results from external needs. In effect, classroom-oriented testing is largely concerned with shaping the processes of teaching and learning; system-oriented testing is concerned with establishing outcomes of these processes for various institutional purposes.

The OTA report points out that tests built around different functions do have a "common feature" in that they all "provide information to support decision making." They differ, however, in the "kinds of information they seek and the types of decisions they can support, and test results appropriate for some decisions may be inappropriate for others" (p. 10). The report then discusses the ways that different functions lead to different kinds of tests. We have summarized the main points of this discussion in Table 1.

Table 1
Test Characteristics Associated with Major Functions

Classroom-Oriented	System-Oriented	
For *managing learning* a test should—	For *monitoring systems* a test should—	For *evaluating students* a test should—
provide detailed information on specific skills	provide general information about achievement	provide information relevant to (1) special needs or (2) future performance
occur frequently	occur infrequently	occur infrequently
be administered and scored according to pedagogical needs	be uniformly and impartially administered and scored	be uniformly and impartially administered and scored
be linked to content that has been taught and have clear and open criteria for scoring	meet reasonable standards of consistency, fairness, and validity	meet particularly high standards of comparability, consistency, fairness, and validity
give feedback quickly in a form helpful to teachers and students	describe group rather than individual performance	provide individual student scores

Table 1 represents a broad selection of practices that have played a role in education in this country and in many other parts of the world. We refer to these practices as *traditional testing* throughout this report. In the United States and in parts of the world where American influence has been strong, the prototypical form of such testing is now the standardized multiple-choice test, which we refer to as *conventional testing*. Conventional testing is the product of an early twentieth-century movement which sought, through the application of scientific principles, to develop tests that were more efficient and equitable than traditional ones (Thorndike, 1913). Achieving these objectives, however, led to a drastic narrowing of the scope and functions that characterized traditional testing. For example, writing ability was effectively redefined, for assessment purposes, as knowledge of vocabulary, usage conventions, and grammatical niceties because the chosen format of conventional testing could not be convincingly used to evaluate actual text that students might be asked to write. In recent years, educators have become increasingly conscious of this reductionism and have proposed alternatives which, in certain respects, represent a return to traditional testing. The written essay, for example, has been reinstated as a useful means not only for assessing writing ability but also for finding out what students know about a particular subject and how well they can apply their knowledge.

The movement toward alternative practices, however, goes well beyond traditional testing. These practices are often referred to as *authentic assessment,* a term that conveys the notion that assessment should be built around tasks that are worth doing for their own sake. In determining intrinsic merit, educators derive standards from academic disciplines but, increasingly, from the larger society, as well. Hence, students are expected to use knowledge and methods of inquiry transmitted in school but engage in tasks that have social relevance. These new expectations have considerably expanded the repertoire of tasks that students can engage in. The written essay is no longer the exclusive form of discourse; other forms of writing are elicited, as well—for example, process journals that students maintain while working on a project. In addition, oral forms of discourse that are crucial to social interaction in school and the workplace are being utilized—for example, explaining how something works or how something is done (Black, Kay, & Soloway, 1987).

Moreover, students are not expected simply to create discourse. As many educators point out, schooling has relied too heavily on written symbols as a vehicle for teaching and learning. As a consequence, students can develop considerable skill at manipulating symbols in school yet have difficulty using their knowledge in more functional contexts. There is, for example, a considerable gap which must be bridged between the stylized categories of word problems and the ability to use mathematical knowledge to solve the kinds of problems arising outside the classroom. It is for this reason that authentic assessment has developed tasks that go beyond the purely linguistic. Students are asked to make useful or pleasing artifacts (e.g., a computer database or an architectural model) or to engage in performances (e.g., a science experiment or a community survey). Indeed, discourses, artifacts, and performances are the central categories in the taxonomy of authentic assessment that we develop in the second section of the report.

We have suggested that assessment is considered authentic when students are asked to engage in activities that have intrinsic merit from either an academic or a social perspective. We believe that it is useful to extend the concept of authenticity to include the perspective of students who are asked to take part in an assessment activity. If they do not find an activity meaningful in some respect, they will not be motivated to engage in it effectively, and it could hardly be regarded as fully authentic. Once we include the perspective of those being assessed, we are brought face to face with their diversity and must consider issues that have to do with gender, ethnocultural identity, and socioeconomic class.

One further way in which assessment should be authentic is in the standards that are used in evaluating students. Like other human institutions, schools have produced a distinctive culture in which the various groups involved—administrators, teachers, students, and parents—have made accommodations which allow them to get along without too much discomfort. As Brown, Collins, and Duguid (1989) point out, one of the byproducts of this process is that

> classroom activity very much takes place within the culture of schools, although it is attributed to the culture of readers, writers, mathematicians, historians, economists, geographers, and so forth. Many of the activities students undertake are simply not the activities of practitioners and would not make sense or be endorsed by the cultures to which they are attributed. . . . What students do tends to be ersatz activity. (p. 34)

This implies that the standards that should be applied are those of practitioners, but the only practitioners mentioned are those involved in various academic disciplines. What if we should apply the standards of practitioners in the workplace, which has often been suggested in recent years? The U.S. Secretary of Labor's Commission on Achieving Necessary Skills (SCANS, 1991) has worked out a scheme of generic skills which are useful in a broad range of workplace sites. They suggest that there are foundation skills (e.g., mathematics, problem-solving skills, and language skills such as reading and writing) and workplace competencies (e.g., interpersonal relationships, time management, information management, and technology use). However, much work remains to be done on these generic skills. As Berryman and Bailey (1992) observe, "The nation needs a technically and politically credible source of information about the foundation and generic workplace skills required across occupations and industries" (p. 121). In addition, research needs to be done to establish the levels of skills required and the extent to which these are actually predictive of performance on the job.

We will not pursue these matters further here but would like to suggest that fully authentic assessment practices have to do with the following:

1. the tasks themselves—do they have intrinsic merit?
2. students' interaction with these tasks—are they actively engaged as they carry out the tasks?
3. the standards that are applied to these interactions—would practitioners of x and y agree that these standards are valid?

In effect, authenticity has to do with what we ask students to do, how they engage with these tasks, and how we evaluate their engagement.

At this point we can return to the OTA scheme with which we began this discussion. How might it be modified to include the emerging practices of authentic assessment? These practices have potential relevance to all three major functions, though they have been primarily associated with the management of student learning. Many advocates of authentic assessment take a broader view of such learning than the OTA report, which presents diagnosis in fairly traditional terms. As these advocates point out, traditional testing—particularly what we describe as conventional testing—tends to take an external view of what diagnosis consists of (i.e., providing teachers and students timely feedback about how well specific knowledge and skills have been acquired). From their perspective, however, diagnosis is more properly concerned with uncovering how students think and learn. Moreover, these advocates focus on other important dimensions of student learning that are fostered by authentic assessment: (1) motivating students to work toward high standards and (2) developing in students a capacity to monitor their own work. This more comprehensive approach to student learning is reflected in Table 2, which presents a more detailed view of the purposes of assessment:

Table 2

Major Purposes of Educational Assessment

Managing Student Learning	• understanding how students think and learn • motivating high levels of achievement • developing the capacity for self-monitoring
Monitoring Educational Systems at Various Levels	• individual schools • school districts • states and countries
Evaluating Individual Students for Institutional Purposes	• selecting • placing • credentialing

A major question facing authentic assessment is whether it can effectively perform the two system-oriented functions. In the next section, we analyze certain uses of authentic assessment within this country to evaluate students from a systems perspective (these uses have largely to do with credentialing). We also analyze how the United Kingdom has used certain forms of authentic assessment to monitor its educational system at both primary and secondary levels. Such uses are quite tentative; it remains unclear to what extent authentic assessment can be used to carry out the system-oriented functions associated with traditional testing.

Endogenous and Exogenous Models of Education

In thinking about how assessment is carried out, we need to consider not only its major purposes but also the larger model of education within which it functions. At the present time, assessment functions within a model that can be described as largely *endogenous*; that is to say, it operates with categories and standards that arise largely from the educational system itself. To put it more concretely, life inside school is divided into familiar categories such as math and history in a way that it is not in other settings; and what counts as desirable within these categories is largely generated from within.

How subject matter is categorized has important consequences for how educational assessment is carried out. For one thing, certain subject matters are more highly valued and thus more thoroughly assessed. One particular scheme of values is reflected in the much publicized *America 2000* (1991) which lists the five core subjects of secondary education as English, mathematics, science, history, and geography. In reviewing the subject areas, some educators have been troubled by the absence of the arts. Others have wondered, noting the presence of geography, why foreign languages are left out. Still others have worried about the neglect of the social sciences, especially psychology and economics. In general, those who support authentic assessment advocate a more inclusive scheme of core subject matter:

- math and science
- language arts (including foreign languages)
- social studies (including history and geography)
- the arts

No matter what method of classification is followed, we need to recognize that different subject matters have evolved distinctive approaches to assessment. A particular subject matter develops its own sense of standards and procedures for insuring that these standards are met. For example, mathematics has provided a congenial home for well-defined problems with a clear right answer, whereas the arts were using portfolio methods long before they became fashionable. These distinctive approaches should not be viewed as the capricious operation of academic tradition; rather, they find a strong rationale in Howard Gardner's (1983) theory of multiple intelligences which posits distinctive ways of apprehending the world as fundamental to human intelligence. Gardner has himself been closely involved with assessment issues and has emphasized that how educators go about the task of evaluation varies necessarily with what they are evaluating: students' ability to think mathematically must be assessed differently from their capacity to construct imaginative narrative. Just as intelligences are various, so our methods of assessing them must be various as well.

There are, however, countervailing forces that mitigate against a differentiated approach to assessment. Certain advocates of authentic assessment are concerned with developing a model of education which is more *exogenous*—that is to say, one that reflects more closely the needs of the larger society. In an economically-oriented version of such a model, schools would integrate more closely what they teach with what the workplace requires. Working from this perspective, Berryman and Bailey (1992) have described a fundamental dualism that permeates education:

> The mismatch between the focus of our K-12 schools and serious, coherent economic preparation for our students is deeply rooted in the dualism between culture and vocation, head and hand, abstract and concrete, theoretical and applied. (p. 8)

As Berryman and Bailey point out, this dualism is perpetuated in the long-standing division of secondary education into the vocational and the academic: the academic is concerned with culture and the head, the vocational with the hand. In such an arrangement, the education of all students is diminished, for it does not foster the active integration of knowledge and skill which is fundamental to learning. Hence, Berryman and Bailey argue for an education that actively unites the vocational and the academic. Indeed, if the fundamental dualism were to be dissolved, major changes would take place in the way curriculum and instruction are organized and in the outcomes that would be valued. Rather than being organized around traditional content areas, assessment would focus more

directly on generically conceived competencies (i.e., the broad knowledge and skills that individuals need to solve real-world problems) (Resnick, 1987a; Stasz, McArthur, Lewis, & Ramsey, 1990). Students would not, for example, be assessed for mathematical thinking on the one hand and writing skills on the other. Rather they would deal with real-world problems that call for them to embed mathematical reasoning within a discursive structure. In effect, students would be asked to integrate knowledge and skills that have been compartmentalized in the endogenous model of schooling. Moreover, they would be forced to go beyond the purely academic and integrate its ways of knowing with technical—and indeed social—ways of doing. Or to expand the metaphor introduced by Berryman and Bailey, the head would be brought together with both hand and heart.

The Testing Paradigm

In educational theory, testing is just one form of a larger enterprise referred to as assessment. In much educational practice, however, testing effectively equals assessment, and certainly the general public often equates educational outcomes with test results. It seems important, then, in developing new methods of assessment to give some attention to how they will be categorized by users—students, teachers, administrators, and the public at large. A particularly interesting question is whether a new way of assessing will be regarded as another kind of test, with all that this implies in our culture, or as something more novel. Answering this question will enable us to add a pragmatic dimension to the taxonomy of authentic assessment that we build. In addition to categorizing assessment practices according to their internal characteristics, we will be able to give attention to how they are likely to be assimilated to the ongoing educational enterprise.

What, then, leads educators—and for that matter, students—to regard a particular activity as a test? We believe that the following four features are especially crucial in establishing a testing paradigm:

1. the activity is conducted in a single time frame of specified duration;
2. the activity consists of prescribed tasks that are presented in stable form;
3. the activity elicits individual responses that are not based on external resources; and
4. the responses are evaluated according to a pre-established scheme.

We should observe that an activity, even when these core features are present, should not be considered a test unless it is designated as such. Teachers often administer tasks that reflect these features in order to prepare students for a final exam, but if their evaluation of student responses does not enter the official record, the activity does not count as a test. Similarly, a teacher who asked students to engage in an activity and then, after the fact, declared it a test would be regarded as violating an important norm. In effect, an activity becomes a test only as it is institutionally constituted as one.

We discuss each of the core features in turn, with particular attention to emerging practices that do not readily reflect the particular feature under consideration. Having considered all the features, we then discuss certain ways in which authentic assessment may be extending the traditional paradigm of testing.

A Single Time Frame of Limited Duration

An activity considered to be a test is ordinarily administered in a single session with a pre-established time limit. When a test does extend beyond a single session, it is ordinarily structured into separate subtests; that is to say, in later sessions, test takers are not allowed to work further on material covered in earlier sessions. Sometimes the time limit for a test is set with the expectation that not everyone will be able to finish the prescribed tasks. Indeed, in the case of many high stakes tests, the time limit is set so that only a few will be able to finish all the tasks. In effect, time itself becomes a means of maximizing differences in student performance.

Other kinds of tests attempt to neutralize the effects of time by deliberately setting a time limit that will enable most students to finish without feeling undue pressure. For many tests, however, a time limit is imposed on the basis of administrative convenience rather than educational considerations. We know that it is unrealistic to put students under time pressure when assessing their writing skill, but our most practical alternative is often to ask them to write a finished essay in a fifty-minute class period. In this way, we can monitor their writing to make certain that it is their own.

Bloom (1976) has pointed out that schools tend to be organized so that all students have the same amount of time to learn the material, with the result that there are large differences in achievement. He has suggested that most students would achieve more in a school that is structured so that there can be flexible inputs of time. As educators have

sought more authentic modes of learning, they have moved away from tests that deal with small bits of content and skill toward evaluations of more complex activities. As assessment goals become more complex, it is often unrealistic to maintain tight time frames. If we ask students, for example, to write about effective policies for dealing with the homeless, we cannot expect to get very effective results in a fifty-minute essay. One attempt to solve this problem at the college level has been the growing use of take-home exams, in which students are allowed an extended period—typically from as little as twenty-four hours to as much as a week or two—to respond to prescribed tasks. These activities are often referred to as exams, but we might question whether they are not better conceived of as projects or term papers for which less than the usual amount of time has been allowed. For those who honor the traditional testing paradigm, the major difficulty with allowing an assessment activity to extend beyond a single setting is that it can no longer be monitored.

Prescribed Tasks in Stable Form

A test typically requires all test takers to do the same set of tasks since this insures that they are all treated equally. In certain instances, however, it may supply a list of tasks from which students can choose. Writing tests, for example, may allow students to select from among three or four topics. Since content is not deemed crucial to judging skill in writing, it is felt that students are better able to show what they can do if they are not forced to write on a subject they find uncongenial. Even tests that deal with content can allow choice so that students will have some leeway in handling gaps in preparation or lapses in memory. A history test, for instance, may require students to write only one or two short essays but provide a fairly broad range of topics.

Some new forms of assessment go beyond providing choice and ask students to construct the tasks to which they will respond. In vocationally–oriented assessment, students may be asked to decide what is the most important question that they need to face as they plan what to do after high school. They are then expected to respond to this question, and their performance is evaluated on the quality of both the task and the response. That is to say, students can be rewarded for formulating an important question, no matter what kind of response they make. This form of assessment is particularly valued by those who consider the ability to ask productive questions as fundamental to critical thinking. Making question formation a part of assessment is an effective way of putting this skill on the educational agenda.

Traditionally, tests have used a written format to insure that test takers are treated equally. Not only does writing make certain that the same words are used for each task, but it also avoids the vicissitudes of spoken language. Even in foreign language testing, where oral tasks are recorded to guarantee stability of pronunciation, problems such as machine malfunction or environmental noise can affect playback and introduce undesirable variation in the way that the task is experienced by test takers.

When we look at recent trends in assessment, we discover that oral exams, despite the concern for task stability, are becoming increasingly prominent. Many educational reform movements have focused on the skills that students need to function well within the larger society, and educators and employers have agreed that oral skills have been neglected in formal education (Natriello, 1989). Certainly, conventional testing has virtually ignored this area. Some of the recent oral exams have utilized tasks that are relatively stable. The Department of Education and Science in England and Wales (1989), for example, has developed tasks in which students listen to a recorded explanation of a diagram (e.g., a tape about language and the brain, based on a diagram of the brain) and then explain what they have learned to a student who has not heard the tape.

In some forms of oral testing, there has been less concern for task stability. For example, members who serve on a jury may question students on particular topics. While they might have specific questions in mind at the beginning, they are forced as the discussion develops to improvise if they are to ascertain the degree to which a student has thought critically about a topic. In effect, rather than presenting a stable task and evaluating a range of performances, these oral formats seek to encourage a desired level of performance by varying the task as necessary. Indeed, a major rationale for oral exams is that they provide an opportunity to probe whether students can think on their feet and express their thinking in socially appropriate discourse.

Individual Response without External Support

That each individual responds independently is basic to the notion of what a test is. If test takers are unknown to the person administering the test, there is often a system to insure that there are no impostors (e.g., an identification card with a recent photo). Precautions are also taken during the test so that students cannot make use of one another's work or have access to notes or other materials.

In recent years, there has been a trend toward allowing the use of external support during testing. Students taking writing tests are often encouraged to bring a dictionary, and perhaps even a handbook, since these are thought of as essential tools for writers in the real world. Mathematics exams often permit the use of calculators for similar reasons. Since calculators enable students to carry out calculations quickly, an additional advantage to using them is that more time can be devoted to mathematical reasoning.

Recent developments in testing greatly extend the range of tools that test takers can use. For example, students, even at the secondary level, have increasing access to a computer as they respond to various tests: in some instances they simply use a word-processing program to facilitate writing; in other instances, they manipulate a database in order to solve a problem (e.g., "What have been the major patterns of weather change in the United States during this century?"). This increase in the use of tools is accompanied by a new rationale: testing must reflect problem-solving activities in the real world and allow for, indeed encourage, the various tools used in these activities. In effect, a fundamental goal of alternative testing is to assess how well students can use technical support in problem–solving.

In real-world problem–solving, however, social support is even more important than technical support. Hence, assessment procedures are increasingly designed to measure the degree to which individuals are able to use the help of others when solving a task. In Feuerstein's model of dynamic assessment (Lidz, 1987), test takers receive an increasing amount of help as they demonstrate a need for it. Within this model, a well-designed task is one in which help is available in increasing increments.

The Department of Education and Science in England and Wales (1989) has developed such a model of assessment at the secondary level. The following excerpts from the test giver's manual give an idea of how this sort of assessment is carried out:

1. Ask: "What is the perimeter of a rectangle?"

2. Present sheet with rectangle ABCD.
   Ask: "Could you show me the perimeter of this rectangle?"
   [If necessary, teach.]

3. Ask: "How would you measure the perimeter of the rectangle?"
   [If necessary, prompt for full procedure.]
   [If necessary, teach] (pp. 33-34)

As can be seen, the task is designed so that students receive a prompt only if they need it. The score that students receive is determined by how much help they require. The six-step evaluation scheme ranges from "unaided success" to "an unsuccessful response despite prompting and teaching."

Other approaches go even further in determining how well an individual is able to use social resources. In an approach known as *collaborative testing,* two or more students are asked to work together on tasks. The Massachusetts Educational Assessment Program assesses how well pairs of eighth-grade students can apply mathematical and scientific concepts by asking students to evaluate how well three different kinds of cups maintain the heat of what they contain. In observing students carry out this task, the test administrator evaluates not only their knowledge of the principles of insulation but how well they work together. The evaluation form includes a category called "attitude" and such subcategories as "willing to cooperate" and "listens to others' ideas and responds."

A Pre-Established Evaluation Scheme

Fundamental to the basic paradigm of testing is a scheme for evaluating student response. At one end of a continuum is the multiple-choice test in which target answers are so well specified that marking can be done by a machine. Such specificity is not achievable with other test formats, but this does not mean that they cannot be evaluated with a certain degree of rigor. Even though the School Certificate exams in England make use of open-ended tasks, evaluation is fairly well-defined. Those evaluating student performance are given a marking scheme which specifies the content of an appropriate response for each task; they then assess each individual's performance with respect to how closely it corresponds to what the scheme specifies. In the case of short-answer tasks such as those used to assess reading comprehension, a content-oriented scheme is constructed. For example, a task such as "State three reasons why diesel tractors are preferred" is based strictly on information that the sample passage contains. Any answer that goes beyond that information is counted wrong, even if it provides a more sophisticated account of the preference for diesel engines.

Content-oriented evaluation is used even with tasks that elicit more extended writing. A task on a social studies exam might require students to write about the major effects of democracy in post-colonial Africa. Those evaluating this task would receive a marking scheme that includes the essential points to be covered. Students lose credit if they

leave out any of these points or if they include additional ones; here, even more than with short-answer exams, students lose credit for independent thinking. Recent research suggests that such marking schemes introduce major problems of equity, particularly in countries where teachers who do the marking have not received adequate training (Hill & Parry, in press).

When test makers evaluate the way something is done rather than specific content, they are forced to develop a more abstract evaluation scheme. Typical is the scoring system developed by the National Assessment of Educational Progress (NAEP) to evaluate writing skills. The goal of each writing task is analyzed to develop a list of features that are important in achieving the goal. Then a scoring scale is developed for each feature. As an example, we can examine the 4-point scoring scale for *persuasiveness,* which is viewed as an important feature in certain kinds of writing:

1 – contains no reasonable argument

2 – has one or two poorly thought out arguments

3 – contains several logically thought out reasons

4 – in addition contains a number of compelling details

Once a scale for a particular feature is established, it is common to locate exemplars—in this case, texts that exemplify each of the four levels of persuasiveness. These exemplars are then used to train scorers so that they can achieve a high degree of reliability.

Current Status of the Testing Paradigm

Our discussion of core features suggests that certain practices of authentic assessment can be seen as fitting within a testing paradigm. For example, with respect to evaluation, the shift toward lists of criteria has major consequences, but these are mostly confined to those who do the actual evaluation. Test takers and the public generally accept test scores without being overly curious about how they are arrived at. It is worth noting, though, that one of the tenets of authentic assessment is that evaluation criteria should be known and understood in advance, and we can anticipate that there will be considerable difficulty in making the workings of abstract evaluation schemes seem clear and fair to test takers. Even if they come to think of evaluation as more or less arbitrary, this is not likely

to affect their view of what constitutes a test. Similarly, assessment activities that present tasks orally or use questions that have not been formulated in advance will certainly seem novel at first, but the activities that include them are likely to be regarded as tests as long as other core features of the paradigm are maintained.

Hence the features most crucial to a test paradigm are probably limited duration and the unavailability of outside help. Assessment activities that extend beyond a fairly brief period are likely to be seen as projects rather than tests. Perhaps this is because the emotions and attitudes associated with testing are intense and cannot persist over a prolonged period. Likewise, collaborative testing and other assessment activities where one has the advice and support of at least one other person are not easily accommodated to the testing paradigm. A sense of being totally alone—which for many test takers leads to feelings of isolation and vulnerability—would seem to be a defining characteristic of taking a test. We should note that technical support can be more readily accommodated than social support; it is increasingly common to allow tools such as a calculator in a testing situation, particularly if it is oriented to real-world problems.

We have provided only preliminary discussion of a complex set of issues that calls for more careful attention than we can provide here. In preparing this document, we have noted how unstable the term "testing" has become in describing assessment activities. The OTA report, for example, is entitled *Testing in American Schools*, even though it discusses a broad range of activities that go well beyond a testing paradigm. The report itself does introduce a more refined set of descriptive terms, but we were not always able to locate the activities they describe in relation to a traditional testing paradigm. The term "performance assessment," for example, was used to describe a broad range of activities: some of them clearly fell within a testing paradigm while others just as clearly fell outside it. We suspect that this terminological instability is symptomatic of the fact that the testing paradigm is undergoing substantial change. Clearly, the term "test" is increasingly applied to a greater range of assessment practices. Consider the phrase "collaborative testing," which when viewed from the perspective of the traditional paradigm is a contradiction in terms since a test is an activity that an individual carries out alone. We suspect, however, that evaluating individuals as they participate in social activity will play an increasingly crucial role in expanding our notion of what constitutes a test. As we become more aware of the deeply social character of individual knowledge and skills, we will seek to measure what an

individual knows and can do in a context (Gee, 1992a). Indeed, the presence of such a context may come to be viewed as fundamental to a new paradigm of assessment.

An Overview of the Second and Third Sections

Before proceeding to the next section, we would like to give a sense of what lies ahead. In the next section, we explore various practices of authentic assessment that are being used in secondary education. We develop a working taxonomy that makes a basic distinction between practices that remain within a testing paradigm and those that extend beyond it. Our focus throughout is the American high school, though we do consider assessment practices in other countries as they illuminate what is going on in this country. In addition, we give considerable attention to assessment practices based on a more exogenous model of education, particularly one that seeks to develop an appropriate relation to the changing workplace. In certain instances, we deal with vocational schools, especially those concerned with developing more generic skills for the workplace; in other instances, we deal with comprehensive secondary schools that are developing closer relations to the workplace.

As we develop our taxonomy, we analyze the strengths and weaknesses of actual practices in relation to the purposes they are designed to serve. Since some practices have been implemented more than others, these analyses vary a good deal in the amount of detail. Moreover, the implementation varies a good deal in the degree to which it has been documented. A major problem with authentic assessment is its lack of adequate documentation; there is a good deal of anecdotal reporting but little in the way of systematic description. We have not hesitated to draw upon our own research where it facilitates a greater depth of analysis. Some of this research deals with authentic assessment in early childhood education and adult education. In both of these areas, there has been more experimentation with assessment than in secondary education. It is as if these areas, one at the very beginning of school and the other beyond it, are less constrained by normative expectations about how education is to be carried out.

In the final section, our perspective becomes more policy-oriented as we draw upon three basic principles—excellence, equity, and efficiency—to assess the major characteristics of authentic assessment. We are particularly concerned with using these

principles in comparing such assessment with conventional testing. We end this section—and the report—by examining testing and assessment from a more comprehensive perspective that we characterize as "ecological." Our concern is to examine various testing and assessment practices in relation to how well they facilitate the fundamental goals of education.

EXPLORING AUTHENTIC ASSESSMENT

A Taxonomic Overview

We begin this section by introducing heuristically oriented categories that we have used to explore authentic assessment in secondary education. In the previous section, we distinguished between emerging practices that remain within the testing paradigm and those that extend beyond it (we pointed out that the paradigm is undergoing change, so it is not easy to determine whether certain activities remain within it). This distinction is basic to any overview of authentic assessment, so we have established two broad categories that we describe as *alternative testing* and *documentation practices*. With respect to the first, we use the word *alternative* in spite of the fact that it leads some people to assume a less rigorous form of testing; as we hope to show, alternative testing is, in principle, more demanding than conventional testing. However, the term does accurately convey the notion that such testing is designed to replace conventional testing which is still dominant within American schools. With respect to the second term, we have borrowed from Chittenden and Courtney (1989) who use the word *documentation* to describe assessment methods in early childhood education that seek to formalize the classroom practices of good teachers.

In Table 3, we display two further levels of categorization that describe various practices developed by advocates of authentic assessment. With respect to alternative testing, we use the terms "discourse testing" and "performance testing" to distinguish testing that calls for discourse alone (e.g., a written test) from testing that calls for performance that goes beyond discourse (e.g., retrieving information from a computer data base). We adopt the widely used term "performance testing" with a certain reluctance since the word "performance" suggests a somewhat arbitrary display of knowledge and skills. This connotation is unfortunate in an assessment movement that is primarily concerned

with authenticity. At the next level, we use the two basic forms of language—speech and writing—to establish two kinds of discourse testing. We then draw on the element of discourse itself to distinguish two kinds of performance testing. Discourse is intrinsic to certain activities but not to others. A science experiment, for example, requires a lab write-up in order to be complete; whereas, a simple activity such as estimating the weight of an iron bar can, in principle, be conducted without discourse. Of course, testing itself imposes a form of discourse by requiring that a response be made public, but such discourse is not intrinsic to the activity itself.[1]

Table 3

Major Categories of Authentic Assessment

ALTERNATIVE TESTING			
Discourse Testing		Performance Testing	
WRITTEN DISCOURSE	ORAL DISCOURSE	INTRINSIC DISCOURSE	NO INTRINSIC DISCOURSE

DOCUMENTATION PRACTICES			
Work Samples		Records	
PORTFOLIOS	EXHIBITIONS	SYSTEM-ORIENTED	CLASSROOM-ORIENTED

With respect to documentation practices, we distinguish between actual samples of student work and the records that are kept about them. In the case of work samples, we distinguish between portfolios and exhibitions. We use the term "portfolio" in a restricted sense, which we later explain, to describe student work that involves discourse, and the term "exhibition" to describe two kinds of student work, that embodied in artifacts and that expressed in performances. By way of contrast, our elaboration of the category *records* is more functionally based. We return to the basic distinction we established when categorizing the main purposes of assessment: we use "classroom-oriented" to describe records used in managing student learning, "system-oriented" to describe records used in monitoring an educational system or in evaluating students for some institutional purpose.

[1]We are here taking a broad view of discourse (i.e., a product whose meaning is transmitted through symbols, mainly words, but does not exclude images or numbers).

A Preliminary Caveat

Before turning to our discussion of alternative testing, we would like to take a look at a current trend in conventional testing. It is disconcerting, though perhaps not surprising, to find standardized multiple-choice tests advertised as constituting authentic forms of assessment. Expressions such as "authenticity," "higher-order skills," "critical thinking," and "real-world processes" are widely used in the advertising brochures and manuals that accompany these tests. What is the status of such claims? Are they mere hype, or do they represent real changes in conventional testing that should be given a hearing?

It is interesting to find a test that is not marketed competitively, the High School Proficiency Test (HSPT)—developed and administered by the New Jersey Department Education (1990)—to assess math, reading, and writing skills, making such claims. According to the manual, the HSPT, when compared to standardized tests previously used in the state, can be viewed as measuring more complex kinds of skills. With respect to the writing subtest, the manual claims that it deals with the

> recursive process of writing [that] activates several dimensions of thinking. In the process of writing any set of words in a composition, the hierarchical sub-processes of the act require a review of what already has been written. (p. 7)

To insure that such review takes place, the test makers set up various contexts in which students revise and edit written material; this assumes that the recursive processes used during revision are an effective proxy for those that occur during composition. Here is a context that the manual includes in its sample material for students at the eleventh-grade level:

> Vera has written this biographical essay. You have been paired with her for a revising and editing conference. Read the essay critically, and be prepared to make suggestions to improve the text in organization, sentence structure, grammar, and usage. Feel free to write in the text as you read, revise, and edit. (p. 16)

These introductory remarks are followed by Vera's essay—twenty-two numbered lines that present Thomas Edison as "an improver as well as an inventor." The essay is followed by eight tasks, each designed to measure a specific skill in revising or editing

(these skills range from punctuation to discourse cohesion). Consider, for example, the fifth task:

> 5. In line 14, which of the following would be the best revision for the phrase "adequately provided . . . for"?
>
> A. brightened
> B. shined
> C. glowed
> D. Make no change

In order to deal with this task, students have to return to line 14, locate the phrase that has been elliptically represented in the task stem, then test out each of the first three choices in its place. Here is the sentence in which the elided phrase occurs (we have placed the key phrase in italics):

> In 1879, Edison produced the incandescent bulb, an improvement of a Russian engineer's design for the arc lights that *adequately provided abundant illumination for* the streets of Paris for the past twelve years.

As can be seen, the key phrase is buried within a larger sentence and, hence, can be quite difficult to locate.[2] Having located the phrase, students must then hold its discourse frame constant while they work through the various choices. It is not surprising if they experience cognitive dissonance in carrying out this task. To begin with, the frame itself is complex and difficult to hold in place, but the key phrase is also difficult to work with since it is elliptically represented in the task stem. We are puzzled as to why an ellipsis is used to replace only two words; such a small saving of textual space leads to a rather large increase in task difficulty. Complicating all of this is a semantic contradiction between *adequately* and *abundant;* indeed, the motivation for using the ellipsis may have been to disguise this contradiction.

To determine to what degree students draw on their ordinary revising and editing strategies in responding to such tasks, Coyle (1992) asked forty high school students (1) to edit the test passages independently of the multiple-choice tasks, then after a substantial time delay, (2) to respond to the multiple-choice tasks and explain their responses. With respect to task 5, one-half of the students selected the distracter "Make no change"; yet a little over half of these students had edited the phrase in question when responding to the

[2]It is part of a relative clause that modifies the direct object of a nominalized clause which is itself the direct object in another nominalization: in other words, the phrase to be focused on is at the fourth level of the grammatical hierarchy.

passage without the multiple-choice tasks. Moreover, their editing decisions had been well-targeted. One student removed *adequately*, presumably to avoid overlap with *abundant;* another replaced the latter with *sufficient.* One student even deleted the key phrase entirely and opted for a separate sentence to characterize how the arc lights had been used: "This had been the light that was used in the streets of Paris for 12 years." His explanation of his response to task 5 contains a note of frustration:

> I choose "make no change" because A, B, and C make no sense in the sentence. It seems totally wrong.

Why didn't this student select the target response "brightened"? His editing of the passage contains a clue. He has noticed that the sentence is not grammatical unless the final verb is in the past perfect:

> In 1879, Edison produced the incandescent bulb, an improvement of a Russian engineer's design for the arc lights that [had] *brightened* the streets of Paris for the past twelve years.

It appears that the complexities of this task overwhelmed the test makers themselves; they ended up using a text whose own deficiencies were confusing to students. Added to such deficiencies is the multiple-choice format that short-circuits the ordinary processes of revision that students draw on.

Hill and Parry (1988) point out similar problems with the revised edition of the *Test of Adult Basic Education* (TABE) (1987), the most widely used test in adult education in the United States. The new version of the test is advertised as incorporating authentic material from everyday life, such as letters and advertising in the tests reading component. The *Examiner's Manual* for the TABE contains a diagnostic instrument, called the *Objectives Mastery Summary*, that is designed to measure whether individual students have mastered specific skills. With respect to the reading component, the instrument includes skills such as "analyzing character (feeling, motive, trait)" and "interpreting events (conclusion, cause and effect, outcome)" (p. 2).

Such categories are however, misleading, since they are based only on the passage content rather than on what students must do with that content. Consider the following task, classified under "interpreting events (cause and effect)":

Calvin Simmons was called the "Maestro Kid" because at age eleven he

A. conducted a chorus
B. led the church choir
C. lived in the opera house
D. was an accomplished imitator

The target response is "conducted a chorus," which results in a sentence that directly parallels one in the passage:

PASSAGE SENTENCE:

He [Calvin Simmons] was called the "Maestro Kid" because, by the age of eleven, he was conducting the San Francisco Boys' Chorus.

COMPLETED TASK SENTENCE:

Calvin Simmons was called the "Maestro Kid" because at age eleven he conducted a chorus.

As can be seen, this task involves no real use of cause and effect in interpreting events; it simply recycles material that includes a causal connector.

The classification scheme is especially confusing since it does include a category described as *recalling detail*. Certain recycling tasks that parallel the one above are placed in this category, particularly if they do not have any overt content that deals with the various features in the character or event categories. In effect, the recalling-detail category functions as a kind of dumping ground for leftover tasks when, in reality, nearly all the tasks call for essentially the same thing—the recycling of textual detail.

These low-level tasks can easily mislead inexperienced test takers since the new edition of the TABE, unlike the older editions, uses passage material—letters, advertising, and even poetry—that invites a more personal style of interaction; when these tasks include a distracter congruent with this style, inexperienced test takers often fail to understand that the point of the task is simply to recycle a detail (particularly when, as is often the case, that detail has no major consequence within the passage).

Finally, Hill and Parry (1988) point out a generic flaw in the diagnostic instruments that often accompany the new tests (the HSPT also includes such an instrument):

> It is the target response alone that controls how a given task is classified; and yet anyone who has worked with the multiple-choice format knows that the real demands of a given task are anchored in the total configuration of

> choices (i.e., the reader has to work through each choice, evaluate what its particular demands are, and then come up with some way of evaluating which choice is, in fact, the most appropriate one). (p. 31)

In effect, the real point of such tasks is at a metacognitive level—what tactics are used to select a single choice from among competing ones? Or to use a more familiar term, it is "testwiseness" that these tasks measure. The new tests end up favoring the testwise student even more than the older ones did since inexperienced test takers, given the communicatively oriented material, are more attracted to a task distracter that invites a higher-level response (Aronowitz, 1984).

To sum up, conventional testing is not well-designed to carry out diagnostic work with individual students; indeed, such testing fulfills classroom-oriented functions much more poorly than the emerging practices of authentic assessment fulfill system-oriented functions. Given the problems that we have delineated, we do not view a multiple-choice test, no matter how strong its claims, as authentic; the very machinery of such a test obstructs the real purposes of authentic assessment. Teachers and administrators at all levels need to be wary of the claims that accompany multiple-choice tests. As test makers rush to join the movement for greater authenticity in assessment, they often end up constructing a test that is more dysfunctional than a conventional one.[3]

Alternative Testing

Discourse Testing

As already indicated, our basic distinction in alternative testing is between discourse testing and performance testing. The first, as its name implies, elicits only discourse, and the second is an activity that goes beyond discourse. We initially deal with discourse testing, where we make a fundamental distinction between oral and written forms. In our

[3] Certain test publishers are now experimenting with a hybrid form of testing in which the multiple-choice format is retained but expanded so that students include explanatory discourse about each choice that they make. This expanded format leads to various possibilities in scoring: choice and explanation can be scored separately; they can be scored together; or there can even be a system where the explanatory discourse is scored only when a student misses the target response (this last option is the most efficient one for the test makers). To our mind, this approach, no matter which scoring system is used, is flawed. Rather than students preparing discourse around a real problem, they focus on an artificial one, namely, how they went about deciding which of four choices is the correct one. Hence, such testing focuses even more attention on an activity that has little or no value beyond the world of testing. To our way of thinking, students would be better prepared for real-world problems that call for multiple perspectives.

earlier discussion of the testing paradigm, we pointed out that the rationale for a written test has to do with the greater stability of tasks. Writing also insures that student responses will be more stable and, thus, easier to evaluate. Even though writing stabilizes both task and response, those who construct alternative tests are willing to sacrifice such stability for other values. From their perspective, a written test is unable to deal with the dynamic aspects of discourse competence that are fundamental to everyday use of oral language. Oral testing is thus used to probe not only how students think about a particular subject but their capacity to adapt their thinking to the exigencies of social interaction. When we discuss oral testing, we will be concerned with these broader interactive capacities.

Written Discourse

The use of written discourse in testing has a long history. Its origins can be traced all the way back to the second century B.C. when China introduced a national exam in order to select those who would serve in the civil service. Promising candidates underwent lengthy preparation for this exam built around Confucian literature and were required to respond in a rhetorical form known as *baguwen*, which was viewed as reflecting Confucian values of balance and harmony.[4] In the eighteenth century, European thinkers such as Rousseau, Voltaire, and Adam Smith were strongly attracted to the Chinese model of national exams and argued for its value in developing excellence and equity in education (Madaus & Kellaghan, 1991). As a consequence, such exams, built largely around written essays, were introduced into education systems within European countries and have been in use for more than two hundred years. Even today, students who wish to obtain the Baccalaureate in France or the Abitur in Germany produce written discourse on specified tasks (Eckstein & Noah, 1989).

Within the United States, written discourse was the dominant mode of testing until it was replaced by conventional testing in the early part of this century. After many years of disuse, written discourse has now resurfaced, even in testing that is system-oriented.[5]

---

[4] Ingulsrud (1988) has shown how these Confucian qualities, initially sought in those who sat for the imperial exams, are still valued in China as well as in neighboring Asian countries. He further points out that Western readers often find excess and indirection in such discourse and tend to evaluate it negatively. Such cultural relativity raises serious questions of equity when writing samples are evaluated on an international scale (e.g., the International Baccalaureate includes samples from students in countries throughout the world).

[5] During this period of disuse, written discourse continued to be used in classroom testing; even then, however, its role was diminished because of the powerful influences of conventional testing on classroom practice.

As for monitoring an educational system, the National Assessment of Educational Progress uses a limited set of writing tasks to assess general abilities of representative students (e.g., a time-limited expository essay on a pre-established topic). In evaluating individual students, advanced placement tests use the essay form to measure knowledge and skills in specific content areas. Students are expected to follow a traditional format in writing an expository essay. In literature, for example, they deal with such elements as character, plot, theme, and point of view as they discuss particular works.

The renewed focus on written discourse has been accompanied by greater use of real-world tasks, which are often imported directly from the workplace. The use of these tasks is often justified by the use of terms such as *authentic* or *contextualized.* These terms suggest that test takers should not carry out tasks simply to display particular bits of knowledge or skill; rather they should do tasks that call for the complex thinking and doing that they use when engaged with real problems.

Real-world tasks have become strong candidates for the national exams now being proposed. Here, for example, is a task developed for this purpose by a team of high school teachers from Arkansas, Colorado, Connecticut, and Texas:

> You work in the purchasing department of the PERK corporation. You are given the responsibility to write a report supporting your recommendation on whether to buy or lease new cars for PERK executives.
>
> The best negotiated price on a 1991 Cadillac DeVille is $32,686. You are able to finance the car at an interest rate of 10.5 percent over four years with a 5 percent down payment. The same car can be leased for $499 a month over 60 months with an option to buy that car for $12,224.56 at the end of the fifth year.
>
> Include these factors in your report and as many others as you feel are important in making your recommendation. (cited in Chira, 1991, p. A19)

This task represents a familiar pattern in alternative testing: a context is set up that students must imagine themselves working within; and given an increased emphasis on establishing connections between education and the workplace, this context often has to do with an imagined job. Here students must view themselves as a purchasing agent who has been asked to write a report "about whether to buy or lease new cars for executives."

At the heart of this task is a fairly standard word problem in which the costs of two courses of action are given and students are asked to determine which is cheaper. What is different is that the response is to be in the form of a discourse rather than a calculation. If this were a traditional word problem, students could expect that the figures would be set up in such a way that there would be a clear right answer. Since they are, in fact, asked for a reasoned recommendation, they might anticipate that the numbers themselves will not lead to a clear answer and that we will have to bring in other factors. In this sense, alternative testing does provide an important corrective to conventional testing; it leads to the more open inquiry that characterizes our real-world thinking and doing.

The calculations required seem simple enough. Students are given the cost of leasing fairly directly; whereas for buying, they are given only a few facts. Presumably, if they know how to interpret these facts and have the requisite background knowledge, they can arrive at a cost for buying that will enable them to compare it to the cost of leasing. The cost of leasing can be calculated as follows:

$$\$499 \times 12 \text{ months} \times 5 \text{ years} = \$29,940$$

Owning the car at the end of five years, as indicated by the passage, would require an additional $12,224.56 which makes a total of $42,164.56.

The cost of buying the car, on the other hand, would be a downpayment of $1,634.30 plus $31,051.70 balance plus "10.5 percent over four years." How should students go about calculating this interest? Most schools teach the mathematics of simple interest, but most students quickly learn that interest calculations are seldom simple in the real world. Interest compounded daily, for example, has become common, but no bank figures it out with paper and pencil. In fact, the calculation of interest is so complicated, and consequently so open to abuse, that the Federal government from time to time promulgates regulations that require all lenders to calculate rates in a particular way so that consumers will be in a position to compare them.

In addition to the ambiguities surrounding the calculation of interest, the task does not make clear whether the loan is to be paid back at the end of four years or this is an installment loan, which would typically be paid in equal monthly payments. Again, the mathematics of installment payments are beyond the capacities of an amateur with access

only to paper and pencil. In order to calculate such payments, those working in corporate finance make use of the following kind of formula:

$$P \times \frac{i}{1 - \frac{1}{(1-i)^n}}$$

P = principal
i = rate per month
n = number of payments

This is obviously not a formula that most people carry around in their heads. We presented this task to high school students in two math classes and discovered that none of them made an attempt to calculate compound interest. Rather, most of them assumed simple interest and that the loan was to be paid back only at the end of the four years. Hence they carried out the following kind of calculation:

$$31{,}051.70 \times 10.5\% \times 4 \text{ years} = 13{,}041.71$$

Since this task was designed to achieve workplace authenticity, we might ask how real purchasing agents would handle it. If they were old fashioned, they would probably reach for a book of tables giving monthly payments for various amounts and interest rates and times periods, and find that $30,000 borrowed at 10.5 percent for four years would require monthly payments of $768.10. This would, in effect, reduce the cost of buying to about $39,600 (this lower cost is, of course, related to the obvious but often ignored fact that the principal on which the interest is paid is continuously decreasing over the four-year period). More up-to-date purchasing agents would simply use a Hewlett-Packard business calculator which has all the cost accounting formulas built in.

Using such a reduced figure to compare leasing and buying, we can see that at the end of the five years a buyer would have spent about $10,000 more (nearly $40,000 rather than nearly $30,000) but would, in fact, own a car worth about $12,000. Since these figures are fairly close, we should look for additional factors upon which to base our recommendation. One notable difference between leasing and buying is that the leaser would have an additional $269.10 a month to invest in the business which could earn perhaps ten percent or even more before taxes. A purchasing agent with a Hewlett-Packard would be able to arrive at an estimate of these opportunity costs fairly quickly. This estimated additional revenue would probably be sufficient to tip the decision in the direction of leasing.

Another consideration would be taxes; the cost of leasing would be deductible. As for buying, interest costs could be deductible but not repayment of principal, though some of it could be recovered through depreciation. Still another consideration would be the estimated cost of buying versus leasing at the end of each year and the cost of getting out of either arrangement. One can imagine, for instance, that the high-rollers who work for the PERK Corporation will not regard a four- or five-year old Cadillac as much of a perk and they may want to bail out of either arrangement fairly early on. Other factors to consider are insurance and the cost of maintenance and repairs, but the task doesn't say who is to pay for these.

Those who developed this task were attempting to present the kind of problem-solving typical of the workplace. In that case, though, shouldn't they have included the monthly cost of buying so that students could concentrate on the actual problem-solving? And if they were trying to achieve authenticity, why did they indulge in the pedagogical whimsy represented by the corporate name and the much ado about Cadillacs? Wouldn't the task have been more authentic, though perhaps less colorful, if it had dealt with the best way to obtain Ford Escorts for the sales force?

Given all these considerations, we are led to wonder just what the developers of this task had in mind. What exactly did they aim to assess? Were they aware of the difficulties in working with interest costs? Did they expect students to handle the formulas for figuring complex interest? How did they anticipate evaluating the range of responses it is likely to elicit? What weight should be given to fairly heavy computation as opposed to the higher order thinking represented by the recommendation? What weight would be given to a student's capacity to take on the role of a purchasing agent? These questions raise a much larger one about the current fad for simulating real-world contexts in authentic assessment—and, for that matter, in education at large. Jobs and workplaces are complicated entities, and we should not assume that material taken from these sources will remain authentic outside of the original context. When students are asked to take on a real-world role without the relevant information and the experiential base that ordinarily accompany that role, they can only simulate the expertise, which hardly leads to authenticity. Indeed, one danger of poorly conceived workplace tasks is to send the message that jobs are a good deal simpler—and much less interesting—than they really are.

Berryman and Bailey (1992) make the point that contextualized teaching and learning (i.e., bringing real-world objects and problems into the classroom) does not necessarily mean turning to the workplace. In many ways, it is preferable to contextualize instruction and assessment with material that is closer to hand—the classroom, the school, and the surrounding community. When material from the workplace is introduced into schools, it is more likely to be effective when it is used in more comprehensive ways (e.g., in an extended project rather than in a limited task on a test).

Oral Discourse

The United Kingdom has been far more innovative than the United States in developing oral testing. As early as 1975, it had developed an oral component for national monitoring of both primary and secondary education. Using random sampling techniques, representative students at three ages (eleven, thirteen, and fifteen) were selected throughout England, Wales, and Northern Ireland. Both oral and written components were used to test these students in four areas: mathematics, science, English, and foreign languages (French, German, and Spanish). As far as we know, this is the only use of an oral test to monitor education at a national level.

The oral test, unlike the written one, is administered to students on an individual basis.[6] As a consequence, it is administered to a smaller sample; in each age group, about twelve-thousand students take the written test but only about twelve hundred take the oral test. Moreover, the oral test, unlike the written test, is administered by teachers who do not work directly with the individual students. These teachers participate in a training program in which they first watch videotapes of the test being administered and then practice administering it themselves. Such training is designed to increase standardization which, as indicated in our discussion of the testing paradigm, is difficult to achieve in oral testing.

Teachers follow an interview schedule which specifies not only the questions to be asked but certain answers that are anticipated. Such precoding of answers allows the test administrator simply to check off a predictable response and thus concentrate on recording the less predictable. At the same time, the interview schedule allows for, indeed, encourages, a certain flexibility; a test administrator is free, for example, to ask for clarification as students explain what they are doing. Through skillful intervention, a test

[6] In science, it can be administered to a small group of students who move through workstations together.

administrator is able to elicit crucial information that a written test simply would not deal with.[7]

In the area of English language, each student carries out activities that are selected from five broad categories:

1. Providing instructions or directions
2. Giving and interpreting information
3. Telling a story
4. Describing or specifying
5. Participating in a discussion (either for a collaborative purpose or to justify an individual point of view)

The student does not communicate directly with the test administrator but rather with another student who lacks the information being communicated. This condition encourages a more genuine exchange, though the testing situation clearly constrains what goes on. A larger group of students is present for activities such as discussion, in which case the exchange, given the right topic, can become quite lively. The administrator's role is simply to establish a context for talk rather than to direct it or control it. In addition, the person audiotapes the oral communication and provides an immediate evaluation that concentrates on two fundamental aspects of the student's performance: (1) what was said and (2) orientation to the other(s). This second aspect takes account of how well the student listens as well as speaks. As Widdowson (1978) has pointed out, listening and speaking ultimately cannot be separated when evaluating oral competence; the two are simultaneously at work as we engage in talk in the real world. The concurrent evaluation, necessarily impressionistic, is followed by a more detailed evaluation that a group of specially trained teachers carries out as they listen to the audiotape. These two kinds of evaluation thus provide complementary means of assessing the student's oral competence.

In the area of foreign language, an oral component to testing has been particularly welcome. In the absence of such a component, testing unduly emphasizes written

[7] Before the oral test begins, the administrator tells the student that such intervention does not indicate that a particular answer is wrong; rather, it is just a way of finding out more about the thinking that lies behind the answer.

language, testing which has the unfortunate consequence of reinforcing a grammar-and-translation method in foreign language instruction. Here, as in English, the oral test is built around a functional use of language. The tasks have to be carefully planned to insure that students have the requisite skills to carry them out. As Burstall (1986) points out,

> We have concentrated on the extent to which students are able to function competently in a realistic communicative setting. We have adhered throughout to the principle of positive scoring—that is, scoring criteria that give credit for even the most modest level of achievement and that do not penalize students for what they are unable to do. This break from tradition in foreign language testing has been welcomed by foreign language teachers weary of deducting marks for each small error. (p. 21)

One of the important strengths of authentic assessment is its capacity to elicit from students knowledge and skills that are bypassed in conventional testing.

Many educators in the United Kingdom have been enthusiastic about the role of oral testing in the national program of monitoring education. Within all four subject areas, it has provided greater understanding about what students know and can do; of particular importance is its capacity to probe the student thinking that goes into a particular response. Yet, this testing is increasingly faced with difficulties. Economic retrenchment in the United Kingdom has put severe pressure on the national program for monitoring education, and the labor-intensive demands of oral testing make it a difficult component to sustain. At the same time, the more complex aspects of what students do in oral testing have not been easy to evaluate. Educators are far more accustomed to evaluating the stable discourse that students produce in writing; whenever they are forced to evaluate oral discourse, they feel less secure in their judgments. Hence, it is not clear how effectively oral testing can serve system-oriented assessment as it is carried out on a national scale.

Performance Testing

We earlier introduced the notion of intrinsic discourse to distinguish two kinds of performance testing. We pointed out that a complex activity such as a science experiment includes discourse—the lab report—while a simpler activity such as estimating the weight of an iron bar can, in principle, be conducted without discourse. We then pointed out that testing itself imposes its own form of discourse so that the individual estimates can be made public and thus evaluated. We will describe such test discourse as *imposed* to distinguish it from the discourse that we are calling *intrinsic*.

Performance with Intrinsic Discourse

Performance testing that contains intrinsic discourse is increasingly used in vocational education as well as in academic programs oriented to the workplace. In order to simulate many different kinds of workplace activities, testing must embed some form of social exchange. Consider, for example, the following sequence of activity which represents a type of exchange that is common in the workplace:

1. the test taker answers an incoming call that requests information;
2. the test taker uses a computer to retrieve the information; and
3. the test taker communicates that information to the caller.

The first and last steps involve direct social exchange; in the first, the test taker listens to a request and in the last, responds to that request.

This sequence can be extended to accommodate even more of what takes place when someone calls for information. Consider, for example, a caller who wishes to make a plane reservation. Once the third step is carried out, the caller might then request that a reservation be made, which would lead to at least two further steps:

4. the test taker enters the reservation into the computer; and
5. the test taker provides confirmation of the reservation.

This confirmation is expressed orally; but the test taker, if simulating a real-world routine, can produce a printout to be made available to the caller.

This sequence of activity involves both a human-human interface and a human-machine interface. To perform well in these situations, an individual must not only maintain the two interfaces but transmit information efficiently between them. In the real world, these situations reflect a complexity that is difficult to simulate in a testing situation. To illustrate, a caller often does not know exactly what to request; in the case of an airline reservation, the request is contingent upon factors such as which flights are available, whether discount fares are available, and so on. With appropriate planning, these complicating factors can be incorporated, at least in a rudimentary way, into a performance test.

Performance testing built around workplace routines can be quite difficult to evaluate. If evaluators attend to the performance concurrently, they must attend to many different kinds of behavior at once. With respect to the above performance, they would have to deal with what happens not only at the two interfaces but during the transitions between them. Concurrent evaluation can be supplemented—or even replaced—by retrospective evaluation based on a videotape. In principle, this approach makes the complexity of human behavior more manageable since a videotape allows for repeated viewings. Videotaping such a performance, however, introduces a broad range of problems. To begin with, it is a costly endeavor, particularly in the human time that it demands. Not only must the original performance be recorded, but the videotape often has to be viewed many times before an appropriate assessment can be made. As anyone knows who has worked with videotape, such repeated viewing can be quite time-consuming. Moreover, videotaping introduces many technical problems such as those having to do with adequate reproduction of the original behavior as well as subsequent storage and retrieval of information (see Erickson, 1982, for an overview of problems involved in using videotape to analyze human behavior).

Performance testing is not, of course, limited to workplace routines. It is particularly well-developed in the sciences where students need to know how to engage in a sequence of exacting behaviors. Students are often asked to perform an experiment within a testing framework. Their performance is then evaluated according to basic criteria of scientific inquiry, including the selection of relevant material and technical apparatus, appropriate ordering of steps, and exact measurement of substances.

A science experiment can offer distinct advantages as a performance test. To begin with, it can be an effective means of bridging the academic and the practical. In chemistry, students can conduct a small-scale experiment dealing with industrial pollution—for example, one verifying that water drawn from a particular lake is contaminated. Students can even be asked to report the results of the experiment to a political organization such as a city council. These more complicated demands, however, are difficult to meet in a limited time frame and, therefore, might not be feasible in performance testing. We will deal with more demanding forms of science experimentation, particularly those that call for group work, when we come to documentation practices.

Other major advantages of a science experiment in performance testing are derived from the lab report. To begin with, such a report insures faithful simulation of authentic practice since in the real world a science experiment is not considered complete until it has been written up. Moreover, the lab report facilitates evaluation since it is a well-defined genre that is relatively succinct (succinctness is, in fact, often a basic criterion used to evaluate the lab report). Finally, the report provides a natural means of introducing self-assessment. After conducting an activity, students can be asked not only to describe what they did but to evaluate how well they did it. In the case of the water pollution experiment, they might describe certain strengths and weaknesses in the procedures they followed. For many educators, the fundamental goal of assessment is to develop just this capacity for self-monitoring. Students not only must internalize socially constituted norms of good practice but must know how to use these norms in assessing their own practice. As we will see, such self-monitoring is commonly elicited in portfolio assessment, where students are provided sufficient time to evaluate samples of their work. Yet, it is increasingly elicited in a testing situation as well; if the initial activity is sufficiently brief, students can be asked to evaluate how well they have carried it out.[8]

Performance without Intrinsic Discourse

The use of workstations to elicit student performance is common in domains of science that call for hands-on activities. Rather than requiring students to engage in a single extended activity such as an experiment, workstations elicit a range of activities that sample crucial knowledge and skills. In medical schools, for example, as students move from one workstation to another, they are required to examine various diagnostic materials such as x-rays or slides. The students carry with them a test booklet that contains questions that require them to write short answers. At any one workstation the time is strictly regulated; for example, there is often a buzzer to signal that a student must move on.

Within alternative testing in secondary education, workstations have been used in a different way. In the United Kingdom they have become central to the national testing of math and science. In science, for example, students carry out activities at nine workstations where they are asked a series of questions by a test administrator. Although these questions are controlled by an interview schedule, the administrator has some

[8] Even discourse testing can, at times, include a final component where students are expected to evaluate how well they have done the task. In some instances, students first write and then discuss orally how well they have managed the writing.

flexibility in probing for what individual students are thinking as they respond to a particular question. In effect, the discourse is oral, concurrent, and nonintrinsic, as opposed to the written, retrospective, and intrinsic discourse embodied in the lab report of a science experiment.[9]

In discussing this approach, Burstall (1986) outlines a number of advantages. Since the tasks are administered orally, poor readers are not handicapped. Since students are allowed to clarify points they don't understand, they are able to respond to tasks that they would have to leave blank on a written test. In actually responding to a task, students have the opportunity to amend or even retract their answer (they often discover errors in their thinking when they try to explain an answer). Finally, this approach offers the test administrator an opportunity to observe more directly the problem-solving strategies that students use.

Despite these very real advantages, this approach, as indicated, has been difficult to implement and sustain in a national system of assessment. Given the high costs in time and money, many question it on grounds of efficiency. Those who support conventional testing often challenge performance testing on grounds of equity as well; they claim that an administrator's interventions can provide an unfair advantage to individual test takers, particularly those accustomed to a certain style of academic exchange. Since mainstream students are most experienced at such exchange, students from other ethnocultural backgrounds can be placed at a disadvantage. We are sympathetic to such problems but believe they are, at times, overstated. Our own experience with students from such backgrounds leads us to believe that they often benefit from the give-and-take of oral exchange, given their heavy reliance on it within their own speech communities. Any approach that sheds more light on student thinking can uncover distinctive powers of language, thought, and experience, which ultimately leads to more equitable assessment.

[9] Even at the kindergarten level, workstations have been successfully used in assessment activities. They are, for example, used in an early childhood assessment project in Newburgh, New York, that one of us has directed during the past three years. As children move through the workstations—or as we prefer to call them, *play stations*—they manipulate various kinds of material such as building blocks. Accompanying these manipulations are questions to which they respond orally. Some questions have to do with their capacity to estimate features such as volume or weight; others have to do with their capacity to discern geometric shape and abstract patterning. We have been pleased with children's responses to such hands-on testing; they actively handle the material as they explain how they solved the various problems.

We would, however, like to end with a word of caution. Within this approach, discourse is imposed on the activity rather than being intrinsic to it. As the administrator intervenes with a series of external questions, the person increases the risk of distracting test takers from the task they are engaged in; it is as if the fragmentation that we associate with conventional testing is reinstated. It is for this reason that we prefer performance testing in which discourse is intrinsic rather than imposed by the administrator. In this way, the integrity of the activity is preserved and the testing is able to achieve greater authenticity.

Documentation Practices

In recent years many educators have come to view authentic assessment as an enterprise altogether different from testing—whether traditional, conventional, or alternative, as we have been using these terms. They claim that testing, no matter how much it is reformed, still focuses on how students handle tasks on a single occasion under rather severe time constraints. Such constraints necessarily lead to limited tasks that cannot get at the complexity of what students do when they engage in purposeful tasks over a period of time. An extended timeframe offers the opportunity to document students working on a greater range of tasks, each inherently more complex than what they can be asked to do on a test. Hence, methods have been developed to track student involvement with multiple tasks over time. We use the term "documentation practices" to refer to such methods; that is to say, those not limited to student response to prescribed tasks on a single occasion.

It is not easy to establish a framework for describing these practices since they reflect a good deal of variation. Part of this variation can be traced to the experimentation that comes with any development of new practices; the field of authentic assessment is flooded with ideas that are being tried out in different ways. Some of the variation, however, comes from the nature of the movement itself. Once assessment is strongly committed to authentic practices, it opens up to a vast range of real-world activities that go far beyond what we ordinarily associate with schooling. Any attempt to reflect these activities in assessment practices necessarily leads to variation. In effect, authentic assessment encourages a more exogenous model of education than conventional testing does.

This variation is reflected in the language that educators use to talk about documentation practices. Consider the widely used term "portfolio"; we have come across at least three ways in which it is used. First, a portfolio can be restricted to a selection of student work that is placed in a physical folder; typically, such work consists of written discourse since it fits most readily into a folder. In a second use, the term refers to a larger collection of work samples that are not necessarily located at a single point in space. Such a collection can include artifacts and performances as well as discourse. There is a third use—even broader—that is evidenced in secondary schools such as Walden III in Racine, Wisconsin, that requires students to prepare for exit-level certification. A portfolio is everything that students put together for this rite of passage, and it ranges from collected samples of their work to various records that have been maintained about them throughout their school career. As we talked with educators and reviewed literature for this report, we discovered that it is the first view of a portfolio that is most widely held. Certainly from the perspective of a practicing teacher, it is natural to think of a portfolio as a receptacle in which representative pieces of student work can be stored and easily retrieved.

In our own categorization of documentation practices, we establish a fundamental distinction between "work samples" and "records." The first refers to what students have actually done; whereas, records are *about* what they have done. To put it another way, work samples present students as creative agents; records present them as subjects whose work is to be described. With respect to work samples, our major distinction is between *portfolios* and *exhibitions*. We use the first to refer to pieces of discourse, primarily written, and the second to artifacts and performances. As can be seen, we have adopted the most restricted conception of a portfolio. Since it is the one most widely used by educators, we have adopted it partly as a matter of convenience. This conception also reinforces the emphasis on discourse that we established in our discussion of alternative testing. In effect, portfolios can be viewed as parallel to discourse tests and exhibitions to performance tests (exhibitions also include artifacts, but they have no parallel in alternative testing since it does not provide sufficient time for students to produce an actual object).

We believe our emphasis on discourse is well-motivated since it remains the dominant mode of expression within the emerging practices of authentic assessment. At the same time, we recognize the point that Grant Wiggins (1991b) often makes: it can be unduly limiting to think of a portfolio as a physical collection of work samples—discourse or otherwise—located at a single point in space. He prefers the third conception that we

described above: everything that is collected for assessment purposes—a student's discourses, artifacts, and performances—as well as the recorded judgments that the student and selected others—peers, teachers, or external evaluators—make about what has been collected. From an intellectual perspective, this is an attractive notion, but it is so all-encompassing that it is difficult to use for taxonomic purposes.

Work Samples

Portfolios

Portfolios are the most widely used form of documentation. An increasing number of states include portfolios in a system-oriented assessment program. Alaska, California, North Carolina, Rhode Island, and Vermont all make use of portfolios, though not to the exclusion of conventional testing. In addition, Delaware, Georgia, South Carolina, and Texas have developed plans for implementing a portfolio program. Moreover, individual schools and school districts are actively experimenting with portfolio assessment at more local levels.

We would like to stress at the outset that the basic notion of portfolios is hardly new. They represent what teachers, especially at the elementary level, have long recognized as good practice—helping individual students to keep track of their own work and to present it in an organized way. What is new is the notion that portfolios can be formalized to such a degree that they can replace, or at least supplement, conventional testing. Returning to the fundamental distinction we established with respect to major purposes of assessment, we use the term "system-oriented" to refer to more formalized portfolios used to monitor educational systems and evaluate individuals for institutional purposes, the term "classroom-oriented" to refer to less formalized portfolios used to manage teaching and learning. In order to make more explicit the notion of formalization, we can also return to the core features of the testing paradigm (see pp. 8-15). The first feature—a limited time frame of specified duration—cannot, of course, be used here; the very essence of a portfolio is to break out of a single time frame and provide multiple samples of student work. We can, however, use the other three features to place various kinds of portfolios along a continuum that runs from system-oriented to classroom-oriented. We here present these features as representing the two poles of a continuum (each accompanied by a subsidiary feature germane to portfolio assessment).

Table 4
Contrasting Models of Portfolio Use

System-Oriented	Classroom-Oriented
(1) more insistence on prescribed tasks (can include timed tasks)	(1) more allowance of freely selected tasks (does not include timed tasks)
(2) more insistence on individual work (emphasizes products)	(2) more allowance of group work (allows more attention to process)
(3) more insistence on analytic evaluation (emphasizes external accountability)	(3) more allowance of holistic evaluation (allows more self-monitoring)

System-Oriented Portfolios

Portfolio assessment in Vermont is an exemplar of a system-oriented approach, for it is used not only to evaluate individual students but to monitor education within the state as a whole. In 1989, the state legislature approved the experimental use of portfolios as the primary means of assessing language arts and mathe. ics for fourth- and eighth-grade students.[10] Given our concern with secondary education, we will deal only with portfolio assessment at the eighth-grade level.

Prescribed Tasks. With respect to the first feature in Table 4, the contents of an individual portfolio are well-specified. The language arts portfolio, for example, must contain the following items:

1. a table of contents,
2. a best piece (a personal response to a book, math problem, scientific phenomenon, cultural event, or current issue),
3. a letter that explains the choice of the best piece and the process of its composition,
4. a poem, short story, play, or personal narration, and
5. three prose pieces from any curriculum area other than English or language arts.

[10] At that time, Vermont mandated a statewide system of testing as well; prior to this legislation, the state did not maintain any system of assessment in public education. According to Ross Brewer (1991), a major reason for introducing testing was to provide a means of monitoring portfolio assessment. We would like to sound a word of caution about this approach. When more authentic forms of assessment drive an educational system, students do not necessarily perform better on traditional forms of testing; indeed, quite the contrary might be true—as students become more communicatively oriented in their school work, they may do less well in conventional testing.

Individual students are not allowed to include any item that does not fit within the above categories and cannot, for example, include as their best piece a letter to a friend, a journal entry about the death of a grandparent, or a documented research paper on industrial pollution in Vermont (even though they—and even their teachers—may view such a piece as the strongest writing they have done). By the same token, they cannot write the process-oriented letter about a piece of creative writing (e.g., a poem or a short story), even though they may have especially vivid impressions about composing such a piece. Moreover, they are not allowed to include multiple samples of a particular kind of writing; for instance, students cannot lard their portfolios with poetry even though they have spent most of their time working on it. In effect, the portfolio is not designed to allow students to pursue a special talent or set of interests but rather is set up to provide some kind of balance between personal and academic forms of writing that cut across the school curriculum.

At the same time, students are encouraged to use independently generated work in building their portfolios. In a widely distributed brochure entitled *The Writing Portfolio: Your History as a Writer*, students are offered the following advice:

> Whenever you write something on your own, without a teacher's guidance, whether at home or during a free period at school, consider including that piece in your portfolio. Your portfolio can then become an opportunity for you to let your teachers know about the kinds of writing you do outside the realm of "schoolwork." (Vermont State Board of Education, 1991a, p. ?)

This advice may sound attractive, but we suspect that relatively few students include such work in their portfolios. The fact that portfolios are used for official assessment tends to discourage individual initiative. It would be useful to conduct empirical research on the degree to which students draw on independent writing in building portfolios that fulfill system-oriented functions.

The language arts portfolio in Vermont, unlike some system-oriented portfolios, does not include any timed piece of writing. In many situations where portfolios replace a conventional test (as noted earlier, Vermont maintains a supplementary test), they include writing produced under controlled conditions. In the United Kingdom, for example, the General Certificate of Secondary Education requires three timed pieces of writing alongside six pieces produced independently. Two of the timed pieces are based on reading passages accompanied by tasks (unlike conventional reading tests, certain tasks focus on inferential

and evaluative responses that require extended writing). For the other piece, students listen to a taped talk and respond to both form and content.

In this country, some writing portfolios contain an essay written in class as a final exam. Hunter College has introduced portfolio sections alongside traditional sections in its freshman writing course. The culminating entry in a five-part portfolio is a timed essay written during the final class. The primary purpose of this monitored writing is stated, almost as a warning, to students and is included to insure the authenticity of their unmonitored writing. We understand the instinct to build in this checkpoint but also recognize that many students, particularly foreign students and speakers of nonstandard dialects, experience substantial difficulties when writing under time pressure (particularly when they are aware that it is a testing situation).

Individual Work. The language arts portfolio in Vermont includes only writing that is composed by an individual student.[11] In spite of a fairly strong emphasis on individual authorship, students are allowed to show their work to teachers, peers, and parents before submitting a final copy. Indeed, they are encouraged in the process-oriented letter to describe how they have used others' help in developing what they consider their best piece of writing. When portfolios are used for purposes of external accountability, there is an inevitable conflict between individual authorship and a social orientation to writing. On the one hand, students are cautioned that their work must be their own so that the assessment system can operate fairly. On the other hand, they are encouraged to carry out work according to real-world norms and thus draw on social support. As a consequence, students are often confused about how to use help. If a parent or an older sibling supplies a good deal of felicitous rewording in the margins of an essay, can they directly incorporate it? It is this kind of question that has not been adequately addressed by those who use portfolios for the purpose of external accountability.

In certain approaches to portfolio assessment, there are mechanisms for dealing with the problem of individual authorship. One common mechanism is to require students to document the processes they go through in constructing a portfolio. In the case of a

[11] The mathematics portfolio in Vermont can contain one report of a group activity or project. This report has to be accompanied by a letter that explains each individual's contribution. When we come to consider classroom-oriented portfolios, we will see that such collaborative work is common in a portfolio whose fundamental purpose is to develop a capacity for self-monitoring.

writing portfolio, the documentation of process can take many forms such as informal notes, outlines used at various stages of writing, drafts that include others' comments, and notes on how such comments were used. However, as indicated in Table 4, system-oriented portfolios tend to focus on products rather than the processes that undergird them. The language arts portfolio in Vermont, for example, includes five products, yet students provide process-oriented commentary on only one of them (and even this commentary is not accompanied by any artifacts of the writing process). The rationale for this streamlined portfolio is an obvious one; a more process-oriented portfolio is simply too unwieldy for purposes of official assessment.

Analytic Evaluation. In Vermont the language arts portfolios are evaluated in two phases: (1) classroom teachers use a four-point scoring rubric to rate individual portfolios with respect to five criteria; and (2) a double-tiered system of moderation is used to monitor teacher evaluations.

During the first phase, teachers are trained to evaluate individual portfolios with respect to five criteria:

1. Purpose—the degree to which the writer's response
   - establishes and maintains a clear purpose
   - demonstrates an awareness of audience and task
   - exhibits clarity of ideas
2. Organization—the degree to which the writer's response illustrates
   - unity
   - coherence
3. Details—the degree to which the details are appropriate for the writer's purpose and support the main points of the writer's response
4. Voice/Tone—the degree to which the writer's response reflects personal investment and expression
5. Usage/Mechanics/Grammar—the degree to which the writer's response exhibits correct
   - usage (e.g., tense formation, agreement, word choice)
   - mechanics (e.g., spelling, capitalization, punctuation)
   - grammar
   - sentence structure

Teachers use a four-point scoring rubric which is specified for each criterion. We here provide the specification for "organization":

1. Extensively
   - organized from beginning to end
   - logical progression of ideas
   - clear focus
   - fluent, cohesive

2. Frequently
   - organized but may have minor lapses in unity and coherence
   - transitions evident
   - usually has clear focus

3. Sometimes
   - inconsistencies in unity and/or coherence
   - poor transitions
   - shift in point of view

4. Rarely
   - serious errors in organization
   - thought patterns difficult, if not impossible, to follow
   - lacks introduction and/or conclusion

In addition, a category of "nonscorable" is used in the case of a blank page (i.e., a required piece of writing was not included in the portfolio) or writing that is illegible or incoherent. These rubrics are used to score all six pieces of writing with respect to the five criteria. These six assessments are then summarized in a single score for each of the criteria.

The Vermont scoring system is analytic in the sense that it assumes that writing can be characterized by a limited number of discrete categories which can be evaluated separately. What is surprising is that the categories and the rubrics associated with them are applied irrespective of the kind of writing being evaluated. An examination of the categories shows that they are largely derived from our ideals for expository writing, yet they are applied to all forms of writing in the portfolio—including personal responses; narratives; and even imaginative forms like poems, short stories, and plays. Thus, if we were to follow the rubric for organization strictly, we might be forced to give even a poem we admire the lowest rating because it "lacks an introduction and/or conclusion."

To anchor the four levels in the scoring rubric, teachers are provided writing samples by a statewide benchmarking committee that has access to student portfolios from

past years. The committee provides twenty samples altogether: four for the criterion of "purpose," four for the criterion of "organization," and so on. Many of these exemplars are puzzling, yet there is no accompanying explanation relating specific features of a text to the way it is evaluated. Here, for example, are the first four sentences from a student essay entitled "Booker T. Washington," which the committee placed at the bottom of its scale—it was rated "rarely"—with respect to voice/tone:

> Booker T. Washington was born April 5, 1856, in Franklin County, Va. His mother, Jane Burroughs, was a plantation cook. His father was an unknown white man. When Booker was only a child, he swept yards and brought water to the slaves working in the fields.

At first glance, it is not clear why this text has been given such a low rating. It does manage to maintain a fairly consistent tone which, though not particularly vivid, seems appropriate for the task at hand—a dispassionate reporting of biographical fact. However, the voice/tone category has been set up to value a high degree of "personal investment and expression." This seems to ignore the many kinds of writing in which a strong personal stamp would usually be considered inappropriate, such as that found in encyclopedias, textbooks, recipes, reports of scientific experiments, and much workplace writing. We suspect that a bias is at work among those working in authentic assessment—the more personal the tone, the better the writing. Such a view may be understandable, given the struggle that has gone on in recent years to validate personal forms of expression in classrooms around the country. However, to neglect genres and styles that are widely evidenced and considered appropriate in the larger society is to adopt an endogenous stance that can lead to unfair assessment (for further discussion of the Vermont scoring system, see Hill, 1992).

Once teachers have completed their ratings of individual portfolios, the second phase of the evaluation scheme goes into operation. Moderators at the district level carry out independent evaluations of a random sample to verify teacher evaluations. Wherever there is a discrepancy between the two evaluations, moderators at the state level are called in to make a further independent evaluation. In addition, state-level moderators monitor not only a random sample of the evaluations made at the district level but each other's work as well.[12]

[12] This system—indeed, the term *moderation* itself—was borrowed from the United Kingdom, where it has been used at the secondary level for a number of years.

This massive scheme of external evaluation leaves little room for self-evaluation by individual students. It is true that students, in consultation with their teachers, select the piece of writing to be included. Moreover, in the letter that accompanies their best piece of writing, they can include some evaluation with respect to that piece (some students even squeeze a bit of evaluation into their table of contents). In comparison to a classroom-oriented approach, however, the Vermont system does not encourage much self-evaluation. This point is recognized by Ross Brewer (1991), who concedes that certain values associated with student learning were sacrificed in order to insure that a statewide system could operate with equity and efficiency.

We would also like to question the use of analytic scoring in system-oriented assessments of writing. The usual reason for adopting analytic scoring is that it provides diagnostic detail, but it is hard to see how the general kind of diagnosis produced by large-scale assessment can be put to work in the classroom. How, for instance, would one go about beefing up instruction in "voice/tone" or "purpose"? In a discussion of various approaches to writing assessment, White (1985) concludes that "analytic scoring is uneconomical, unreliable, pedagogically uncertain or destructive, and theoretically bankrupt" (pp. 123-124). On the other hand, to use the kinds of holistic scoring that White finds more reliable would entail major changes in approaches like that in Vermont. Programs, mostly at the college level, that have achieved reasonable reliability with holistic scoring have depended upon a precise description of the writing to be done, the development of a scoring rubric that matches the assignment, and thorough training of a team of evaluators. These factors seem to rule out approaches in which students have input into what is included in their portfolios. In recent years, however, there has been substantial progress in the development of holistic scoring, and further research may well produce solutions to current dilemmas.

Classroom-Oriented Portfolios

When portfolios are not used for purposes of official assessment, they tend to take various shapes. Such variation is in keeping with the philosophy of those who use portfolios to support classroom teaching and learning. From their perspective, portfolios must be adapted to local conditions if they are to achieve their full potentiality. It is thus difficult to focus on any one adaptation as an exemplar of classroom-oriented portfolios. We will, however, draw heavily on the Arts PROPEL Project in Pittsburgh public schools since its portfolio use contrasts sharply with the system-oriented use in Vermont. Arts

PROPEL is a project funded by the Rockefeller Foundation that focuses on three areas in the secondary school curriculum: music, visual arts, and writing (to maintain parallelism with Vermont, we will deal only with writing). Researchers from the Educational Testing Service and Harvard Project Zero have been involved in the project from its inception; their primary concern is to develop forms of assessment that are closely integrated with teaching and learning.

Freely Chosen Tasks. As a project centered in the arts, Arts PROPEL is concerned with maintaining maximum flexibility for students. They are encouraged to develop their own writing projects which often extend over a period of weeks and even months (in most system-oriented portfolios, the specification of an exact range of tasks leads to shorter pieces of writing produced in less extended time periods). Moreover, the various projects are accompanied by process notes, outlines, and working drafts; even here students are free to select the material that documents their writing processes. Out of the various projects, students are expected to develop a major piece of writing that is described as a *domain project.* For this project, they provide more extensive documentation of the various processes they have engaged in; they submit along with a final copy of the project a narrative that guides readers through various materials that document how the project developed. This concerted attention to how writing develops has led researchers associated with Arts PROPEL (Gardner, 1991; D. Wolf, 1988, 1989) to use the term *processfolio* to characterize the work that students do.

Students are also asked to prepare a writing autobiography which is initiated by their response to a set of questions given out at the beginning of the semester (these questions deal with attitudes brought to writing, particularly in its relation to reading, speaking, and listening). Although this task is more constrained than the others, students are encouraged to respond to it in a flexible way. They can, for example, use speech, writing, or a combination of the two to record their responses (they can include an audiotape and an accompanying transcript in their portfolio).[13]

[13] Classroom-oriented portfolios often include oral alongside written discourse. The state of Victoria in Australia, for example, requires that language arts portfolios include an oral as well as a written report about an independently selected piece of literature. These reports are then evaluated according to criteria appropriate to each modality (Wiggins, 1991b).

Emphasizing the importance of students deciding what writing will go into their portfolios, Roberta Camp (1990) reports what a number of teachers have to say about this freedom. As one of them put it:

> When a student decides which pieces to select, he's thinking about how to show what kind of writer he is. The portfolio then becomes more to the student that just a storage place for his writing. (p. 8)

Another teacher emphasized the importance of students feeling in control:

> Like most middle school students, I guess, they [her eighth-grade students] are trying to get command of their writing. They need to see how they can be in control. When they look at writing only one piece at a time, they can't really take ownership of it. But if all of their work is collected in one place, they can see all the pieces, can see the continuity, and can begin to see themselves in control. (p. 8)

From the perspective of these teachers, the freedom of students to plan the presentation of their own writing is crucial to successful use of portfolios. As we have observed, such freedom is difficult to maintain when portfolios are used, as they are in Vermont, for system-oriented assessment.

Collaborative Work. There are two kinds of collaboration that are elicited by classroom-oriented portfolios. The first involves individual students receiving extensive help from a teacher or peers, and the second involves two or more students jointly preparing a portfolio. This first kind of collaboration is evidenced in the Arts PROPEL project. While developing their various pieces of writing, students engage in extensive consultation with their teachers, other students, and even the researchers from the Educational Testing Service and Harvard Project Zero. Within Arts PROPEL, writing is viewed as a social process in which a writer continuously incorporates the reactions of other readers. Hence, the question does not arise, as it does with system-oriented portfolios, whether an individual student has received too much help from others. Indeed, the capacity of a student to absorb help is viewed as a strength and appropriate assessment should view it as such. Certainly in the real world, successful writers are often those who have learned how to use social support effectively (e.g., research assistants and editors). Hence, a project such as Arts PROPEL seeks to develop in students a disposition to draw actively on such support.

The second kind of collaborative work is more common in science than the humanities. In scientific fields it is increasingly common, even in system-oriented assessment, for individuals to work together on a project. Such collaboration is encouraged by various factors; perhaps chief among them is the inherent complexity of scientific investigation which calls for the highly specialized skills of different individuals. In order to socialize students to carry out joint work, portfolios in science often require at least one collaborative project. In the state of California, a science portfolio must include at least six pieces of work, one of which represents a small-group investigation. Each member of the group must participate in preparing a written report that documents the problem worked on, the difficulties encountered, and the conclusions reached.[14]

Attention to collaborative work is also present in the more broadly conceived portfolios required for graduation from high school. At the Jefferson County Open High School in Colorado (Wiggins, 1991b), students must prepare a final portfolio that covers the following areas:

- quest for meaning
- creativity
- career exploration
- global awareness
- practical skills
- logical inquiry

In covering these areas, students must work collaboratively on at least one project, and it is not surprising that collaborative projects are frequently carried out in the last area. Many students satisfy the demands of logical inquiry by working together on a science project.

Holistic Evaluation. It is with respect to evaluation that classroom-oriented portfolios can be most sharply distinguished from system-oriented ones. The classroom-oriented approach places far greater emphasis on self-evaluation. In the writing portfolio for Arts PROPEL, students are active participants in evaluating their own work. Apart from

[14] A collaborative project is also required for the mathematics portfolio. Mathematics, like imaginative writing, is often thought of as an individual activity, but it, too, involves a good deal of collaboration, especially when it is applied in various domains of science.

evaluating individual pieces of writing, they include in their autobiographical sketch an overview of the writing they managed to do during the course of the semester. In characterizing their own work, they evaluate both its strengths and weaknesses. They may analyze, for instance, why a certain piece of writing does not work even though they put a great deal of effort into it.

How students evaluate their own work is taken into account when teachers come to evaluate portfolios. This teacher evaluation is not based on any scoring rubric but proceeds holistically with attention to the following areas :

1. Range of work
   - What does this writer understand about differences in genre?
   - What does this writer understand about adapting writing to fit different circumstances?
2. Development of work
   - How effectively does this writer use criticism?
   - How effectively does this writer sustain a complex piece of writing?
3. Style
   - Is this writer's basic approach merely imitative or original?
   - How effectively can this writer play with ideas and images?
4. Craft
   - How effectively does this writer control basic forms of written English?
   - How effectively does this writer revise?
5. Audience
   - How sensitive is this writer to the reader's needs?
   - To what degree does this writer provide adequate transitions and supporting detail?
6. Reflection
   - How effectively does this writer use writing for the purposes of learning?
   - To what degree does this writer recognize his/her own strengths and weaknesses? (Wiggins, 1991b)

As can be seen, the last area focuses directly on students' capacity to evaluate their own writing. In general, classroom-oriented portfolios are less concerned with norm-referenced evaluation or, for that matter, criterion-referenced evaluation. The focus is rather on self-referenced evaluation; hence, the major concern is to determine what kind of growth the individual student achieves over time. Indeed, the use of self as a reference point rather

than other students or predetermined criteria is increasingly widespread in the movement toward authentic assessment.[15]

Exhibitions

We have reserved the term "exhibition" for the display of artifacts and performances and have thus extended its familiar use within the arts where the term ordinarily refers only to the display of artifacts such as a drawing, painting, or piece of sculpture. Our extension has been primarily motivated by parsimony; we wanted a single term to describe any personal creation, artifact, or performance which departed from what traditional forms of assessment have called for—the production of written discourse.

Artifacts

It is perhaps the arts themselves that present the greatest range of artifacts to be displayed such as drawings, paintings, murals, sculpture, photographs, collages, posters, costumes, scenery for plays, and even musical instruments. As has been long recognized, appropriate evaluation depends upon actually viewing the created objects rather than simply photographic reproductions of them. Of course, some of the artifacts listed above can be placed in a portfolio (i.e., a physical folder) though such placement may not show them to best advantage. For purposes of large-scale assessment, however, students are often forced to use portfolios in exhibiting what they have created.

For the past twenty years, the Advanced Placement Examination for studio art has used portfolios to certify that high school students have produced work that can earn them first-year credit at the college level. Students can submit two kinds of portfolios which they prepare during a year-long advanced placement course: (1) a drawing portfolio and (2) a general portfolio. The first kind of portfolio contains six original drawings which cannot exceed sixteen inches by twenty inches and a maximum of twenty slides in an area of special concentration such as self-portraiture or cubist still-life. The second kind of portfolio contains four original works and twenty slides that document a student's skills in four areas: drawing, color, design, and sculpture.

Despite the fact that these portfolios are highly regulated and depend heavily upon reproductions of original work, they have won the respect of students and teachers at both

[15] In the United Kingdom, such self-referenced assessment is often described as *ipsative*.

the high school and college levels. The Educational Testing Service has evolved a system of evaluation that is highly respected; a panel of judges, consisting of high school and college teachers of studio art, use exemplars as a means of anchoring scores on a scale of one to five (students must receive a score of three or higher to receive college credit). The exemplars selected to represent each level are prominently displayed throughout the evaluation process. What is distinctive about this approach is that it avoids any general scoring rubric or analytic scales similar to those commonly used to assess writing. Rather, the panel relies on the accuracy of their professional judgment based on long years of experience. As one of them put it:

> What you're really after is a mind at work, an interested, live, thinking human being. You want to see engagement. Recognition of it comes after long years of experience and you intuit it. (Askin, 1985, p. 25)

In comparing what they do with what is done in other disciplines, this teacher makes the following point:

> You can make those judgments as accurately as you can in mathematics or in writing or in any other subjects. These other subjects frequently have much more difficulty than we do in the visual arts in agreeing on standards . . . .You get a sense for copied work, a sense when there's engagement, when inspiration, belief, direct involvement are present or absent. (p. 25)

We suggest that those who evaluate writing might explore a comparable approach. Holistic judgment based on professional experience is less likely to distort evaluation than an analytic scheme based on specific features (K. Wolf, 1991).

Exhibiting artifacts—or reproductions of them—is not limited to the arts; this method of assessment is becoming increasingly prominent in mathematics and the sciences. At the Walden III School, students have presented a wide range of artifacts to the final jury: scale models, computer-generated graphics, terrariums, dioramas, and even scientific instruments of a simple nature (Feeney, 1984). In our own assessment project in Newburgh, we have introduced exhibitions alongside portfolios in the area of mathematics. Some teachers experimented with including photographs of artifacts in portfolios but discovered that the actual display of the artifacts has far more impact. Since we are working in early childhood education, the artifacts are often geometrically shaped figures constructed from building blocks (as well as the ubiquitous rods and wheels of tinker toys).

Within vocational education the exhibition of artifacts has long been used as a means of exit-level certification. It is in the making of a useful artifact that students are expected to display the requisite knowledge and skills in their field of specialization. With the new focus on combining applied and academic work, vocational schools are now leading the way in developing more integrated approaches to assessment. One exemplary approach can be found at the Paul M. Hodgson Vocational-Technical High School (PMH) in New Castle County in Delaware. As an early member of the Coalition of Essential Schools, PMH is now providing leadership in integrating the vocational and the academic. Central to its assessment procedures is a senior project that was developed by a team of teachers who attended a summer institute sponsored by the Coalition of Essential Schools. These teachers produced a document, *The Senior Project: An Exhibition of Achievement*, which outlines three components that each student must carry out:

- a research paper that requires students to expand their knowledge in a specialized area where they plan to create an artifact;
- an artifact that students design and construct in the specialized area; and
- a public presentation based on both the discourse and artifact that they have created (Godowsky, Scarbrough, & Steinwedel, 1991).

To document how the Senior Project is now working at PMH, Darling-Hammond and Ancess (1992) carried out a case study in which they describe what students have done in various specialized areas. Students have constructed a wide range of artifacts to satisfy the project. For instance, a full set of dentures, a catered school dinner, a kit for successful dressing, a stone fireplace, an irrigation system for a residential house, dugouts for a baseball field, and even portable houses for the low-income housing program in New Castle County. Darling-Hammond and Ancess observe that the project has been especially effective in assessing students who have not performed well on conventional testing. They present the case of one student who, plagued with severe dyslexia, had struggled throughout his school career. For his senior project, this student built an eighteenth-century-style pencil-post mahogany bed which was so finely crafted that it was valued at $2,500. Moreover, this student's commitment to bed-making led to unaccustomed academic success—he put together a well-documented research paper on the historical dimensions of bed-making. Despite his reputation as someone lacking oral skills, he managed an engaging presentation of his project before a committee that included both his carpentry teacher and his English teacher.

As illustrated by this example, the PMH assessment model uses the making of an artifact as an effective bridge to academic skills. Once students are committed to building an artifact, they are motivated to acquire the academic knowledge that supports their practical work. To return to the metaphors we introduced in the first section, the work of the hand leads to engagement of the head. In closing the discussion of artifacts, we would like to note that the senior project at PMH brings together the three kinds of work—discourse, artifact, and performance—that we are outlining in this section. Indeed, it is the effective integration of these three that distinguishes the PMH approach to assessment.

Performances

We have just described how a system of exit-level certification calls for individual students to make an oral presentation of work before a jury. This kind of culminating performance is becoming increasingly widespread within secondary schools committed to authentic assessment. Students are called upon to present work—discourse, artifact, or some combination of the two—which they have developed over an extended period of time. In this sense, this culminating performance resembles the oral defense of a doctoral dissertation because the students are expected to answer critical questions about the work they have engaged in. As in the case of a dissertation, the questions are posed by a jury composed of members who have worked directly with the student (in this case, classroom teachers) as well as members who are independent. Moreover, the student, just like the doctoral candidate, is often asked to revise the work that has been submitted. It is perhaps more fitting, therefore, to describe this final presentation not as an individual performance but as a constructive exchange designed to improve the work that the student has submitted.

Contrasting with such an exit-level procedure is a form of performance assessment in which students engage in hands-on activities to demonstrate their competence in a particular content area. Within mathematics and science, for example, such performance assessment has become especially popular. A number of states such as California, Connecticut, and New York have pioneered large-scale assessment programs in which students actually conduct experiments to demonstrate their competence. These programs draw heavily on a pilot project that the National Assessment of Educational Progress (NAEP), working with Great Britain's Assessment of Performance Unit, conducted in 1986. During this project, thirty different kinds of hands-on tasks—group activities, workstation activities, and even complete experiments—were field tested. After the

completion of the pilot project, NAEP made available detailed descriptions of the thirty tasks so that science educators throughout the country could adapt the ideas.

In order to illustrate one adaptation, we will examine the work of the Connecticut Multi-State Performance Assessment Collaborative Teams (CoMPACT) which consists of high school teachers from seven states: Connecticut, Michigan, Minnesota, New York, Texas, Vermont, and Wisconsin (Baron, 1990). The CoMPACT group, working under a forty-five–month grant from the National Science Foundation, has developed fifty performance tasks. About two-thirds of these tasks are in science (biology, chemistry, earth science, and physics) and one-third in mathematics (applied mathematics, algebra, geometry, and advanced mathematics).

Each task consists of three phases with initial and final phases of individual work being separated by a phase of group work. During the first phase, the task is presented to students and they are asked to formulate a hunch, an estimate of how it might be solved, and a preliminary design for an experiment. Individual work during this phase is important for several reasons: it organizes the students' preliminary thinking, serves as a catalyst for subsequent group discussion, and provides the teacher a baseline of where individual students are when they first approach the task (such a baseline is particularly important in self-referenced assessment that focuses on the individual student's progress).

In the second phase of the task, students must plan and work together to carry out a number of experiments to test certain hypotheses. A range of assessment tools are used to document individual contributions to these group activities such as written checklists, journals, logs, portfolios, and even videotapes of group discussions. Teachers use a preestablished form to rate individual students on a four-point scale with respect to each of the following six criteria:

1. group participation
2. staying on the topic
3. offering useful ideas
4. consideration of others
5. involving others
6. communicating with others

Appropriate assessment of how well individuals work within a group is one of the major challenges of this project.

During the third phase of the task, the focus returns to the individual. Students are presented a related task and must transfer what they have learned from the group experience. Hence teachers are in a position to evaluate to what degree individual students are able to apply socially based learning in a new context. In this way, the assessment reflects the alternation between individual and group experience that we commonly experience in our everyday lives. Such rhythmic alternation may become increasingly important in assessment practices if educators can find ways of meeting the formidable challenge of evaluating what individuals do within the group.

To illustrate how these three phases work, we here outline a science task described as "exploring the maplecopter." This task was designed for physics classes in order to assess how well students understand basic laws of motion, aerodynamics, and air resistance and use models in explaining scientific phenomena. During the first phase, individual students are asked to perform two tasks:

1. Throw a maple winged pod into the air and observe it while it floats down (describing the motion of the pod in as much detail as possible).
2. Explain how the pod spins as it falls and why it does so.

During the second phase, students working in groups are asked to do the following seven tasks:

1. Describe the motion of the pod (using the observations of the entire group).
2. Establish the variables that might affect its motion.
3. Design and implement experiments to test each variable.
4. Construct various kinds of paper helicopters to further test each variable.
5. Based on the results in 4, design and implement further experiments.
6. Synthesize the results from 3 through 5.
7. Provide a broad explanation of the motion of the pod, focusing on particular biological advantages of its structure.

Once this group activity has been completed, the final phase of the task focuses on individual students. They enter a competition in which each constructs a paper helicopter that is designed to remain in the air for a maximal length of time. In addition to constructing the helicopter, they must describe the factors that are crucial to its aerodynamic design.

Competitions

There is a well-established tradition of using competitions to assess the performance of both individuals and groups. In prestigious academic areas such as mathematics and science, national competitions for individuals are held annually. Certain of these competitions such as the Westinghouse Science Talent Search are well-funded, with the winners being assured of college scholarships. Competitions are also common within the arts, especially music, where students compete not only as performers but also as composers. Many of these competitions are concerned with group rather than individual performance—choruses, bands, and orchestras all participate in competitions at various levels, including school district, state, and national.

Within vocational education, the emphasis on team performance is especially well-developed. Team competitions are common in many vocational areas, ranging from home economics (e.g., the culinary arts) to technical specialties (e.g., machinery maintenance and repair). In agricultural states, these competitions are often built around farm machinery; in the state of Ohio, for example, the Agricultural Educational Service sponsors an annual competition in tractor troubleshooting. Each school enters a team of two at the district level, and the winners at this level then compete at the state level. The two team members attempt to diagnose and actually correct engine malfunctions at five workstations. At each workstation a tractor is set up with two malfunctions related to the air, fuel, electrical, or hydraulic systems on its engine (gasoline or diesel). These malfunctions are designed to be fairly typical (e.g., burnt points, a faulty condenser, shorted spark plugs, or defective wiring or carburetion). The two students are supplied with a standard toolkit that includes various screwdrivers, pliers, wrenches, sockets, extension bars, and more specialized tools such as a spinner handle, a ratchet wrench handle, and a torquemeter handle. The students have only twenty minutes at each workstation to identify the two malfunctions and repair them.

The team performance at each station is scored in four major areas: (1) workmanship, (2) identification and repair of the malfunctions, (3) tractor performance (as measured by a dynamometer), and (4) time used to complete the work. The first area is the most difficult to score since the judges must closely observe student performance along a number of dimensions such as how well they use tools, how carefully they observe safety practices, and how skillfully they employ troubleshooting procedures (do they, for example, follow a systematic plan of action so that they avoid troubleshooting a component before they have reason to suspect that it is faulty?). This area, taken as a whole, is more heavily weighted than the others (it constitutes about a third of the total score). The other three areas, each constituting about a fifth of the score, can be more easily scored. For the second area, identifying the malfunctions and repairing them are given equal weight; for the third, a dynamometer is used to determine the actual horsepower of the engine once the students complete their work; and for the fourth, the actual number of minutes they work is recorded. After the students are finished with the five workstations, they take a written test (it constitutes only about a tenth of the composite score).

Such vocational competitions represent a peculiar mix of values. On the one hand, they embody basic values in authentic assessment since students not only work together as a team but they confront a real problem (as mentioned, strong emphasis is placed on typical malfunctions). On the other hand, certain values are carried over from conventional testing: for example, students receive a substantial reward for completing the work rapidly (they earn credit for every minute saved under the allotted twenty minutes). In the real world, however, speed in tractor repair is seldom all that important. In fact, this emphasis on speed works against other values that are rewarded in the competition (e.g., following appropriate safety practices or insuring that the repaired engine operates with full power).

We suspect that such a contradiction can be partly traced to the fact that the assessment procedures take the form of a competition. It is traditional within most competitions to place a premium on speed—who can run a mile, swim two hundred meters, or solve a mathematical problem most rapidly? Certainly speed in human performance has a primitive, even esthetic appeal, but we question its value in assessment procedures where it militates against human values such as safety and high-quality workmanship.

In closing this discussion of competition, we would like to mention two other problems. First, a competition tends to focus on only the most talented students. From Wiggins' (1991a) perspective, such focus is not necessarily negative since it can be an important means of setting high standards and thus raising the general level of performance among all students. In practice, however, students are often discouraged from participating in a particular activity because they judge their own performance according to the standards of those who excel in a specialized activity. In effect, competitions can lead to standards that are not realistic for most students. Second, organized competitions can undermine certain aspects of cooperative learning that authentic assessment seeks to foster. It is true that team competitions are designed to encourage students to work together toward a common goal. Since the goal is for one team of students to outperform all others, these competitions contribute to an ethos that works against cooperation in a broader sense.

Internships

In closing our discussion of performance in authentic assessment, we would like to call attention to the increasing use of workplace internships within authentic assessment. Since such internships involve student performance over an extended period of time, they involve complicated problems of documenting what students actually do. One way of addressing the problems is to require the students themselves to provide substantial documentation. This approach is being used by the International High School, which is supported jointly by the New York City Board of Education and the City University of New York. This school enrolls nearly five-hundred students from more than fifty countries and has developed an innovative approach for teaching English through various content areas (there is no separate program in English as a second language). Central to this approach is a personal and career development program that extends throughout the three years. Each year, students take a course in career development where they "examine the workplace from a sociological perspective, studying such issues as hierarchy and authority, gender issues, and organizational structures" (Darling-Hammond & Ancess, 1992, p. 36). In addition, students participate in an internship (usually for a half-day, four days a week during one trimester) that is accompanied by a weekly seminar in which students together explore their experiences in the workplace. The seminar is also designed to help them with practical tasks such as preparing for a job interview and writing a resume and business letters. Students usually serve internships within the social services where they are able to draw on their first-language skills. At the same time, the internships are designed to develop their practical skills in English, build up their workplace competencies,

and strengthen their ability to cope with the social and cultural environment of a large urban area.

As a means of providing documentation about their internships, students are required to construct an *internship album* which consists of four parts:

1. students identify their personal and career objectives and articulate a rationale for their choice of an internship;
2. they describe their job interview, their job duties, a typical day of work, and the organizational structure, relationships, and roles at their workplace;
3. they present interviews with their supervisors and coworkers about their own education, career development, and job satisfaction as well as the salary range, benefits, and employment opportunities in the particular field; and
4. they reflect on their own work attitudes, the skills and competencies they are acquiring (or need to acquire), and what they have learned about their work as well as themselves.

Students read each other's albums as they are developing them and make suggestions as to how they might be improved. Once the albums are complete, they are placed in a small library where they become a curriculum resource (they are particularly valuable to students who are trying to decide about internship placements).

The albums are then evaluated according to criteria that have to do with content, language usage, and rhetorical presentation. Student peers as well as teachers use these criteria to produce scores for the albums. It is not only the internship albums that are evaluated; student performance in the workplace itself is evaluated by workplace supervisors. They use a five-point scale (ranging from excellent to poor) to evaluate individual interns in nine areas: attendance, promptness, quality of performance, dependability, cooperation with both coworkers and supervisors, ability to learn, initiative, and growth. In addition, the students themselves fill out a questionnaire in which they evaluate their own performance, their internship album, and even the assessments that have been made by their student peers, teachers, and supervisors. Whenever their own assessments differ, the students are required to provide evidence for their own evaluations. As the final stage in the assessment process, the teacher takes account of what the student has provided and develops an overall assessment of both internship performance and the

album that has been created around that performance. Unless they are carefully managed, multiple assessments, especially those from other students, may force those being evaluated into a defensive posture. In our own experience, peer evaluation is better used during the process of developing work. At this stage, such evaluation can, if managed well, encourage a sense of collaboration among students. Once the work is complete, we question the wisdom of involving peers in the final evaluation.

Records

We can begin this section on records by returning to the basic distinction we established with respect to documentation practices. Work samples consist of what students have actually done; whereas, records are *about* what they have done. In assessment models that emphasize self-monitoring, these two kinds of practices, as we have seen, are not held separate. Self-monitoring calls for students to maintain records about the work samples they assemble, and these records can, of course, be viewed as constituting further work samples. In assembling a portfolio, for example, students are often required to provide a letter that orients the reader to their work. This letter can be viewed as a record because it tells how they decided on which pieces to include, how they organized the individual pieces, the strengths and weaknesses of these pieces, and so on. At the same time, the letter constitutes a further sample of work, one which for the sake of clarity can be described as *indexical* to the other work.

Indexical work samples are particularly important in authentic assessment. They provide important evidence about two fundamental capacities—the extent to which students can organize and present their work as well as the extent to which they can assess its strengths and weaknesses. As students keep records about their own work, they develop the very capacities those records are supposed to document. We here observe the power of reflexivity within these documentation practices; the means used to document the growth of certain capacities itself facilitates such growth.

System-Oriented Records

In many high schools, student record-keeping is central in exit-level certification procedures. In order to graduate, students are required to assemble and present evidence that they are competent in a number of basic areas. Consider the graduation requirements

of the Central Park East Secondary School (CPE) in New York City. Students are required to provide documentation in fourteen basic areas (Henderson & Meier, 1990):

- postgraduate planning
- autobiography of school experience
- community service
- practical knowledge and skills
- media
- literature
- history
- geography
- foreign languages
- ethics and philosophy
- fine arts and aesthetics
- math
- science and technology
- physical fitness

For all areas, students are required to maintain records. For the postgraduate plan, students submit an initial written plan that they update throughout their school career and also include appropriate personal references and recommendations to document their developing interests. In the area of community service, they provide records of particular work they have done outside of school. This often takes the form of a written resume of work experiences which can be used in future job searches. As for a specific internship in the workplace, students include appropriate documentation from their employers as to matters such as punctuality and reliability. They can also present a work log in which they have recorded their daily experiences on the job.

In the more academic areas, students submit various records along with work samples. In literature, for example, students can keep records about what they have read for particular courses as well as on their own. These records can take the form of journal

entries in which they present their basic reactions to what they have read. Even in an area such as physical fitness, students use records of various physical activities they have been involved in to supplement their demonstration of fitness through performance. They are especially encouraged to document their participation in activities such as hiking or bicycling that can be maintained throughout their lives.

As students develop documentation in the various areas, they are guided by the five *Habits of Mind* that are central to the educational philosophy of CPE (and other schools associated with the Coalition of Essential Schools founded by Ted Sizer and colleagues):

1. Weighing evidence
   How do we know what we know? What is the evidence and is it credible?
2. Analyzing varying viewpoints
   What viewpoint are we hearing or seeing? Who has established it, and what are her/his intentions?
3. Establishing connections and relationships
   Where have we encountered this before? How is it connected to other things?
4. Speculating on possibilities
   Can we imagine alternatives? What if we changed certain aspects of the situation?
5. Assessing value both socially and personally
   What difference does this make? Why should we care about it? (Henderson & Meier, 1990)

These habits of mind are also used by those who evaluate the students. They constitute a set of five criteria that evaluators use to score what individual students do in the fourteen areas. Independent evaluators, who are primarily teachers at CPE, first rank students on a scale of one to four with respect to each criterion; they then meet to negotiate a score for each criterion and eventually establish a composite score for student performance (this score can be maximally twenty if a student receives a four for all five criteria). Individual students are ranked in all fourteen areas according to the following scale:

1. 18-20 distinguished
2. 15-17 satisfactory
3. 12-14 minimally satisfactory

If students score below twelve in a particular area, they must revise their work and submit it again for a second evaluation. In a recent case study of CPE, Darling-Hammond and Ancess (1992) claim that this "process of evaluating and revising and re-evaluating makes . . . assessment . . . fundamentally a learning process, one that promotes both self-evaluative capabilities and habits of work" (p. 22). In effect, this assessment approach is classroom-oriented—supporting student learning—at the same time that it is system-oriented—providing certification that individual students have developed the requisite knowledge and skills. In this sense, the approach of CPE is exemplary in secondary education, which needs to develop assessment models that can fulfill both classroom-oriented and system-oriented functions.

Within vocational education the role of record-keeping is also central in exit-level certification. We earlier discussed the Paul M. Hodgson Vocational-Technical High School (PMH) which requires students to document their competency in a given specialization. We pointed out how students write a research paper that supports their practical work, but they are also expected to provide documentation that can eventually be used in securing employment such as a clear résumé of relevant work experience and detailed letters of recommendation from those who have supervised a workplace internship. Moreover, students are expected to use such documentation in contacting potential employers. The PMH program maintains strong linkages with a broad range of employers in New Castle County such as the major corporations of DuPont, General Motors, and Chrysler as well as more than two thousand small businesses.

Michigan is experimenting with a system of exit-level certification that also integrates the vocational and the academic. In 1987, the state commissioned an Employability Skills Task Force, combining educators and business leaders to set up a system that requires students, working with teachers and employers, to keep records in three basic areas:

1. Academic skills—the ability to read and understand written materials, charts, and graphs
2. Teamwork skills—the ability to express ideas to colleagues and then negotiate mutually acceptable goals
3. Personal management skills—the ability to meet deadlines and handle details in an organized way

Students compile their own records in these three areas, drawing on the resources of teachers and employers (e.g., both provide letters of recommendation that focus on student competencies in one or more of the three areas).

This assessment program was recently piloted in twenty-two school districts and, like any new program, received a mixed reception. Some schools have adjusted well to the greater integration of school and the workplace, but others have experienced considerable difficulty. Crucial to the success of this program is a strong commitment from the participating employers who must be willing not only to create meaningful work for students but to provide appropriate supervision. Apparently a number of employers have been unwilling to make such a serious commitment because of the high cost and the rapid turnover of student interns. One way of encouraging a more serious commitment is to help employers recruit talented students for long-term employment. From the student perspective, a long-term commitment would be more attractive if workplace internships could be integrated with higher education; such internships would not only defray the financial burden of such education but introduce a practical component, as well.

Classroom-Oriented Records

The use of teacher records about student performance has a long history. Consider, for example, a report card—it has always been a place where teachers could display certain kinds of information they recorded on a daily basis. To take the familiar example of student's attendance, a report card usually includes three boxes—P for present, A for absent, and L for late—in which teachers transfer information from their daily records about daily homework, weekly quizzes, and semester-long projects. As to the less academic aspects of student behavior, teachers can record such information as whether students are neat in their work, cooperative with the teacher, and helpful to other students.

What the traditional report card lacks are more extended spaces where teachers can record observations about a student's academic work. Some advocates of authentic assessment have developed schemes to help teachers make such observations. In early childhood education, Chittenden and Courtney (1989) have developed a moderately formal system for teachers to use in recording their observations of children in a number of stable classroom settings. These researchers claim that such documentation is valuable precisely because it is anchored in what teachers are already doing. In effect, it provides teachers a means by which they can observe more closely—and hence be better prepared to work

with—individual children. These researchers also claim that such documentation can, if sufficiently formalized, "replace test scores as accepted forms of evidence" (p. 11).

These observational schemes are not as easy to apply as students progress in school. As their work becomes more involved with symbolic systems, it becomes more internal and less accessible to direct observation.[16] Older students can, however, be encouraged to observe their own work; they have more access, at least in principle, to their own use of symbolic systems and are, therefore, in a position to make observations that an external observer cannot. Thus a scheme for self-observation is sometimes included in an assessment model, especially one concerned with evaluating students' capacity to monitor their own learning.

The schemes that guide such observation can vary greatly in the degree to which they specify what students are to observe; they range from highly focused instruments such as questionnaires to loosely structured ones such as an informal interview. The more focused instruments yield more comparable forms of data, but they often lead to impoverished forms of observation. A loosely structured interview can elicit from students a richer account of their own work.

In some instances, a scheme to guide observation is developed that a teacher or workplace supervisor as well as the student can draw on. We are currently developing such a scheme for use in a broad model of assessment for the Edwin Gould Academy, a secondary school established jointly by New York State and the Edwin Gould Foundation. The academy is located in a rural environment about thirty miles north of New York City to provide a supportive environment for students in foster care who have not been successful in public education within the city. Crucial to this environment are activity centers that provide real-world contexts for students to acquire experientially-based knowledge and skills:

[16] A major challenge in the current movement toward cognitive apprenticeships is for experts to develop methods for making visible the various cognitive processes in which they engage (Collins, Brown, & Holum, 1991). If such methods can be successfully developed, they will undoubtedly play an important role in authentic assessment. We are intrigued by the movement but puzzled about certain claims of authenticity. As we examine what actually takes place during these apprenticeships, we are struck by how the external display of cognitive processes necessarily introduces a certain arbitrariness.

1. a *vocational education center* where students are exposed to the contemporary workplace (with focus on integrating social and technological skills)

2. an *environmental center* where students learn about the natural world (with focus on an ecological approach to natural resources)

3. a *performing arts center* where students are involved in arts such as music, dance, drama, and storytelling (with focus on creative use of traditional resources within modern society)

4. a *community service center* where students are active participants in the social services provided in a modern society (with focus on integrating interpersonal and organizational skills)

Within these centers, students are involved in a variety of apprenticeships. For example, in the vocational educational center, students serve apprenticeships that require them to use computers, operate photographic and video equipment, service and repair cars, employ carpentry and masonry skills in construction projects, and work in food services. These apprenticeships are linked to everyday support of the academy where the students are in residence so that they can acquire a firm sense of the practical consequences of their work. In the community service center, students are involved in apprenticeships, with the support of graduate student interns from the School of Social Work at Columbia University, that range from daycare for young children to support services for senior citizens.

As students work in the activity centers, they will be required to keep systematic records of what they do. In the vocational education center, for example, they will keep daily logs of the work they do. The students store these logs in computer files then draw on them when meeting with their supervisors. In this way, students participate in continuous assessment of their own work. A major benefit of such record-keeping is that it helps students develop the range of skills needed for the workplace (i.e., literacy skills in creating the records, technological skills in storing them in computer files, and communication skills in retrieving them for use in a social setting). As students participate in public discourse about their own performance, they learn to internalize the standards crucial to good work.

In community service work, students will be required to keep process journals of their work—in working with young children, reading stories to them or helping them to

write and illustrate their own storybooks and, in the case of old people, reading newspapers to them or helping them to write letters or even autobiographical sketches for their families. Supplementing these records will be videotaped samples of students actually performing work. Even in this second kind of assessment, students will be required to evaluate their own work, drawing on heuristically–oriented guidelines developed for this purpose. The graduate student interns who supervise the students' work will use these same guidelines to establish an independent evaluation of the videotaped samples of work. Once the two separate evaluations have been made, the students and the interns will work together to achieve a mutually satisfying joint evaluation. This is a crucial stage, for it involves a complex process of ethnocultural negotiation such as students explaining their own sense of what they thought they were doing, and professionals explaining how such actions might have been perceived differently from another ethnocultural perspective. This negotiatory process allows students to express their legitimate concerns about whether they are being judged by fair standards. It also helps to facilitate their awareness of the professional standards that are required for effective performance in the workplace.

The assessment model that we envision for the Edwin Gould Academy is valuable for three reasons. To begin with, it is designed to integrate students' school experience with meaningful work experience; indeed, boundaries between school and the workplace are viewed as relatively permeable within this model. Moreover, the model requires students to use various methods in assessing their own work experience, ranging from records that they themselves keep to analysis of videotaped samples of the actual experience. In this way, they are forced to develop their own standards for judging the quality of their work. Finally, the model leads students and the professional who supervises them to negotiate a mutually satisfying assessment of their work performance. In this sense, the model reinforces students assuming active responsibility for their own assessment but learning to integrate their judgments with those of a professional community. At the same time, it encourages both students and professionals to recognize the ethnocultural dimensions that are present in any educational assessment. In recognizing these dimensions, both parties are better able to respect the evaluation they arrive at. They come to appreciate how contingent it is on ethnocultural values, values that have nevertheless come to play a dominant role within the larger society.

In closing our discussion of documentation practices, we would like to raise certain problems that extensive record keeping entails. First of all, it can be quite difficult to

convince either teachers or students of the value of keeping records. Most human beings have a natural inclination to leave things unrecorded. This is partly a result of the amount of work it takes to maintain effective records; it is tedious to keep track of details on a regular basis. Our resistance to record keeping can also be traced to a human preference for vagueness; if there are no records, then what we do cannot be easily investigated. Moreover, there is the problem of what to do with records once they have been produced. Proponents of authentic assessment recommend that records follow students in school so that future teachers can benefit from them. This is a worthy idea, but we have found it extremely difficult to implement. No matter how carefully records are maintained, future teachers find them difficult to use. They cannot easily establish the context in which the information was originally recorded, and even if they could, the recorded information often lacks relevance to their present concerns.

Despite all these difficulties, we still believe that records, when used judiciously, constitute an important component in authentic assessment. We particularly value the records that students keep themselves. As they engage in appropriate record-keeping, they are developing the skills that are crucial to their success in school and ultimately in the workplace (i.e., the capacity to organize information and store it in such a way that they can retrieve it). In effect, the documentation practices themselves teach the very skills that they are designed to assess.

Finally, we would like to emphasize that the capacity to keep efficient records is increasingly important as students prepare to participate in a society that is undergoing rapid technological change. Technology has vastly increased our capacities to keep track of the work we do as well as to present it in a more organized fashion. Moreover, it has greatly enhanced our capacity to do collaborative work. Consider, for example, how technology facilitates the work that a team of researchers is able to do. As they gather information, they can store it in a hierarchically–indexed system to which they all have continuous access. As they add new information to the system, they can immediately view updated displays of what it contains (at the same time they can maintain access to earlier states of information). As many of us have learned from experience, the power of these new technological systems is easy to abuse; unless they are used judiciously, they can spawn masses of useless information. In a technological society, students must learn not simply how to maintain records but to develop good judgment about how to use the information

they contain. In the absence of such judgment, technology simply amplifies the human capacity to maintain burdensome records that have little functional value.

ASSESSING AUTHENTIC ASSESSMENT

We have shaped this final section by drawing on three principles that are frequently used in discussions of testing and assessment policy: excellence, equity, and efficiency. In the earlier part of this century, these principles were invoked by educators such as Thorndike (1913) to support the development of what we are calling conventional testing and to develop tests that would maintain high standards yet be easily administered throughout the nation. They viewed their work as a scientific enterprise which, if properly implemented, could protect students from the inevitable inequities of local evaluation. They claimed, as indeed many do today, that within a heterogeneous society, individual teachers can make unfair judgments, especially when they are dealing with students from diverse ethnocultural backgrounds.

In recent years, the use of these principles has shifted dramatically; they are now used to make a case for authentic assessment as opposed to conventional testing. In this closing section, we lay out how the principles are commonly used in constructing this case. In doing so, however, we also use these principles to examine certain difficulties that arise within an alternative approach. Assessment is such a complex enterprise that any approach, no matter how well-intentioned, is difficult to implement; in the case of authentic assessment, the difficulties are especially pronounced with respect to equity and efficiency.

Excellence

Those who support authentic assessment place the pursuit of excellence at the heart of their policies (Darling-Hammond, 1990; Neill & Medina, 1989; Newmann, 1991; Resnick, 1987a, 1989; Wiggins, 1991a). They point out that such assessment, in contrast to conventional testing, has the following characteristics:

1. it requires students to construct responses rather than select among preexisting options;

2. it elicits from students higher-order thinking in addition to basic skills;

3. it uses direct assessment of holistic projects rather than indirect measures; and

4. it is integrated with classroom instruction rather than separated from it.

As we discuss each of these characteristics, we first consider the ways in which conventional testing can be viewed as diminishing excellence.

Constructed Responses

In conventional testing, students are forced to adopt a reactive posture when responding to tasks. They select from a set of preexisting options which in many instances does not include what they would choose if the task were open-ended (Freedle & Duran, 1987; Hill, Anderson, Watt, & Ray, 1989). Such forced selection can encourage the use of mechanical procedures in which students attend primarily to idiosyncrasies of surface format; they may, for example, select an option simply because it differs from the others in either length or syntactic structure. As Gee (1992a) points out, such selection also forces them to draw on a highly restricted body of knowledge, beliefs, and values. He reports a study which found that skilled test takers are able to select the target response for certain questions without reading the passage that they accompany and simply select the particular response that accords most closely with idealized forms of what they know, believe, and value. Gee goes on to point out that the students who are able to select the target response are often those who are least likely to use these idealized forms in their real-world reading.

When students construct their own responses, they are forced to adopt, at least to some degree, the more active stance that characterizes everyday thinking. They must decide on what is relevant, organize it in some way, and then work out its presentation. This more active role, according to the OTA (1992) report, makes possible a "closer examination of learners' thinking processes. When students write out the steps taken in solving a proof, or a list of how they reached their conclusions, the students' thinking processes can be examined and scored" (p. 215). Those who support testing reform rightly emphasize the use of tasks that can shed light on student thinking. In constructing such tasks, however, we must not assume that simply requiring students to construct a response insures access to their thinking. We need to bear in mind that any response is, after all, a highly conventionalized representation. In a testing situation, the gap between

response and actual thinking can be especially pronounced. As we noted above, a test tends to elicit from students conventionalized forms of response.

From the practical perspective of motivating the pursuit of excellence, constructed-response tasks do offer a clear advantage. Various studies have found that students spend substantially less time preparing for a test with multiple-choice tasks than one that calls for constructed responses (D'Ydewalle, Swerts, & De Corte, 1983; Traub & MacRury, 1990; Warren, 1979). As Traub and MacRury put it, students prefer multiple-choice tests because they "are easier to prepare for, are easier to take, and hold forth hope for higher relative scores" (p. 42). In contrast, students expend more effort not only in preparing for a constructed-response test but also in actually taking it, and this additional effort is crucial in maintaining a high level of performance. However, it is documentation practices that motivate students to work at the highest levels (Resnick, 1987a; Wiggins, 1991a). Since students work on a portfolio or exhibition over an extended period of time, they become more invested in the project. More importantly, they are usually able to select the projects that they work on, and the power of choice facilitates commitment and ultimately more substantial achievement.

Higher-Order Thinking

A frequent claim is that conventional testing compromises excellence by its undue emphasis on basic skills (Archibald & Newmann, 1988; Haney, 1984; Hoffmann, 1962; Resnick, 1987b). The failure to satisfy the principle of excellence derives at least in part from efforts to satisfy the other two principles. By using the multiple-choice format, test makers can, at least on the surface, avoid the equity problems that result from evaluating individually constructed responses. By the same token, the format leads to greater efficiency since its forced-choice responses can be machine-scored.

In this quest for equity and efficiency, test makers inevitably end up compromising excellence. The machinery of multiple-choice testing calls for a task with a clear right answer, so it becomes difficult to construct one that elicits the complex thinking used in real problem-solving. In fact, if test takers do engage in such thinking when responding to a multiple-choice task, they are likely to be led astray. On reading tests, for example, they are often attracted to distracters that invite richer inferencing. As our research (Hill & Larsen, 1983) has shown, such inferencing is fundamental to real-world reading but is best avoided when responding to a test because it leads readers to construct a context that is too

powerful for what the tasks ordinarily call for—the recycling of discrete bits of information.

Those who support authentic assessment have placed major emphasis on developing tasks that elicit higher-order thinking (Haney & Madaus, 1989; Newmann, 1991; Resnick, 1987b). They claim that such tasks should be central not only in documentation practices but even in alternative testing. Hence, an essay test tends to be constructed around a real-world problem (e.g., how to protect the natural environment in a modern economy). The inherent complexity of such a problem demands that students engage in higher-order thinking, though as we have pointed out, the restricted time frames imposed by testing militate against effective displays of such thinking.

The concern with higher-order thinking is crucial to authentic assessment, but it introduces problems of its own. To begin with, it can lead to an unwarranted opposition between basic skills and higher-order thinking. In public pronouncements on testing and assessment, these two are often set against each other; to our way of thinking, they are better viewed as complementary since students cannot demonstrate their higher-order thinking effectively unless they have mastered certain basic skills. To illustrate this point, we can draw on foreign language learning in which there has been a strong focus in recent years on communicative competence (i.e., being able to use a language to carry out real-world tasks) rather than linguistic competence (i.e., simply knowing the forms of the language). This movement has provided a number of important correctives in foreign language teaching. At the same time, however, teachers have become increasingly aware that a genuinely communicative approach requires learners to exercise even more control over language form. When learners acquire language in a highly specific context, they are forced to be more attentive to constraints on language form (e.g., the tendency to use *I'll*, as opposed to *I will* in everyday oral communication). What is important is that learners acquire this control as they carry out real-world tasks; in this way, they are more likely to retain it as they engage in new activities (Widdowson, 1978). In effect, authentic assessment must not ignore basic skills but rather embed them—and evaluate them—within more complex activities.

An additional problem that stems from the focus on higher-order thinking is setting unrealistic expectations. Consider, for example, a testing situation in which students are expected to write extemporaneously about a complex problem in fifty minutes. Obviously

this time constraint forces students to move hurriedly through the topic at hand; yet the standards for evaluating such writing—at least as they are embedded in official rubrics—are often those we associate with the considered work of experienced writers.

Even documentation practices can reflect a lack of realism about what students are expected to accomplish. It is true that students have a longer period of time in which to prepare their work; and they are encouraged, in many instances, to seek a good deal of social support. Still they can be asked, in the name of excellence, to take on projects that exceed their capacities. At a recent workshop on authentic assessment at Teachers College, Columbia University, Grant Wiggins (1991b) recommended projects in which high school students design, and in some instances actually construct, such artifacts as a scientific instrument, a working roller coaster, or a textbook on American history. In our own experience, high school students generally lack the maturity and experience to take on such projects; indeed, many of the projects mentioned at the workshop would perplex adults, even those who have developed specialized skills. In rightly avoiding low expectations, we can introduce expectations that are inordinately high and that have negative consequences. On the one hand, students may be discouraged and make little effort; on the other, they may be encouraged to overvalue what they do accomplish. Such inflation runs counter to a central concern of assessment reform—namely, to develop in students a capacity for reliable assessment of their own work.[17]

Direct Assessment

It is commonly accepted that conventional testing proceeds by indirection. Rather than requiring students to solve a complex problem, such testing presents tasks that elicit discrete bits of knowledge or skill. The theory is that a limited number of tasks, if appropriately designed, can provide adequate coverage of the knowledge and skills that students need to perform effectively in a given domain.

By way of contrast, authentic assessment is committed to students carrying out holistic projects that have intrinsic merit. It is, of course, difficult to implement such

[17] These higher standards seem more appropriate in collaborative models of testing and assessment. Certainly there have been impressive accomplishments by high school students who have worked together: for example, students from Conval High School in New Hampshire managed to build and race a solar-powered car (National Council on Vocational Education, 1990). This achievement was not, however, carried out within an assessment framework, and in any case, it is not clear how much scaffolding was provided.

projects within alternative testing since the time limits necessarily constrain what students can do. Moreover, a test radically alters the context so, by definition, the assessment becomes indirect. Students do a task not for its own merits but rather to display certain capacities that need to be assessed. The powerful effects of testing cannot be neutralized simply by introducing tasks with an authentic provenance.

This problem of direct testing becomes particularly acute as workplace tasks are simulated. These tasks, though advertised as authentic, force students to make quite inauthentic responses if they lack an experiential base from which to respond. From our vantage point, workplace tasks are not especially promising candidates for alternative testing since they generally call for highly embedded knowledge and skills. An activity that an individual is able to perform quite routinely on the job can present formidable obstacles when it is transferred to a testing situation (Berryman & Bailey, 1992; Carlson, 1990; Lave, 1988; Scribner, 1988, 1991).

It is easier for students to carry out holistic projects apart from a testing situation. Hence, documentation practices appropriately emphasize projects that involve extended discourse, complex artifacts, and multifaceted performances. The first two—discourse and artifact—often yield stable products that can be evaluated more reliably, though even here a certain dynamism may be present and complicate the task of evaluation. The discourse may be oral or the artifact—say, a computer program—structured so that it must be experienced in a temporal sequence. In the case of performances, such dynamic character is, of course, unavoidable; and since many different elements are working together, evaluators can be easily overwhelmed by detail. This problem can be alleviated by videotaping the performance so that retrospective evaluation becomes possible. But then videotaping introduces a host of complications that we earlier described (see pp. 33). Despite the formidable problems, retrospective evaluation of videotaped performance remains a promising development in authentic assessment. In particular, it provides a useful method for sampling what students do in a workplace internship and for encouraging them to become involved in assessing their own work.

Integration with Classroom Instruction

A major argument against conventional testing is that it undermines what goes on in the classroom. This is especially true in the case of a high stakes test such as the SAT, which by design is not integrated with any particular curriculum. Nevertheless teachers

feel obligated to prepare students for a test that has important consequences in their lives. Hence, they often end up using inordinate amounts of class time helping students to become testwise. Students are trained to follow a highly specialized set of procedures: read the questions before the passage; try to get the choices down to two before guessing; once you've decided on an answer, don't second-guess yourself; skip over difficult questions and come back to them later; and so on. It seems evident that these procedures do not have much value beyond the world of testing (they do, of course, provide a certain practice in performing under pressure, a condition that many unfortunately confront in the workplace).

Here is how Mary Lee Smith (1991) describes the effects of testing on classroom time in a large-scale study that she conducted in Arizona:

> Time required for the Iowa Test of Basic Skills and state-mandated criterion-referenced tests, the time teachers elect (or principals require) to take the test, and time spent in recovering from the tests amounted to about a 100-hour bite out of instructional time in the schools we studied. (p. 10)

Even teachers committed to a broad curriculum, one designed to develop critical thinking and expressive power, often end up spending too much time on low-level skills that are not connected to their larger goals. As a teacher in Mary Lee Smith's (1991) study put it, "I wanted to keep my literature program, and I knew if my scores were low, they would make me go back to the basal, so I drilled them with Scoring High worksheets [that match the format and objectives of the Iowa Test of Basic Skills]" (p. 9).

From the perspective of Grant Wiggins, teachers should not lament the gap between what they want to teach and how they must test. Rather they should work to develop a strategic connection between the two. In an article entitled *Teaching to the (Authentic) Test*, (Wiggins, 1989a), claims that testing, if properly aligned with curriculum and instruction, can become a powerful means of driving student performance. Once authentic forms of testing are put in place, the phrase "teaching to the test" takes on positive rather than negative connotations. Teachers are then in a position to use testing as a means of achieving fundamental goals. For example, the right kind of testing practices would help the teacher described above support the literature program that she is committed to.

Certainly Wiggins is right to recognize that testing can be a crucial means of motivating student performance. We would like to call attention, however, to certain

problems with this approach. To begin with, we question whether testing, even if radically reformed, can ever be a sufficient means of motivating students to achieve excellence. The testing paradigm, at least as it is presently constituted, is simply too restricted to provide the necessary challenges. Consider the severe time constraints it imposes—how can students carry out a demanding task in the short span of time that a test allows? It is the limitations of the testing paradigm that have spawned the rich range of documentation practices that we have described in this report.

There is, however, a further problem that remains even when assessment incorporates documentation practices alongside alternative testing. Assessment, by its very nature, tends to be conservative. Once tasks are put in place, they tend to take on a life of their own. This inertia is the result of a number of forces. To begin with, so much work goes into developing and refining tasks that they are not easily discarded; moreover, as data on student performance accumulates, the tasks tend to be held in place to insure longitudinal comparability. On the other hand, curriculum and instruction are subject to continuous pressures to reform education so that it is more closely aligned with the changing demands of the larger society. At this point we can usefully return to our distinction between endogenous and exogenous models of education. As educators attempt to shift toward a more exogenous model, they often claim that an appropriate use of authentic assessment can provide crucial leverage in bringing about reform. We are sympathetic to this claim—and, indeed, often make it ourselves—but we would like to warn that any assessment activity, no matter how authentic, can reinforce an endogenous orientation to education.

In closing our discussion of assessment in relation to excellence, we recommend caution about fostering the expectation that assessment reform necessarily leads to higher standards. This point has been forcefully made by Koretz, Madaus, Haertel, and Beaton (1992) in congressional testimony that criticizes the report *Raising Standards for American Education*, published by the National Council on Education Standards and Testing (1992). As Katz et al. observe, "Using test-based accountability to drive education is hardly a new idea. This approach has been tried many times over a period of centuries in numerous countries, and its track record is unimpressive" (p. 2). They point out that assessment, by its very nature, cannot be as broadly based as curriculum and instruction; whatever it emphasizes acquires a certain weight for students—and teachers, too—and thus constrains what they do in the classroom. Apart from this pressure to narrow curriculum and

instruction, the authors identify other negative outcomes such as test cramming, high dropout rates among at-risk students, and the false accountability that Cannell (1987, 1989) and Koretz (1988) have called attention to and that can result from relying on testing to drive educational excellence. As Madaus puts it in an interview that recently appeared in the *New York Times*, "We are not going to test our way out of the nation's educational problems." He goes on to suggest that if we are serious about achieving excellence in our schools, "we need to look at materials teachers are using, how teachers are trained, the support teachers have, the social support kids have, a whole series of complex, interrelated factors that have to be tackled" (cited in Chira, 1992, p. A19).

Equity

Educators have made frequent claims about how authentic assessment will lead to greater equity. They point out that such assessment, in contrast to conventional testing, has the following characteristics:

1. it uses samples of student work collected over an extended period of time;
2. it is based on clear criteria of which students are made aware;
3. it allows for the possibility of multiple human judgments; and
4. it is more closely related to what students learn in the classroom.

Multiple Samples of Extended Work

Conventional testing can be challenged on grounds of equity because of the limited sample of student work that it elicits. Since such tests are usually designed to fit into a normal school period, they have difficulty covering very much material. Typically, they opt for breadth rather than depth of coverage and require students to move quickly from one unrelated task to another, a cognitive skill not often called for in other contexts. The anxiety that many students experience as a result of the strict time limits is an additional impediment to their being able to show what they can do under normal circumstances.

By way of contrast, authentic assessment rewards sustained attention to an extended task. In the case of discourse testing, students typically prepare an essay on a single topic; and in the case of performance testing, they are usually asked to carry out

basic steps of a well-defined task such as a science experiment. As to documentation practices, the samples of student work are not limited to a single sitting and so can reflect the refinements more characteristic of real-world work. In the case of writing, for example, students can engage in the multiple processes that real writers work through, like gathering crucial information, exploring ideas, developing heuristic frames for writing, producing writing within those frames, revising the frames so that they form a coherent structure, producing more careful writing that adheres to this structure, and editing for style and cohesion. As students engage in a task over an extended period of time, they are able to produce more substantial work; thus, evaluation can, in principle, be more equitable.

In addition to including tasks that elicit sustained effort, authentic assessment usually leads to a more diverse sampling of what students can do. In the case of a writing portfolio, for example, many different kinds of work can be included such as formal modes of writing, including narration or exposition, as well as informal modes such as process journals. Such diversity insures more equitable evaluation, for it provides students a chance to demonstrate areas of strength that can be bypassed in more restricted forms of testing. This is a point that has been made by Smitherman (1991) in a longitudinal study of African-American student performance on the writing component of the National Assessment of Educational Progress (NAEP). According to her analysis, the writing samples of many African-American students reflect discourse strategies strongly marked for oral communication (Gates, 1987). She identifies, for example, their greater use of dramatic resources of language (e.g., "Darkness is like a cage in black around me, shutting me off from the rest of the world."). As the evaluation of their writing samples indicate, these strategies function more positively within narrative as opposed to exposition (i.e., African-American students have consistently performed better on the narrative task). It is for this reason that Smitherman is critical of NAEP for removing the narrative task from the writing test (particularly since the longitudinal data clearly show this is an area of strength for African-American students).

In emphasizing multiple samples of student work over an extended period of time, authentic assessment does introduce equity issues of its own. A fundamental issue has to do with individual responsibility for work. For instance, how can one determine the extent of help that a student has received in preparing a particular piece of discourse, an artifact, or a performance? As we earlier observed, this question can be approached in different ways. In certain instances, safeguards are built in to insure that the work belongs to a single

individual; a portfolio can, for example, include a monitored piece of writing (e.g., an essay written in class) which can be calibrated with the unmonitored pieces of writing (in carrying out such calibration, we need to be sensitive to how a testing situation, particularly the time pressure, limits what students are able to achieve). Moreover, students can be asked to document the development of at least one piece of writing in the portfolio; it is difficult for students to fabricate documentation for processes in which they have not engaged.[18]

In other instances, there is simply less concern that work samples be the product of a single individual. Indeed, a fundamental concern for many who advocate authentic assessment is to determine to what degree students can work effectively with others. One way of assessing this is to examine whether they can draw effectively on various kinds of social support in preparing a portfolio or exhibition. The means of assessing such student capacity is not well-developed, but it promises to become increasingly important in the years ahead. Especially important to determine is whether students, when offered critical evaluation by their teacher or other students, are able to use it in improving their work. Clearly the capacity to respond positively to critical evaluation is crucial to successful performance in the workplace and many other endeavors.

Public Criteria

In this country, conventional testing has long been shrouded in secrecy (Schwartz & Vor, 1990). The official rationale for this policy is that the integrity of individual tests must be insured. Since the same test is used on separate occasions, those who administer it go to considerable trouble to insure that what it contains is sufficiently protected. Test material can, in principle, be used in a scholarly publication such as this one. In practice, however, test publishers are often reluctant to grant such permission, claiming that any use of their material threatens test security. In preparing a recent publication, we requested permission to use a test item but were told we could use only a small portion of it. We appealed to a basic principle of discourse analysis—a text is most fairly evaluated when it is

[18] There are some signs, at least in a large urban area such as New York City, that new assessment practices may be engendering a new kind of commercial product: a prefabricated essay accompanied by various kinds of documentation on the processes of its development. As education varies its demands, the marketplace usually finds a way to respond.

treated as a whole—but our appeal was denied, and we were thus blocked from any serious intellectual discussion of the test material.[19]

Advocates of authentic assessment are generally opposed to the secrecy that surrounds conventional testing. They claim that students should know not only how they are evaluated but should receive adequate help in meeting the required standards. As we observed in our discussion of portfolio assessment in Vermont, the evaluative criteria are made public. Moreover, they are accompanied by samples of student work that illustrate how they are to be applied; these samples, in principle, help both teachers and students internalize the standards used in evaluation.

As admirable as a policy of public disclosure may be, it introduces a range of problems that have not been adequately addressed. In discussing the portfolio system in Vermont, we pointed out the difficulty of evaluating such features as voice and tone. When we examined the exemplars provided by the state benchmarking committee, we were often not clear about how the committee had applied the public criteria (see pp. 43-45); ideally, exemplars need to be accompanied with an explanation of how specific characteristics of the text have been used in reaching a judgment with respect to particular criteria.

Even if we find ways of making the use of criteria less problematic, our public disclosure of them still may not be as effective as anticipated. For one thing, it is no easy matter to present a set of criteria so that a broad range of students can understand them and apply them in their writing. One danger is that students will overreact to such standards as "organized from beginning to end" and produce writing that is chockfull of transitional words and other surface manifestations of organization. And what are fifteen-year-old students to make of the feedback that they do not have "a mature tone?" From a broad educational perspective, we question whether criteria built around analytic traits can really capture the complex ways in which readers respond to texts. It is for this reason that we have expressed a preference for a more holistic approach to evaluation (see pp. 45, 51).

---

[19] Some countries do not follow this policy of secrecy. In Japan, for example, the major exam for university admissions is published in the newspapers after each administration. Such publication stimulates a good deal of public debate about the legitimacy of certain tasks. The newspapers even publish challenges to certain tasks and the test makers' responses to these challenges (see Ingulsrud, 1988, for further discussion). In this country, the state of New York has passed legislation that requires public disclosure of college and university admissions tests such as the SAT; this legislation has not, however, led to newspapers actually publishing such tests, though they do occasionally discuss a controversial task or two.

The advantages of holistic scoring apply especially in those situations when the level of writing ability in a population is to be assessed, as is often the case in system-oriented assessment. In much classroom-oriented assessment, however, an analytic approach may well be preferable, especially when a teacher wishes to place pedagogical focus on particular aspects of writing, either for an individual student or for the class as a whole. The teacher can attend to highly specific features without being forced to assign them an arbitrary weight; these features can be properly explored as a means of sharpening a student's skills as a writer.

Apart from these evaluation problems, there are broad policy issues that accompany public disclosure of assessment tasks. An important issue is whether such disclosure may lead teachers to focus unduly on these tasks and thus narrow the curriculum in undesirable ways. Certainly those of us who work with alternative practices continuously face this problem. One way of handling it is to develop a sufficiently broad range of tasks so that teachers are less tempted to focus on a narrow set; this approach, however, introduces accountability problems of its own. To begin with, it takes an extraordinary amount of time—and talent—to develop good tasks, but even if we can develop a broad set of tasks, we are then faced with another equity issue—how can we reliably compare student performance on various kinds of tasks? This question becomes especially acute when assessment is used to carry out system-oriented functions such as placement or certification.

Multiple Human Judgments

In order to insure strict impartiality, conventional tests are designed so that human judgment is removed from evaluating test takers' responses. In actuality, human judgment is not so much removed from the evaluation process as it is displaced to an earlier stage. Someone still has to decide which tasks are to be included on the test, how they will be weighted, which answers are correct, and so on. These decisions get buried in the course of constructing a test and thus no individual remains who can be held accountable (Hill & Parry, in press). It is thus difficult for students to make an effective appeal when they believe that tasks on a conventional test are unfair.

By way of contrast, authentic assessment is designed so that students can appeal the evaluation they receive. Moreover, two or more individuals are generally required to evaluate student work. Even when a single individual does the evaluation, there is a

moderating system in place to review the initial evaluation; that is to say, a system in which individuals at higher levels use random procedures to check work at lower levels. In effect, greater equity is insured by an evaluation process that relies on multiple human judgments.

Nevertheless, certain problems are inevitable when assessment is carried out by teachers who work directly with the students. These teachers are, at times, the ones who initially evaluate student work. Once their judgment is in place, it can take on a life of its own. Most teachers operate—and rightly so—with a strong instinct to support the students whom they teach. The problem is that this support can lead to unfair assessment; students often receive a more positive evaluation than their work, strictly speaking, warrants. By the same token, teachers can allow negative feelings about an individual student to influence their judgment. Unfortunately, such feelings are difficult to monitor since they often remain unconscious, particularly where they operate across gender and ethnocultural lines. If the burgeoning research literature on teacher bias in the classroom is even moderately reliable, we are faced with a troubling question—how can we guard against such bias since gender and ethnicity are more palpably expressed in authentic assessment? Even if student names are removed in discourse testing, gender and ethnicity can still be revealed in style as well as substance; and in many assessment practices, the individual students are actually present in which case they can be directly observed.

There is a further problem with multiplying the number of individuals involved in evaluation. It may be designed to insure fair play, but the students being evaluated can develop a defensive posture. This problem is accentuated by the fact that they often associate on a daily basis with the evaluators who are, for the most part, their classroom teachers. Unless the evaluation process is carefully handled, students can develop a sense of being continuously judged. This problem becomes especially acute if student peers become involved in the evaluation process. Based on our own experience with peer review within a university setting, we are not altogether sanguine about its prospects among adolescents. It is for this reason that we have questioned the wisdom of including peer judgments in any kind of summative evaluation (see pp. 59-60). If peer evaluation is to be used, we recommend that it be used only in formative evaluation (e.g., in a seminar setting, students can offer constructive criticism on each other's work while it is in the process of development).

Relation to Classroom Learning

A major criticism of conventional testing, on grounds of equity as well as excellence, is that it is not sufficiently aligned with the curriculum. This criticism is, of course, less justified in the case of achievement testing such as the New York State Regents Exams which are deliberately constructed to reinforce a statewide curriculum, so they are, in principle, aligned with classroom teaching. In the case of nationally administered achievement tests, however, there is an attempt to test only what is widely shared by curriculums in different parts of the country. An aptitude test like the SAT which seeks to measure qualities that have been developed over an extended period of time is kept as independent of a particular curriculum as possible.

Those who support authentic assessment have chosen a quite different approach. They attempt to integrate assessment as fully as possible with what goes on in the classroom because they are convinced that such an approach not only motivates students to achieve excellence but provides them with a greater sense of fair play. Certainly any teacher is well aware that students are centrally concerned with whether they have been adequately prepared to do what assessment practices call for. One of their most common complaints is that they are assessed about matters on which they have not been properly taught.

This concern with adequate preparation raises difficult questions for educators who support the emerging practices of authentic assessment. Since these practices generally make greater demands on students, schools must be prepared to provide the support that enables students to meet these demands. For example, consider the increasing use of portfolio assessment with students who come from diverse ethnocultural backgrounds. A frequent criticism of conventional testing has been that it discriminates against such students, yet we have observed that these students can be overwhelmed by the intensive demands of a portfolio approach. They often do not have adequate resources at home with which to respond to these demands such as their own computer, a good supply of reference materials, ample space and time in which to work, and individuals who are prepared to provide feedback. Moreover, a portfolio approach can encourage these students to draw on discourse strategies that are effective in their own communities but do not fit what has been traditionally expected in an academic environment. In responding to these problems, schools must be prepared to provide resources that may be lacking at home; it is crucial, for example, that a teacher or tutor who is sensitive to varying discourse strategies respond to

what these students write. If these students are not provided adequate resources, authentic assessment may turn out to be even more inequitable than conventional testing.

We would like to close our discussion of equity on an optimistic note. If ethnoculturally diverse students are provided adequate support, we believe that the movement toward authentic assessment can lead to greater equity. Unlike conventional testing, it provides multiple samples of student work that are developed over time; but more importantly, it encourages students to draw on their own language, thought, and experience which if they are properly guided, can lead to more authentic expression. To our way of thinking, such authenticity is a key to developing the knowledge and skills that are required for successful participation in school and, ultimately, in the rapidly changing workplace.[20]

Efficiency

According to its proponents, conventional testing is highly efficient for two main reasons: (1) the multiple-choice format and (2) the statistical techniques known as psychometrics. The multiple-choice format facilitates both administration and scoring, making it possible, for example, to administer a single test in less than two hours to thousands of high school students throughout the nation and then score it and send the results to hundreds of college and university admissions offices in a matter of weeks—all at relatively low cost. Psychometric techniques, on the other hand, facilitate test construction as well as interpretation of test results. The performance of a sample population can be used in selecting appropriate tasks; whereas the performance of an actual population can be used in interpreting the performance of an individual test taker (i.e., the individual can be ranked according to norms derived from the larger group).

From the standpoint of those who support authentic assessment, such notions of efficiency are superficial and misleading. It masks the massive inefficiency that arises in an educational system when fundamental goals are distorted by inappropriate methods of

[20] In providing adequate support for ethnoculturally diverse students, schools can also counteract a major source of inequity that is developing around high-stakes testing in our society—the use of commercial coaching schools by students who can afford to pay for their services. Despite the evidence that coaching schools can be quite efficacious, this problem is virtually neglected in public discussion of testing and assessment policy (National Commission on Testing and Public Policy, 1990). This problem is even more acute in Japan, where coaching schools, called *juku*, constitute virtually an alternative system of education.

assessment. They thus argue for a deeper notion of efficiency, one that has to do with the degree to which assessment fosters good educational practices. We thus prefer a deeper notion of efficiency that can be expressed by the term "ecology." Indeed, this term can be used not simply in place of efficiency but in place of excellence and equity as well. It conveys an overarching principle that forces us to examine the complex role that assessment plays within the larger educational enterprise. As Darling-Hammond (1990) has pointed out, authentic assessment can provide crucial leverage in bringing about and sustaining educational reform. From an ecological perspective, its capacity as an agent of change could even be included as a major function of assessment alongside the others that we listed in the first section (see pp. 1–5).

Ecological Goals

In concluding this report, we would like to delineate three fundamental goals of education that authentic assessment can help to achieve:

1. reforming curriculum and instruction
2. improving teacher morale and performance
3. strengthening student commitment and capacity for self-monitoring

We will focus on the crucial role that authentic assessment can play in achieving each of these goals; we will also call attention to the difficulties that often arise when authentic practices are introduced.

Reforming Curriculum and Instruction

In using assessment to drive school reform, we must be continuously aware of how easily it can reinforce an endogenous model of education. When discussing assessment in relation to the principle of excellence, we pointed out that assessment practices, no matter how authentic, tend to be conservative. To counteract this tendency, we need to explore ways of linking such practices to a more exogenous model of education. One promising way of doing this is to set up student internships in appropriate settings.

In developing these internships, we must be sensitive to the flexible knowledge and skills that modern society demands. Technology is continuously restructuring what people do in the workplace, so it is not appropriate for students to concentrate on a highly specific

set of skills. Rather they must be prepared to acquire generic skills that allow them to participate effectively in what is often referred to as "the information society." On the one hand, they must know how to work with the information that technology produces (the human-machine interface); on the other, they must know how to use that information in communicating with others (the human-human interface).

Consider the role of a internship in facilitating useful knowledge and skills around computers. Ideally, students should plan and implement a project in the workplace. Such a project would be evaluated not only with respect to technical standards but according to how well it meets an actual need. A student assigned to a real estate agency might work on a database that agents can use to identify specific properties for prospective customers on the basis of factors such as cost, size, and location. In focusing on actual needs, authentic assessment acclimates students to confronting the kinds of questions they will eventually face once they enter the workforce. Does a database do what it is supposed to do? Can information be retrieved from it with simple procedures? Can it easily accommodate new information?

There are serious problems in developing appropriate internships in the workplace. To begin with, students can end up doing work that is not sufficiently challenging. Even when they are given challenging work, they often lack the kind of supervision they need. Many workplaces cannot easily support the supervision of student interns. It seems evident that the schools themselves have a critical role to play here. They provide the kind of detached environment that can encourage students to reflect on what they are learning during their internships. Such reflection is crucial if the students are to develop the more generic knowledge and skills that can be used in different work settings. The ultimate goal of any authentic assessment is for students to develop the critical habits of mind that will enable them to participate effectively in the larger society (Sizer, 1984).

Improving Teacher Morale and Performance

One of the major benefits of authentic assessment is that classroom teachers assume a central role in evaluating students. In traditional systems of education, teachers were accustomed to performing this role; but with the advent of conventional testing, their responsibility and, concomitantly, their authority was greatly reduced. One unfortunate consequence of this policy was that teacher morale—and in some instances, competence—was damaged. They came to feel that their own judgment about students was not trusted

and that conventional testing was developed as a kind of surrogate. It is for this reason, among others, that many teachers harbor a basic mistrust of such testing.

The emergence of authentic assessment has altered this situation. Alternative practices have restored to teachers their right—indeed, their responsibility—to evaluate students, for most of these practices depend upon teacher judgment. Moreover, teachers are actively involved in developing and implementing these practices. In many school districts throughout the country, teachers meet regularly to develop new assessment methods. Our own assessment project in Newburgh can be viewed as a pooling of the best practices that teachers engage in. Our regular meetings provide a powerful forum in which teachers can share ideas about the most effective ways of evaluating students. Such teacher collegiality is crucial in developing the commitment that must undergird authentic assessment (Hill, 1992).

There is a further benefit in the more active participation of teachers. Authentic assessment calls for a clinically–oriented style of observing student work. In the case of a portfolio, for example, teachers are involved in helping individual students plan various pieces of writing. Throughout the semester they read early drafts and make suggestions that students can work with as they prepare final drafts. As the semester nears the end, teachers can help them select pieces of writing to include; and once the portfolio has been submitted, teachers can annotate various pieces in preparation for a final conference. Such methods of assessment powerfully transform what classroom teachers do—they become clinicians who facilitate individual work. Within the best models of authentic assessment, teaching and evaluation become virtually indistinguishable.

It is important to stress that as teachers take on an increasing role in evaluating students, they must be provided adequate support. This support can take many forms, but it is crucial that they have sufficient time to insure that assessment practices are properly planned and implemented. In our own experience, nothing compromises assessment reform more than hurried and formulaic work. Moreover, teachers need academic support so that they are well informed about the technical aspects of assessment procedures. They also need to be sensitized to the range of responses that students are likely to make; it is especially important that teachers become more fully aware of ethnocultural differences in student response. We have claimed that authentic assessment can provide powerful leverage in reforming education, but this potential depends upon firm and continuous

support of teachers. In the absence of such support, the movement toward school reform will become simply another educational experiment that never achieved its potential.

Strengthening Student Commitment and Capacity for Self-Monitoring

It is generally accepted that some form of external evaluation is crucial in motivating students to work. We think this position should be reexamined, for it does not take sufficient account of the strong commitment that is engendered when students accept responsibility for their own work. In many models of authentic assessment, students are expected to select the projects on which they will be evaluated. Once they have made their selection, they are responsible for making sure that the necessary work gets done. The teacher is available for consultation, but students are expected to demonstrate self-reliance, which is, in fact, a crucial quality to be evaluated.

Moreover, students are increasingly expected to evaluate their own work within many approaches to authentic assessment. Particularly attractive is a model in which the individual student and teacher—or a professional supervisor in the case of a workplace internship—initially make independent evaluations of work that the student has done. Once these evaluations are in place, the student and teacher work together to achieve a mutually satisfying assessment (see pp. 65-67 for discussion of this model).

This model is particularly appropriate for urban schools, where students often come from ethnoculturally diverse backgrounds. The negotiating process can help to facilitate intercultural understanding. The teacher or work supervisor gains greater understanding of ethnocultural influences on how students carry out work; and the students, in turn, become more aware of the ways in which standards—whether in the school or the workplace—reflect ethnocultural norms that they may not have internalized. Such increased awareness can be quite liberating for both parties in negotiated evaluation. On the one hand, students discover that standards are culturally constituted practices that can have a certain functional value. On the other hand, those who evaluate students no longer carry the burden of imposing an absolute judgment on another human being.

Such an approach also provides a more realistic understanding of the role of assessment in human affairs. Students learn that assessment is intrinsically difficult to do, that it is dependent on many contingent factors and, thus, cannot be absolute. They come to accept continuous assessment, by themselves as well as others, as a natural and

necessary part in any work that they do. Such acceptance enables them, at least in principle, to acknowledge their own limitations yet strive to achieve work that makes full use of their knowledge and skills. In the final analysis, the major goal of any assessment should be to develop in students the capacity—and the commitment—to monitor their own work. Students will not learn to produce work of superior quality as long as standards remain external. It is only as they internalize standards that they are able to engage in the rigorous monitoring that insures excellent work. Ultimately, the commitment to excellence, though it may be richly nourished by external evaluation, comes from within.

A final word of caution—authentic assessment, when compared to conventional testing, makes far greater demands on both students and teachers. Indeed, those who support conventional testing have argued that authentic assessment consumes so much time that instruction is inevitably shortchanged. This way of thinking misconstrues the symbiotic relation between instruction and assessment. Each, when done properly, becomes the other; indeed, we can think of no more vital form of instruction than clinically–oriented assessment that teaches students how to monitor their own work. If this clinical spirit is to be humanely achieved in our classrooms, the very structure of education must be transformed. Such a transformation depends not only on our vision of what education should be but on the resources we are willing to commit to achieving this vision.

WORKS CONSULTED

America 2000: An education strategy. (1991). Washington, DC: U.S. Department of Education.

Archibald, D., & Newmann, F. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools.* Washington, DC: National Association of Secondary School Principals.

Aronowitz, R. (1984). Reading tests as texts. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 245-264). Norwood, NJ: Ablex.

Askin, W. (1985). *Evaluating the advanced placement portfolio in studio art.* Princeton, NJ: Advanced Placement Program.

Baron, J. (1990, June). *Blurring the edges among assessment, curriculum and instruction.* Paper presented at a symposium on "What's New in Large Scale Authentic Performance Assessment," Boulder, CO.

Berryman, S., & Bailey, T. (1992). *The double helix of education and the economy.* New York, NY: Teachers College, Columbia University, Institute on Education and the Economy.

Black, J., Kay, D. M., & Soloway, E. M. (1987). Goal and plan knowledge representations: From stories to text editors and programs. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive psychology and human computer interaction* (pp. 36-60). Cambridge: MIT Press.

Bloom, B. (1976). *Human characteristics and school learning.* New York, NY: McGraw-Hill.

Brewer, R. (1991). *Portfolio assessment in Vermont.* Presentation at a conference on portfolio assessment, Teachers College, Columbia University, New York, NY.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Burstall, C. (1986). Innovative forms of assessment: A United Kingdom perspective. *Educational Measurement: Issues and Perspectives, 5*(1), 17-22.

Camp, R. (1990). Thinking together about portfolios. *Quarterly of the National Writing Project and the Center for the Study of Writing and Literacy, 12*(2), 8-14, 27.

Cannell, J. (1987). *Naturally normed elementary achievement testing in America's public schools: How all fifty states are above the national average.* Daniels, WV: Friends for Education.

Cannell, J. (1989). *The "Lake Woebegone" report: How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Carlson, C. (1990). *Beyond high school: The transition to work.* Princeton, NJ: Educational Testing Service.

Chira, S. (1991, August 7). Educators draw outline for nationwide testing. *New York Times*, p. A-19.

Chira, S. (1992, October 16). Study finds standardized tests may hurt education efforts. *New York Times*, p. A-19.

Chittenden, E., & Courtney, R. (1989). Assessment of young children's reading: Documentation as an alternative to testing. In D. Strickland & E. Morrow (Eds.), *Emerging literacy: Young children learn to read and write* (pp. 107-119). Newark, DE: International Reading Association.

Coalition of Essential Schools. (1990). Performances and exhibitions: The demonstration of mastery. *Horace, 6*(3), 1-12.

Collins, A., Brown, J., & Holum A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator, 6*(12), 38-46.

Cooper, C. R. (Ed.). (1981). *The nature and measurement of competency in English.* Urbana, IL: National Council of Teachers of English.

Cooper, P. (1984). *The assessment of writing ability: A review of research.* Princeton, NJ: Educational Testing Service.

Costas, A. (1989). Reassessing assessment. *Educational Leadership, 46*(2), 13-18.

Coyle, M. (1992). *The New Jersey High School Proficiency Test in Writing: A pragmatic face on an autonomous model.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York, NY.

Cuban, L. (1991). *The misuse of tests in education.* OTA Contractor Report. Washington, DC: U.S. Government Printing Office.

D'Ydewalle, G., Swerts, A., & De Corte, E. (1983). Study time and test performance as a function of test expectations. *Contemporary Educational Psychology, 8*(1), 55-67.

Darling-Hammond, L. (1990). Achieving our goals: Superficial or structural reforms? *Phi Delta Kappan, 71*(2), 286-295.

Darling-Hammond, L. (1991). The implications of testing policy for educational quality and equality. *Phi Delta Kappan, 73*(3), 220-225.

Darling-Hammond, L., & Ancess, J. (forthcoming). *The implications of performance-oriented assessments for secondary school curriculum and teaching.* Berkeley: National Center for Research in Vocational Education, University of California at Berkeley.

Darling-Hammond, L., Lieberman, A., & Frelow, F. (1990). *Performance assessment: Guidelines and examples.* New York, NY: Center for School Reform, Teachers College, Columbia University.

Darling-Hammond, L., & Wise, A. (1985). Beyond standardization: State standards and school improvement. *Elementary School Journal, 85*(3), 315-336.

Department of Education and Science in England and Wales. (1989). *National curriculum: Task group on assessment and testing: A report.* London, England: Author.

Eckstein, M., & Noah, H. (1989). Forms and functions of secondary-school leaving examinations. *Comparative Education Review, 33*(3), 296-310.

Educational Testing Service. (1987). *Learning by doing: A manual for teaching and assessing higher order thinking in science and mathematics.* Princeton, NJ: Author.

Erickson, F. (1982). Audiovisual records as a primary data source. *Sociological Methods in Research, 11*(2), 213-232.

Farr, R., & Carey, R. (1986). *Reading: What can be measured?* Newark, DE: International Reading Association.

Feeney, T. (1984). *Rites of passage experiences.* Racine, WI: Walden III School.

Fillmore, C., & Kay, P. (1983). Text-semantic analysis of reading comprehension tests. Berkeley: University of California at Berkeley, Institute of Human Learning.

Flood, J., & Lapp, D. (1989). Reporting reading progress: A comparison portfolio for parents. *Reading Teacher, 4*(2), 508-513.

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

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Freedle, R., & Duran, R. (1987). *Cognitive and linguistic analyses of test performance.* Norwood, NJ: Ablex.

Gardner, H. (1983). *Frames of mind.* New York, NY: Basic Books.

Gardner, H. (1991). *The unschooled mind.* New York, NY: Basic Books.

Gates, H. (1987). *The signifying monkey: An Afro-American theory of criticism.* New York, NY: Oxford University Press.

Gee, J. (1992a). *The social mind: Language, ideology, and social practice.* New York, NY: Bergin & Garvey.

Gee, J. (1992b). *What is reading? Literacies, discourses, and domination.* Newton, MA: Literacies Institute.

Godowsky, S., Scarbrough, M., & Steinwedel, C. (1991). *The senior project: An exhibition of achievement.* New Castle, DE: Hodgson Vocational-Technical High School.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*(2), 119-132.

Haertel, E., Ferrara, S., Korpi, M., & Prescott, B. (1984). *Testing in secondary schools: Student perspectives.* Washington, DC: American Educational Research Association.

Haladyna, T., Nolen, S., & Haas, N. (1991). Raising standardized test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

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

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683-687.

Henderson, H., & Meier, D. (1990). *The senior institute handbook.* New York, NY: Central Park East Secondary School.

Hill, C. (1992). *Testing and assessment: An ecological approach.* Inaugural lecture as Arthur I. Gates Professor of Language and Education. New York, NY: Teachers College, Columbia University.

Hill, C., Anderson, L., Watt, Y., & Ray, S. (1989). *Reading assessment in adult education: Local detail versus textual gestalt* (LC Report 89-2). New York, NY: Teachers College, Columbia University, Literacy Center.

Hill, C., & Larsen, E. (1983). *What reading tests call for and what children do.* Washington, DC: National Institute of Education.

Hill, C., & Parry, K. (1988). *Reading assessment: Autonomous and pragmatic models of literacy* (LC Report 88-2). New York, NY: Teachers College, Columbia University, Literacy Center.

Hill, C., & Parry, K. (1989). Autonomous and pragmatic models of literacy: Reading assessment in adult education. *Linguistics and Education*, *1*(2), 233-283.

Hill, C., & Parry, K. (1992). Test at the gate: Models of literacy in reading comprehension tests. *TESOL Quarterly*, *26*(3), 36-53.

Hill, C., & Parry, K. (in press). *Testing and assessment: International perspectives on English language and literacy.* Harlow, UK: Longman.

Hoffmann, B. (1962). *The tyranny of testing.* New York, NY: Crowell-Collier.

Ingulsrud, J. (1988). *Testing in Japan: A discourse analysis of reading comprehension test items.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York, NY.

Koretz, D. (1988). Arriving in Lake Woebegone: Are standardized tests exaggerating achievement and distorting instruction? *American Educator*, *12*(2), 46-52.

Koretz, D., Madaus, G., Haertel, E., & Beaton, A. (1992). *National educational standards and testing: A response to the recommendations of the National Council on Education Standards and Testing*. Santa Monica, CA: RAND Corporation, Institute on Education and Training.

Lave, J. (1988). *Cognition in practice*. Cambridge, England: Cambridge University Press.

Lee, C. (1988). Testing makes a comeback. *Training*, *25*(2), 49-59.

Lidz, C. (Ed.). (1987). *Dynamic assessment*. New York, NY: Guilford Press.

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

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the National Society for the Study of Education* (p. 83-121). Chicago, IL: University of Chicago Press.

Madaus, G. (1991). The effects of important tests on students: Implications for a national examination or system of examinations. *Phi Delta Kappan*, *73*(3), 226-231.

Madaus, G., & Kellaghan, T. (1991). *Student examination systems in the European Community: Lessons for the United States*. OTA Contractor Report. Washington, DC: U.S. Government Printing Office.

Medina, N., & Neill, D. (1988). *Fallout from the testing explosion*. Cambridge, MA: FairTest.

Mitchell, R., & Stempel, A. (1991). *Six case studies of performance assessment*. OTA Contractor Report. Washington, DC: U.S. Government Printing Office.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill, MA: Author.

National Council on Education's Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: Author.

National Council on Vocational Education. (1990). *"Time for action": A business, industry, and education forum*. Washington, DC: Author.

Natriello, G. (1989). *What do employers want in entry-level workers? An assessment of the evidence* (Occasional Paper 7). New York, NY: Teachers College, Columbia University, National Center on Education and Employment.

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

New Jersey State Department of Education. (1990). *Grade 11 High School Proficiency Test examiner's manual*. Trenton: Author.

Newmann, F. (1991). Linking restructuring to authentic student achievement. *Phi Delta Kappan, 72*(2), 458-463.

Nickerson, R. (1989). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7.

Nuttall, D. (1992). Performance assessment: The message from England. *Educational Leadership, 49*(8), 54-57.

Office of Technology Assessment (OTA). (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Palincsar, A., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Perkins, D., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16-25.

Raizen, S. (1989). *Reforming education for work: A cognitive science perspective* (MDS-024). Berkeley: National Center for Research in Vocational Education, University of California at Berkeley.

Reif, L. (1990). Finding the value in evaluation: Self-assessment in a middle school classroom. *Educational Leadership*, *47*(2), 24-29.

Resnick, D., & Resnick L. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher*, *14*(4), 5-21.

Resnick, L. (1987a). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L. (1987b). Learning in school and out. *Educational Researcher*, *16*(9), 13-20.

Romberg, T., Zarinnia, E., & Williams, S. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions.* Madison, WI: National Center for Research in Mathematical Sciences Education.

Schwartz, J., & Viator, K. (1990). *The prices of secrecy: The social, intellectual, and psychological costs of current assessment practices.* A report to the Ford Foundation. Cambridge, MA: Harvard Graduate School of Education, Educational Technology Center.

Scribner, S. (1988). *Head and hand: An action approach to thinking.* New York, NY: Teachers College, Columbia University, Institute on Education and the Economy.

Scribner, S. (1991). *Knowledge acquisition at work.* New York, NY: Teachers College, Columbia University, Institute on Education and the Economy.

Secretary's Commission on Achieving Necessary Skills (SCANS). (1991). *What work requires of schools.* Washington, DC: U.S. Department of Labor.

Secretary's Commission on Achieving Necessary Skills (SCANS). (1992). *Learning a living: A blueprint to high performance.* Washington, DC: U.S. Department of Labor.

Shavelson, R., Baxter, G., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347-362.

Shavelson, R., Baxter, G., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), 22-27.

Shepard, L. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *73*(3), 243-247.

Sizer, T. (1984). *Horace's compromise: The dilemma of the American high school.* Boston, MA: Houghton Mifflin.

Sizer, T. (1986). Changing schools and testing: An uneasy proposal. In *The redesign of testing for the 21st century* (pp. 1-7). Princeton, NJ: Educational Testing Service.

Smith, M. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, *20*(5), 8-11.

Smitherman, G. (1991, October). *African-American students and writing tasks on the NAEP.* Presentation at Conference on African-American Language and Communication, Teachers College, Columbia University, New York, NY.

Snow, R. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, *18*(9), 8-14.

Stake, R. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan*, *73*(3), 243-247.

Stasz, C., McArthur, D., Lewis, M., & Ramsey, K. (1990). *Teaching and learning generic skills for the workplace* (MDS-066). Berkeley: National Center for Research in Vocational Education, University of California at Berkeley.

Test of adult basic education. (TABE). (1987). Monterey, CA: CTB Macmillan/McGraw-Hill.

Test of adult basic education: Examiner's manual. (1987). Monterey, CA: CTB Macmillan/McGraw-Hill.

Thorndike, E. (1913). *Educational psychology* (Vol. 1). New York, NY: Teachers College Press.

Traub, R., & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. In K. Ingenkamp & R. Jager (Eds.), *Tests and trends 8: Jahrbuch der padagogischen diagnostik* (pp. 128-159). Weinheim and Basel, Germany: Beltz Verlag.

U.S. Department of Labor. (1987). *Workforce 2000.* Washington, DC: National Institute of Education.

Valencia, S. (1991). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *Reading Teacher, 6*(3), 338-340.

Vermont State Board of Education. (1991a). *Analytic assessment guide.* Montpelier, VA: Author.

Vermont State Board of Education. (1991b). *This is my best: Vermont's writing assessment program pilot year report, 1990-91.* Montpelier, VA: Author.

Warren, G. (1979). Essay versus multiple-choice tests. *Journal of Research in Science Teaching, 16*(6), 563-567.

White, E. M. (1985). *Teaching and assessing writing.* San Francisco, CA: Jossey-Bass.

Widdowson, H. (1978). *Teaching language as communication.* New York, NY: Oxford University Press.

Wigdor, A., & Green, B. (Eds.). (1991). *Performance assessment for the workplace* (Vol. 1). Washington, DC: National Academy Press.

Wiggins, G. (1989a). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

Wiggins, G. (1989b, May). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wiggins, G. (1991a). *Current approaches to portfolio assessment.* Presentation at a conference on portfolio assessment: Teachers College, Columbia University, New York, NY.

Wiggins, G. (1991b). Standards, not standardization: Evoking quality student work. *Educational Leadership, 49*(2), 18-25.

Wolf, D. (1988). Opening up assessment. *Educational Leadership, 45*(4), 24-29.

Wolf, D. (1989). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(7), 35-39.

Wolf, K. (1991). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. *Phi Delta Kappan, 72*(1), 129-136.