Assessment: Authenticity, Context, and Validity

Let us at least demand, Mr. Wiggins suggests, that test-makers recognize their obligation to link their tests to the tasks, contexts, and "feel" of real-world challenges — in all their messiness.

BY GRANT WIGGINS

THERE IS an inescapable tension between the challenges presented by contextualized performance and conventional, large-scale, generic testing. Understanding is not cued knowledge; performance is never the sum of drills; problems are not exercises; mastery is not achieved by the unthinking application of algorithms. In other words, we cannot be said to understand something unless we can employ our knowledge wisely, fluently, flexibly, and aptly in particular and diverse contexts. As Lauren Resnick has put it, the two key assumptions of conventional test design — the decomposability of knowledge into elements and the decontextualization of

GRANT WIGGINS is president and director of the Center on Learning, Assessment, and School Structure (CLASS), a nonprofit education organization in Geneseo, N.Y. This article is adapted, with the permission of the publisher, from his book Assessing Student Performance: Exploring the Purpose and Limits of Testing, published in October by Jossey-Bass. For ordering information, write to Jossey-Bass Publishers, 350 Sansome St., San Francisco, CA 94104; ph. 415/433-1767 or fax 415/433-0499. ©1993, Grant Wiggins.
knowing (whereby it is assumed that if we know something we know it in any context) — are false. It may in fact make better sense to think of understanding as being more like good judgment or a disposition than the possession of information.

In other words, something more than face validity is at stake. The simplest way to sum up the potential harm of our current tests is to say that we are not preparing students for real, "messy" uses of knowledge in context — the "doing" of a subject. Or, as one teacher expressed it to me a few years ago, "The trouble with kids today is that they don't know what to do when they don't know what to do." That is largely because all our testing is based on a simplistic stimulus/response view of learning.

Two key words in this analysis are context and judgment. Competent performance requires both. It makes no intellectual sense to test for "knowledge" as if mastery were an unvarying response to unambiguous stimuli. That would be like evaluating trial judges only on the basis of their knowledge of law or doctors only on the basis of their recall of biochemistry. What we should be assessing is the student's ability to prepare for and master the various "roles" and situations that competent professionals encounter in their work.

But we should keep the test-maker's dilemma in mind: fidelity to the criteria on situations maximizes the complexity and ambiguity of the task requirements and maximizes the freedom to respond — both conditions that work against standardization and reliability. While this conflict between validity and reliability must never be construed as an either/or choice, it remains a design problem to be carefully negotiated.

Today, we are failing to negotiate the dilemma. Modern, professionally designed tests intended for national and state use tend to sacrifice validity for reliability. In other words, test-makers generally end up being more concerned with the precision of scores than with the intellectual value of the challenge. Thus the forms of testing and scoring used are indirect and generic, designed to minimize the ambiguity of tasks and answers.

But such forms of testing simply do not tell us what we need to know: namely, whether students have the capacity to use wisely what knowledge they have. This is a judgment that we can make only through tasks that require students to "perform" in highly contextualized situations that are as faithful as possible to criterion situations.

Deeper educational issues underlie these concerns. To assume that tests should assess whether all students everywhere have the same "knowledge" is to short-circuit a vital educational dialogue in a pluralistic and diverse society. Genuine intellectual performance is inherently personalized, and the meanings, strengths, and aspirations that we derive from an education are inherently idiosyncratic. Common knowledge is not the aim of any robust education for lifelong learning. If competence is more like contextual insight and good judgment than inert knowledge, we will need to rethink our reliance on short-answer, unambiguous items and one-time tests.

We should be seeking a more robust and authentic construct of "understanding" and a more rigorous validation of tests against that construct. We can begin by keeping in mind that the aim of education is to help the individual become a competent intellectual performer, not a passive "selector" of orthodox and pre-fabricated answers.

WHAT IS PERFORMANCE?

The word perform in common parlance means to execute a task or process and to bring it to completion. Our ability to perform with knowledge can therefore be assessed only as we produce some work of our own, using a repertoire of knowledge and skills and being responsive to the particular tasks and contexts at hand.

One way to illustrate the difference between drilled skills and performance ability is with an anecdote from my soccer-coaching career. It is common in soccer and other sports to practice drills related to exploiting a numerical advantage on offense. Every coach routinely does what are called "2 on 1," "3 on 2," or "4 on 3" drills — drills in which the offense has the ball and a numerical advantage. But mastery of these drills does not automatically translate into mastery in a game, as the following tale reveals.

Once, during a game early in a season, one of my better players had a series of opportunities to exploit such a numeri-
and required reading? You will be assessed on how well you support your claim about the accounts in the text, in response to the question, Are the accounts biased, inaccurate, or merely different from our usual viewpoint? [The students are given an extended excerpt from the "textbook," actually a text used in China in 1970, along with the following questions to consider.]

1. What can be said to be the most likely political influences on the authors' point of view? What evidence is there of those influences? How do they affect the authors' choice of language? Does the language reflect bias or an acceptable (but different) point of view? Explain your reasoning.

2. Why does it make sense, given the authors' perspective, that they pay particular attention to a) the Committee of Correspondence, b) the contribution of women, and c) the plight of "Indians" and "Negroes"? Are the facts accurate? Do they warrant that much attention in your view, or does such selective emphasis suggest a biased treatment? (How are these topics treated in the current text, and is the treatment there less biased or selective?)

3. You will be judged on the accuracy, aptness, and convincing qualities of your documentation and on the rhetorical effectiveness of your case. Be fair, but be an effective speaker and writer! A six-point scoring scale will be used for each dimension to be assessed: persuasiveness of evidence, persuasiveness of argument, rhetorical effectiveness of speech, and support material.

Leaving aside the feasibility of using such tasks in large-scale assessments (though this is done in other countries), the question raised by such a task is, Isn't this what we mean by performance? Clearly, students must "do" history in order to master the task. Mere control over what was in the textbook would neither prepare students adequately for doing such a task nor adequately represent the "criterion situation" of "doing" history. To merely "understand" what the textbook said is neither to understand the events themselves nor to understand what it means to do historical research on such events. In fact, when I have used this task in some fine suburban high schools, many students have been stunned to discover that the passage disagrees with their own textbooks— a sure indication that they have not "understood" or "done" history.

All tests must always point toward and be "enabling" of adult performance, in the sense suggested by Robert Glaser: "To place tests in the service of learning in a marriage of cognitive and psychometric theory, we must consider assessment . . . as measures of skills and dispositions that are essential to further learning. Once mastered, the skills and knowledge of a domain can become enabling competencies for the future." Thus we will need a better understanding of how understanding develops in a subject, as modeled by Harvard University psychologist William Perry's more general scheme for intellectual development in college.

Or, as Glaser has put it, "Modern learning theory is taking on the characteristics of a developmental psychology of performance changes . . . . In the future, achievement measurement will be designed to assess these changes."

We must begin to do a better job of testing for emerging competence by moving backward from the ultimate criterion performance, even when the student's current knowledge is rudimentary. We need to generalize from such approaches as those of Berlitz and other immersion language courses, which get the learner speaking and listening in context immediately and working toward the ultimate criterion of fluent contextual performance.

Not all hands-on work involves performance. Doing simplistic tasks that merely "cue" us for the desired bit of knowledge is not a creative employment of knowledge and skill, but a drill or exercise out of context. Consider, for example, a task from the original hands-on science test given a few years ago to all fourth-graders in New York.

In the first of five "stations," the student is given a cup, water, a thermometer, and other instruments. The student is expected, among other things, to accurately measure the temperature of the water. Toward what end? To what degree of precision? No one measures anything in general. Are we simulating the measuring of body temperature or of roasts in the oven? Are we making gravy or medicine? Purpose and context matter in our assessment of skill.

The test also requires that the answer

"I think a lot of this information should be on a 'need to know' basis."
be correct within two degrees above or below the "actual" water temperature. But why two degrees? Why such generosity? (And why is the student not told of this tolerance margin?) With no consequence or purpose to the measuring, there can be no appropriate margin of error. (A better form of this task would be to choose a recipe that requires just the right temperature of a liquid for a specific result to occur.)

UNDERSTANDING AND HABITS OF MIND

Another way to see how contextual understanding might be better tested is to consider the idea that understanding is inseparable from certain habits of mind. Habit is a word rarely used to describe academic mastery. We tend to reserve it now for the more pressing and oppressing concerns of affective issues or personal addictions. That is a pity, because a case can be made that academic learning and our assessment of it cannot be understood unless we see our aim as the formation of good habits of mind in each subject. 7 Yet we continue to do large-scale testing in a one-event format—a format inherently incapable of revealing whether a student is in the habit of performing up to standards—at our students' peril. 8

Through the work of Piaget, David Hawkins, and Eleanor Duckworth and in the literature on naive misconceptions in science (all well summed up and extended by Howard Gardner in The Unschooled Mind), we are reminded that many "obvious" adult ideas are counterintuitive. 9 Dewey, for example, was adamant that our ideas—even logical rules of inference—should be thought of as habits, sometimes not easily given up even when wrong: "The history of scientific beliefs shows that when a wrong theory once gets general acceptance, men will expend ingenuity of thought in butressing it with additional errors rather than surrender it." 10

The word habit suggests what assessment needs to become if it is to be directed toward thoughtful and effective understanding. 11 A higher-order habit is an intelligent proneness, not a reflex, in an inherently ambiguous situation. To say that academic learning aims ultimately to develop the habit of employing knowledge effectively alerts us to the fact that we need more than an assessment of learnedness; we need to assess for intellectual character. Any test of understanding should make it possible for us to know whether the student can accurately and willingly adapt knowledge to varying situations. This requires forms of testing that evoke and require for success the disposition to be critical in an ill-structured situation. Consider, for example, such open-ended prompts as this one used in a recent 12th-grade performance assessment in Connecticut: "How much does it cost to take a shower?"

To become progressively self-disciplined as a thinker, one needs more than an imposed rigor and the fruits of someone else's studies. One needs to acquire the habit of inquiring and engaging in discourse with care and thoroughness. What follows for assessment should be clear to anyone who grasps such an objective. How can I ever assess such a disposition in a one-sitting event? How can I acquire evidence as to whether students have "discipline" in a discipline without asking them to conduct inquiries and present findings? How can I assess students' learning unless I see whether they have learned how to learn in the subject in question? How can I assess their understanding without assessing their ability to ask and persist in trying to answer the right questions? Thus it is clear why a predominance of ill-structured tasks is essential for assessment of understanding: the lack of structure in the answering process is the only way I can discern whether a student has the necessary intellectual habits. If a focus on understanding is our goal, then secure, one-event, well-structured tasks with arbitrary criteria are at best dysfunctional.

It is one thing to learn how to respond to an unambiguous stimulus; it is another to become disposed to invoke the right habits of mind in a fluid performance context. Good teacher/coaches have students constantly moving back and forth between drill and a "whole" performance. In this way, students can learn what it feels like to be in the habit of skillful performing and can see the value of developing the newer, more difficult habits. We develop a repertoire of habits by continually practicing strategies in performance contexts, by using our judgment as to what works (and when and why), and by being constantly tested (through real or simulated performances).

UNDERSTANDING AS REVEALED THROUGH GOOD JUDGMENT

If performance requires a larger purpose, a rich context, and a repertoire wisely used, then effective performance is impossible without good judgment. Thus competence is testable only through tasks that demand good judgment in the use of knowledge. To test for understanding is to see if knowledge can be thoughtfully adapted. "Acquiring information can never develop the power of judgment," according to Dewey. "Development of judgment is in spite of, not because of, methods of instruction that emphasize simple learning... [The student] cannot get power of judgment excepting as he is continually exercised in forming and testing judgments." 12 Rather than merely having knowledge of general principles and of unambiguous cases to which they apply, "to be a good judge is to have a sense of the relative values of the various features of a perplexing situation." 13

Judgment involves effective adaptation to specific roles and situations; that is what we mean by competence. 14 We should recall that Binet defined intelligence as good judgment, "practical sense, initiative, the faculty of adapting one's self to circumstances"—what the French call bon sens, or good sense. 15 To develop a thoughtful control over performance depends not so much on learning and employing "knowledge and skills" but on having our judgment awakened and empowered through problems that demand judgment.

Judgment certainly does not involve the unthinking application of rules or algorithms—the stock in trade of all conventional tests. The effective performer, like the good judge, never loses sight of either relative importance or the difference between the "spirit" and the "letter" of the law or rules that apply. Neither ability is testable by one-dimensional items.

All performers must be judges. In a new case, how do we know which rules to apply? The performer, like the judge, "has to innovate, and where he innovates he is not operating from habit." 16 That is precisely what the soccer player mentioned above could not do: she could not discern when her "2 on 1" knowledge
should be applied. Unending drill tests prevent the development of the perception and adaptive intelligence needed to meet that sort of challenge.

Consider a different perspective on good judgment in the employment of skills—the account by the OSS (Office of Strategic Services) staff of what they called behavior "of the highest order of effectiveness." In the view of the OSS, effective use of knowledge and skill requires someone to "perceive and interpret properly the whole situation that confronts him...and...to coordinate his acts and direct them in proper sequence. . . ." [T]hey all require organization. . . . Consequently, [a test for effectiveness requires] tasks and situations which cannot be properly solved without organization." The OSS staff elsewhere described these tasks as requiring "mental operations on a higher integrative level; and since there is a difference between 'know-how' and 'can-do' we made the candidates actually attempt the tasks with their muscles or spoken words."1

In other words, any test must be judgment-based. After all, when asked to evaluate the correctness of an answer in real-world settings, we typically respond, "Well, it depends on. . . ." But this response is the antithesis of testing as we know it, where the aim is to use well-defined, unambiguous problems with a single apparent variable and a single correct answer so that we can ensure reliability and cost-effectiveness.

Let us now consider two tests—both designed for use in a general science course for middle school students, both from published texts/curricula. The first, an ill-structured test that requires thoughtful performance, is known as the "Sludge." In the Introductory Physical Science course, it is one of the major events.

In this multiday test, students have to chemically analyze a sludgelike mixture of unknown solids and liquids. (In one New Jersey school district, the Sludge takes up the last two weeks of class in June, serving as a very elaborate performance test of the year's work.) This is an ill-structured and authentic task par excellence: though the methods and criteria are quite clear to all students in the course, there are no pat routines, procedures, or recipes for solving the problem. Thus the test faithfully simulates a wide range of real-world "tests" of chemical analysis.

By contrast, consider a brief sampling of traditional items taken from a summative test of 200 items in a science textbook, to be given in a 90-minute period at the end of the eighth grade:

1. A general statement based on a hypothesis that has been tested many times is
   a. a conclusion
   b. scientific law
   c. scientific knowledge
   d. a theory
2. 125. Green plants and algae are
   a. omnivores
   b. herbivores
   c. consumers
   d. producers
3. 191. The level in the classification that is broader than species but narrower than family is
   a. class
   b. order
   c. genus
   d. phylum

Such exhaustiveness may well provide a certain superficial "content" validity, but the test bears little relationship to the practice of science, which is the ultimate "test" of one's knowledge.

Norman Frederiksen made this point a decade ago in writing about the "real bias" that exists in testing.18 His point was that the "tests" of life are more like the "Sludge" or the cost-of-the-shower problem than any neat and clean multiple-choice item. The "real bias" in testing is that tests are inherently restricted to unambiguous items, and their design ends up influencing what is taught and what is thought to be a problem.

A critic might respond, "But surely, before students can perform, we must give them drills and then tests concerning their mastery of the drills." This logical fallacy has probably done more harm than any other operant principle in American education. Look how Bloom's Taxonomy was (and still is) improperly construed as a chronology, though its authors warned against regarding it as such. Look how often syllabi unendingly postpone the student's exposure to genuine performance with knowledge in the name of "lessons" whose meaning is opaque and whose interest value is minimal.

All one has to do to see the fallacy of this way of thinking is to look at the adult performance world: musicians, athletes, architects, and doctors learn how to "perform with knowledge" by practicing the criterion performance. The Little Leaguer gets to play baseball; the medical student makes hospital rounds; the young artist starts by drawing. Any drill testing for these young performers is a means to an end; it is certainly not to be confused with the important performance itself.

Some roles and activities common to professional life might serve as possible "templates" for better test design. Perhaps a school district or state agency might store them in a nonsecure "bank." For example, to give students a taste of the role of museum curator, they might be asked to design an exhibit on a given topic or to compete with one another in designing grant proposals. To sample the problems faced by engineers, students might be given projects that require them to bid and meet specifications or to design and build a working catapult or herbarium. As a sample of the world of design and marketing, students might be asked to design an advertising campaign for a product (real or imaginary) or to design book jackets, blurbs, and so on for the books used in class. The possibilities are endless.

For assessment purposes, similar roles and situational challenges might be gen-
eralized into a more sophisticated and feasible classification system (or set of representative tasks), linked to role-based research such as that by Robert Gagné.20 Many school districts are now developing sets of learner-outcome criteria and standards that point in the same direction. Situational accuracy depends on the kind of “job analysis” rarely applied to K-12 education. Yet this kind of “role analysis” is at the heart of competency-based education: “The competency-based approach begins with the definition of the knowledge, skills, and attitudes required for successful performance in a particular role.”20 Or, as Gerald Grant put it in a book thoroughly analyzing competency-based education at the collegiate level, “Competence-based education . . . derives a curriculum from an analysis of a prospective or actual role in modern society.”21 Distinguished psychometricians have been making these points for years.22 Persistent unwillingness to design K-12 tests to match ultimate criterion situations was at the heart of David McClelland’s influential critique of standardized testing 20 years ago. He argued that validity coefficients for tests were too often derived from other indirect tests or from such predictors as college grades, which are still not the real measure. The best testing, he argued, is “criterion sampling . . . [T]here are almost no occupations that require a person to do word analogies. The point is so obvious that it would scarcely be worth mentioning, if it had not been obscured so often by psychologists.” The solution? “Criterion sampling means that testers have to get out of their offices . . . and into the field where they actually analyze performance into its components. If you want to test who will be a good policeman, go find what a policeman does.” McClelland knew well the difficulty of getting testers to honor his concern, however: “The task will not be easy. It will require psychological skills not ordinarily in the repertoire of the traditional tester. What is called for is nothing less than a revision of the role itself — moving it away from word games and statistics toward behavioral analysis.”23

Frederiksen made the same point a decade ago: “We need a much broader conception of what a test is if we are to use test information in improving educational outcomes,” he argued.24 Or, as the OSS staff put it, anticipating all the discussion well: “The best that can be done [given the limits of predictive testing] is to expose a man to a variety of situations of the same type as those he will find in the field.”25 The OSS staff noted particularly that traditional forms of paper-and-pencil tests failed to test for “effective intelligence” — the ability to select goals and to recognize the best means in context for attaining them coupled with quick resourceful thinking or good judgment. They were “prompted to introduce realistic tests of ability” when they observed inconsistencies between paper-and-pencil tests of from those that best predict job performance.”28 Our school-based tests are inaccurate and inappropriate as predictors for the same reason. It is time that test-makers were held more accountable for their methods of validation and required to do more careful trait and job analyses to justify both the form and content of tests they design.

AUTHENTICITY

What we require, therefore, are more general design criteria that can be used to frame challenges that are psychometrically useful but also more “authentic” —

The criterion of a good test is its congruence with reality.
insufficient or irrelevant. Problems require a repertoire of knowledge, good judgment in determining which knowledge is apt when and where, and skill in prioritizing and organizing the phases of problem clarification and solution.

• Tasks that require the student to produce a quality product and/or performance.

• Transparent or demystified criteria and standards. The test allows for thorough preparation as well as accurate self-assessment and self-adjustment by the student; questions and tasks may be discussed, clarified, and even appropriately modified, through discussion with the assessor and/or one's peers.

• Interactions between assessor and assessee. Tests ask the student to justify answers or choices and often to respond to follow-up or probing questions.

• Response-contingent challenges in which the effect of both process and product/performance determines the quality of the result. Thus there is concurrent feedback and the possibility of self-adjustment during the test.

• Trained assessor judgment, in reference to clear and appropriate criteria. An oversight or auditing function exists: there is always the possibility of questioning and perhaps altering a result, given the open and fallible nature of the formal judgment.

• The search for patterns of response in diverse settings. Emphasis is on the consistency of student work—the assessment of habits of mind in performance.

We might summarize these points by using the perhaps oxymoronic term “authentic simulations” to describe what we should be after in the design of educational tests. As Robert Fitzpatrick and Edward Morrison put it 20 years ago in their comprehensive review of issues in performance testing, we seek two things in authentic simulations: the “fidelity” of a simulation and the “comprehensiveness” with which the many different aspects of situations are replicated. Of course, any simulation, like any test, involves choices and compromises, and the purposes and budgetary or logistical constraints under which we test may cause us to settle for a lesser degree of each than is optimal.

But the problems are more than just practical. As Fitzpatrick and Morrison put it, “The dilemma of simulation is that increasing fidelity and comprehensiveness appear to increase validity but on the other hand with decreasing control [over the situation and possible responses] and thus reliability.”26 Tests are simplified of contextual “noise” to make scores more reliable; yet we need to maximize the fidelity and comprehensiveness of the simulation for reasons of validity.

We can thus learn to negotiate the dilemma in a way that is educationally sound only by gaining better insight into what it is we must be more faithful to: the setting in which the challenge is embedded or the constraints under which the student is expected to operate. If “generic performance” is a contradiction in terms and if judgment-based performance should be a major part of testing, then testers are going to have to think through the role of context in testing.

CONTEXT, CONSTRAINTS, AND AUTHENTICITY

In real life we use our intellect and acquired knowledge and skills in particular contexts to solve particular problems. As Arthur Chickering and Charles Claxton, researchers of competency-based learning, put it:

Competence is . . . situational and personal. This is the most critical principle. Competence levels and qualities are dependent upon situations and contexts. Particular contexts and situations interact with particular clusters of pre-dispositions and abilities brought by the person. The outcomes depend upon these complex interactions . . . A person who is “literate” in one culture can at the same time be “illiterate” in another.31

Over the past few years of thinking these matters through, I have come to believe that this claim is so true that testers should pay most attention to the second of my nine criteria of authenticity—that is, replicating or simulating the diverse and rich contexts of performance.

As I noted in talking about judgment, being competent requires sensitivity to context. For example, a doctor is not expert merely because he or she possesses a set of general rules and propositions or habits in the muscles—rules and habits called “medicine.” The doctor knows (or does not know) how to adapt relatively abstract guidelines of pathology and technique to each individual patient.

Support for heightened attention to contextual detail can be found in a variety of research sources. In addition to the report of the OSS staff, the competency-based model used at Alverno College makes “contextual validity” an essential part of the design problem.32 And John Seely Brown, Allan Collins, and Paul Duguid have argued the more comprehensive point that all cognition is “situated” in cultures and contexts, rendering decontextualized learning and assessment invalid and dysfunctional—not fully productive of useful learning. Since schoolwork is “very different from what authentic practitioners do” and since learning and testing tend to reflect the culture of school and not the culture represented by the field of study, “contrary to the aim of schooling, success within this culture often has little bearing on performance elsewhere.”33

A test may always be a contrivance, then, but it should not feel like one. Consider the best professional training and testing. Doctors and pilots are confronted with situations that replicate the challenges to be faced later, including vital complexities of human interaction. (For example, many of the simulations used for recertification of professional airline pilots involve working effectively with other crew members, because no one crew member has all the necessary information.) A context is thus realistic to the extent that we so accept the premises, constraints, and “feel” of the challenge that our desire to master it makes us lose sight of any contrivances or extrinsic factors—factors such as the reality that someone is evaluating us—in the same way that Outward Bound exercises and the publishing of a school newspaper for a journalism course do not feel contrived. Researchers have consistently found that this verisimilitude and the chance to feel efficacious are essential not only to producing one’s best performance but to student motivation.

Here is a simple example of one high school teacher’s initial attempt to design a performance task and of how the task evolved as a concern for context was introduced. The original task (in a global studies course) required students to design a trip to China or Japan. The purpose was to determine whether they had learned from their reading the most im-
portant things about one of the countries. But what kind of trip should be designed? For what customers? With what constraints of budget or time?

The teacher then refined the task: each student had a $10,000 budget to design a one-month cultural-exchange trip for students his or her age. Okay. But the purpose was still too abstract. What must the tour designers accomplish? Are they trying to design a tour in the abstract or a tour to really attract young people? The students were finally charged to be travel agents who would develop an extensive brochure, fully researching the cost and logistical information using the SABRE computer-reservations system (available through the school computers). One student noted during the project, "Boy, this is hard. Is this what real life is like?"

Paradoxically, the complexity of context is made manageable by contextual clues. For the student to have a clear sense of what kind of answer fits the problem at hand, detail is essential. Think how difficult it would have been for students to design a trip with no contextual clues. Put differently, in the best case studies, the problem is solvable only within the context provided. This is why business and law school cases are so difficult to write: they must be as faithful as possible to both the important and the unimportant facts of the situation. "A case typically is a record of a business issue which actually has been faced by business executives, together with surrounding facts, opinions, and prejudices upon which executive decisions depend."34 In other words, any criteria and standards required by a performance task should be clear and natural to the situation.

**THE AUTHENTICITY OF CONTEXTUAL CONSTRAINTS**

The most vital aspect of contextual fidelity has to do with the authenticity of the constraints placed on performance by the demands of mass testing. But most educational testing involves constraints that have little to do with fidelity to the criterion situation and everything to do with maintaining standardization of task and procedure. It is time that we looked at the validity questions raised by this unbalanced trade-off.

There are typically four kinds of constraints facing any performer. There are demands placed on us by others, whether or not we would make such demands of ourselves; there are limits on the time available to complete the task; there are limits (sometimes because of the situation and sometimes because of the time limits) on the human and material resources at our disposal; and there are limits on our ability to get guidance and feedback as we proceed.

We can pose the following, then, as a set of design questions: What are appropriate limits on the availability of time, of reference materials and resource people, and of prior knowledge of the tasks, criteria, and standards to be mastered? I am certainly not arguing that students should have unlimited access to all resources during testing. But let us ask, What kinds of constraints authentically simulate or replicate the constraints and opportunities facing the performer in context? When are constraints authentic, and when are they inauthentic? It is often a matter of degree, but the principle needs to be maintained and defended.

Consider the following guidelines for testing for synthesis from Benjamin Bloom, George Madaus, and Thomas Hastings:

> The student may attack the problem with a variety of references or other available materials as they are needed. Thus, synthesis problems may be open-book examinations, in which the student may use notes, references, the library, and other resources as appropriate. Ideally synthesis problems should be as close as possible to the situation in which a scholar (or artist, or engineer, etc.) attacks a problem he or she is interested in. The time allowed, conditions of work, and other stipulations should be as far from the typical, controlled examination situation as possible.35 (Emphasis added.)

Thus whatever assessors are testing in a 20-minute essay, it most certainly is not the ability to write. As those of us who write for a living know, writing is revision, a constant returning to the basic questions of audience and purpose — a process that is missing from standard writing tests (where there is no audience, no opportunity to reflect on each draft, and no real purpose).

The amount of time allowed for performing is not always what determines whether time constraints are reasonable or unreasonable; sometimes the issue is how that time is allotted. Is the limiting of a test to one sitting authentic? If writing is indeed revision, why not allow the writing assessment to occur over three or four days, with each draft graded? Many districts now do so, including such large school districts as those in Jefferson County, Kentucky, and Cherry Creek, Colorado.36

Restrictions on access to texts and human resources, on time for revision and reflection, and on opportunities to ensure that one's answer is apt and understood would seem to change what a test is measuring. What are we really testing in, say, the Advanced Placement exams, in which we deny students access to reference materials and human resources — despite the obvious availability of such things in almost all criterion situations that could be imagined in the particular subject matters? What can the exam results possibly tell us about students' ability to bring research to fruition, to sift through facts to discern the significant from the insignificant, or to use knowledge to good effect?

We do not need to keep all books and other materials from students if the task is genuinely authentic. For example, in many of Connecticut's performance tasks in mathematics, the key formulas are given to the students as background to the problem. And why not allow students to bring notes to an exam? Is this not precisely the sort of thing we really want to find out about students — whether they are organized, well-prepared, and effective at using what they know? (The test-makers' defense — that I am seeking to measure something different from what they claim to be measuring — is a dodge. What they are measuring is inappropriate if our aim is to see what students understand.)

Authenticity in testing, then, might well be thought of as an obligation to make the student experience questions and tasks under constraints as they typically and "naturally" occur, with access to the tools that are usually available for solving such problems.

**THE RELATIONSHIP BETWEEN AUTHENTICITY AND VALIDITY**

Attention to the authenticity of purposes and constraints in context and to the nature of understanding makes clear
why a performance-based task is not necessarily a valid or authentic test. Have we sampled the performance domain fairly or comprehensively? Would scores be different if we used different prompts or different kinds of tasks? Have we gathered sufficient evidence, using diverse forms and diverse settings, of the pattern of responses that indicates competence? These questions demand every test-maker’s attention.

And the problem is not limited to the selection of tasks. Most of the scoring rubrics that I have encountered seem invalid to me. In English classes, we score what is easy, uncontroversial, and typical — not necessarily what is apt for identifying exemplary writing or apt for the situational demands of real-world writing.

Consider New Jersey’s scoring criteria for essay writing and the descriptor for the top score on the scale. (New Jersey’s writing assessment is typical of many state and district rubrics now in use.) The criteria are: organization/content, usage, sentence construction, and mechanics. The descriptor for a top score reads as follows:

Organization/Content: Samples have an opening and closing. The responses relate to the topic and have a single focus. They are well-developed, complete compositions that are organized and progress logically from beginning to end. A variety of cohesive devices are present, resulting in a fluent response. Many of these writers take compositional risks resulting in highly effective, vivid responses.

Sentence Construction: Samples demonstrate syntactic and verbal sophistication through an effective variety of sentences and/or rhetorical modes. There will be very few, if any, errors in sentence construction.

Mechanics & Usage: Few, if any, errors.

What a bore. Little in this scoring system places a premium on style, imagination, or ability to keep the reader interested. Only the top score description mentions “effective and vivid” responses, instead of those criteria being woven through the whole rubric. Yet we see this limitation in almost every writing assessment, including those of the National Assessment of Educational Progress (NAEP). For example, in a review of student stories contained in portfolios as part of a pilot project to score locally completed work, assessors using the NAEP rubric were restricted to formal criteria. Here is a descriptor for a story that merits a score of 6 (the top level): “Paper describes a sequence of episodes in which almost all story elements are well developed (i.e., setting, episodes, characters’ goals, or problems to be solved). The resolution of the goals or problems at the end are [sic] elaborated. The events are presented and elaborated in a cohesive way.” Surely this is not the best description possible of a good story.

But habits of testing for merely formal problems run deep. In working with an English department on some schoolwide scoring rubrics, it took me two long sessions to get the teachers to admit that whether or not a paper was “interesting” was of primary importance. But they had never been willing to grade on that criterion, nor were they confident that such a criterion should be in a formal rubric.

The reader should not infer that I believe that criteria of the sort listed above do not matter. Of course, they are important. But they are merely necessary, not sufficient. As examples of better criteria that are more closely linked to the reason writers write and the effects writers hope to have on an audience, while still being mindful of formal criteria, consider clarity, persuasiveness, memorability, and enticingness, criteria offered by Allan Collins and Dieter Gentner. Note that these criteria, which flow from a careful analysis of the purpose of the task, will probably not be met if there are distracting errors of organization or mechanics. Yet these criteria correctly alert the writer to the fact that writing ought to be worth reading, not merely formally “correct.”

For scoring rubrics to be valid, the criteria have to be more than “face authentic.” A test should enable us to effectively and validly discriminate between performances of different degrees of quality. As a result, scoring rubrics must be based on a careful analysis of existing performances of varying quality. We must possess models of exemplary and not-so-exemplary performance and be able to tell the two apart on the basis of apt reasons. Our discriminations must be valid, not merely reliable. Thus we should be basing our judgments on the most salient and educative distinctions — not on those that are easiest and most uncon-

Many dissertations with 400 footnotes have nothing to say.


The context of testing

Contextual issues also relate to test administration itself. There is no such thing as an invariant and generic test situation — one in which students can be assumed to always reveal what they “know.” Contextual factors in the test situation itself can affect what is really being measured — especially if the test does not permit the student to explain answers. Messick and others have argued the issue more broadly by making the point that validity must be analyzed in terms of the context in which testing occurs and the consequences that accrue from it. Not only is the meaning of test scores important, but so are the “relevance, utility, import . . . and the functional worth of scores in terms of the social consequences of their use.”

Once we grasp the fact that student responses are colored by the particular task and setting, the implications, according to Messick, are considerable: “We are thus confronted with the fundamental question of whether the meaning of a measure is context-specific or whether it generalizes across contexts [since the] very nature of the task might be altered by the operation of constraining or facilitating factors in the specific situation.” That is why validity inheres in the interpretation of a score and is not truly a property of the test itself. Messick soberly concludes by urging that “the role of context in test interpretation and test use be repeatedly investigated or monitored as a recurrent issue.” Can test-makers honestly say that they do such monitoring regularly?

Consider the following example of what happens when the context of testing is not considered. The kindergarten teachers in Ellenville, New York, were puzzled by the results of a commercial standardized test used in the district. Almost every student had gotten one simple question wrong. The question that caused the students so much trouble seemed easy enough: “Which one of the animals is a farm animal?” (In standardized testing for young children, the multiple choices are in the form of pictures, the “right” one to be selected after the test administrator reads each question.) The choices: pictures of a whale (or a porpoise?) diving in the water, a giraffe, and a chicken. Why didn’t the students select the chicken? Because not more than 20 miles from Ellenville is the Catskill Game Farm, where the star attraction (also represented on its large billboards in the area) is — you guessed it — a giraffe.

What seemed the most reasonable answer to the students was correct in context; what seemed the only apt answer to the test-maker turned out to be wrong in a specific context. Therein lies an inherent problem with tests that are both generic and nonresponsive. In a country with no agreed-upon universal syllabi or texts, test questions must be stripped from their natural setting. Yet, by depriving students of situational feedback and detail, we violate one of the most basic norms of social interaction. All questions are normally asked in context, often assuming a purpose, a culture, an audience, and various situational constraints. Suppose your math teacher says, “Do the odd problems for tonight’s homework.” We assume that she means the odd-numbered ones, not the most bizarre ones (though students can be masters at exploiting such inherent ambiguity to their benefit).

If we view the giraffe story as evidence of a mistake by the test company, we fail to grasp the danger of traditional generic testing. As Messick’s argument about context suggests, we may be assuming far too much about the stability and transferability of student knowledge and too little about the influence of testing conditions. The particular “mistake” concerning the giraffe may be a sign that important arguments about the fundamental question of context have not gotten an adequate hearing. Tests that are designed to yield stable scores — without regard to the local syllabus, culture, and milieu

"Boy! A few more like that and I’ll be ready for Gamblers Anonymous."
— may end up testing only a trivial residue of a context-bound education.

Ironically, tests of this sort make test items harder for students than they normally would be, because all typical contextual cues and responses are removed. And those researchers who caution us in the use of performance tests — advising that there is inadequate generalizability when tasks vary slightly — may be looking through the wrong end of the telescope. Intellectual performance may be more contextually sensitive — and hence unstable — than we have heretofore been able to see or willing to admit.

Thus it is not proper to say that a student either does or does not "possess" knowledge. Rather, the test-taker acts knowledgably or ignorantly — in context. Context not only enables the student to know whether "chicken" or "giraffe" is the right answer in this case; it also lets the scorer know whether an answer is right or wrong in this case. What, then, is being assessed when the student answers a question but the judge and the environment are mute? Certainly not competence.

The use of the word understand as opposed to know makes the point more clearly: we do not understand things in general; we understand (or misunderstand) a person or an answer in context. Perfectly bright and able people who are effective in one setting or job can screw up in another: competency is context-bound. There is an obligation on testers, therefore, to seek each test-taker's rationale for answers. If they do so, there is no good epistemological or cognitive reason to assume that test scores should be stable (as most test programs tacitly assume) if we vary the task or context even slightly.

The argument for authenticity that I and others have made should thus be understood as something more substantial and less naive than some measurement folks would have us believe. If validity refers to the implications or consequences of the inferences made, these issues cannot be ignored. As Messick puts it, in quoting Cronbach:

"The bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences. You . . . may prefer to exclude reflection on consequences from meanings of the word validation, but you cannot deny the obligation." But we would prefer a somewhat stronger phrasing, because the meaning of validation should not be considered a preference.

**TASK WORTHINESS AND INCENTIVES: FACE VALIDITY REVISITED**

In a student-centered view of assessment, perhaps we ought to resurrect an old concept that is now often pooh-poohed by psychometricians: face validity. Is the test, "on the face of it," a proper test of the test is judged relevant to its objectives . . . can affect examinee cooperation and motivation. . . . Therefore, it is argued that face invalidity should be avoided whenever possible.47

Anna Anastasi has probably written the most about face validity and its role in validation. In the most recent edition of her textbook on psychological testing, she argues that, while face validity "is not validity in the technical sense" (since it refers "not to what the test actually measures but to what it appears to measure"), it is vital for "rapport and public relations." In fact, she argues that "face validity itself is a desirable feature of tests."

Perhaps we ought to resurrect an old concept: face validity.

She too cites examples of the negative reactions of adults to items that were too much like school items; they were "frequently met with resistance and criticism." She goes on to say more forcefully that, "if test content appears irrelevant, inappropriate, silly, or childish, the result will be poor cooperation, regardless of the actual validity of the test" as (emphasis added).

It is this last phrase that needs to be pondered. To take an extreme case, why would we assume that a test is "technically valid" if it is universally ridiculed and resisted by those who take it? How can an inference about a score not be conditioned by the user's response to the test? For example, if performance tasks are in fact "far more likely to elicit a student's full repertoire of skills," as Howard Gardner's research shows, then why is the validity of tests that do not evoke these responses not open to question?49

What if indirect forms of assessment so distance some students from contextual "tests" of understanding that they lose interest in the ultimate criterion? While the relative merits of directness and indirectness in test construction have been argued on technical grounds in the testing liter-
nature, there is no mention of the possible effect of this technical factor on the test-takers.30

Although face validity should be considered, to focus only on it is to miss a more important point about the incentives to perform well that might inhere in more authentic forms of assessment and that might change the implications of scores. For example, Gardner observes that assessment in the context of students' working "on problems, projects, or products which genuinely engage them" can "hold their interest and motivate them to do well," suggesting that a more substantive question about validity is at stake.51 And John Raven is prepared to question the validity of any test that ignores motivational issues: "Important abilities demand time, energy, and effort. As a result, people only display them when they are undertaking activities which are important to them. It is meaningless to attempt to assess a person's abilities except in relation to their valued goals."52 Raven acknowledges that such views are in "sharp conflict" with traditional views, but there is a commonsense appeal here that deserves better exploration at least.

It therefore seems not only reasonable and fair but also contributive to more comprehensive validation to give the (older) student and/or the teacher a chance to judge a test's appropriateness. (New York does give teachers this opportunity in the Regents Examinations.) I have devised a simple questionnaire to enable teachers to assess their students' sense of a test's aptness and fairness. It can be given out after every major test in middle or high school.53

I am not proposing that students have the final say, nor would I claim that their judgment is necessarily accurate or technically informed. What I am asking is that we take account of the test-taker's point of view as one factor in considering the matter of validity. Many respected testing specialists have advocated such techniques as part of all pilot testing.54

How one's work and talents are judged is of paramount concern to everyone. Thus fairness demands that the test-taker's responses be solicited and pondered. We routinely assume that adults have the right to some say in performance appraisal — often through formal negotiations. So why should students be perpetually shut out of the discussion?

THE 'VALUE' IN CONCERNS ABOUT VALIDITY

One reason that concerns about validity are so easily finessed is that they demand more than technical expertise in measurement. Any inference drawn about test results is a complex act of judgment that involves the consideration of different kinds of data and our intellectual values. I am not arguing that indirect tests are inherently defective or prone to invalid inferences about performance. I am fully aware that certain constructs, such as critical thinking or reading comprehension, do not easily admit of direct testing. I am also aware that an indirect test can yield positive correlations with some criterion situations (e.g., vocabulary tests as predictors of verbal-role success). My fear is that validity in mass educational testing has deteriorated into an excessive concern with content validity, the use of questionable methods for obtaining responses that illustrate certain constructs, and mere correlations with other indirect tests (in an endless circle of results on questionable tests being used to validate other questionable tests).

At bottom is a major philosophical problem about the purpose of schooling — and hence of testing. Is schooling meant to yield common knowledge? If so, then it makes perfect sense to think of tests as properly focusing on what students hold in common. But what if education is seen to be a personal, idiosyncratic affair, where the meaning and personal effectiveness that I derive from coursework is more important than what knowledge we all hold in common? In that case, any kind of standardized, indirect test would make no sense. What could we possibly mean by a standardized test of the meaning of educational experience?

Consider Albert Shanker's view that we should think about achievement as scouts think of merit badges: achievement should be validated by a person's demonstrated ability to use knowledge in the field. "That's the kind of knowledge that doesn't leave you," according to Shanker.56 The result would be a personalized collection of badges — even if each badge requirement is standardized.

This is the deeper issue raised by the report of the Secretary's Commission on Achieving Necessary Skills with regard to the meaning of a transcript.56 If a transcript is really better thought of as a résumé, as the SCANS report claims, then what does that suggest about testing, about teaching, about curriculum design, and about the current myopic search for national standards? At the very least it suggests that a penchant for testing everyone on the same things is misguided.
While the issue is debated further, let us at least demand that test-makers recognize their obligation to link their tests to the tasks, contexts, and “feel” of real-world challenges—in all their messiness.


8. Readers might think I am forgetting about test reliability. On the contrary, the standard psychometric conception of reliability assumes that the performer is always reliable and that only scores can be unreliable. This is a fundamental epistemological error as well as a violation of common sense.


11. Let me offer a caution to the language-sensitive reader: I am speaking about higher-order habits—what are often referred to as “dispositions.” For a good account of the technical differences between “habits” and “dispositions,” see Gilbert Ryle, The Concept of Mind (London: Hutchinson House, 1949); and J. Passmore, The Philosophy of Teaching (Cambridge, Mass.: Harvard University Press, 1980).


20. This definition, used by the Fund for the Improvement of Post-Secondary Education, was formulated by T. Corcoran and is quoted in Nicke et al., p. 10.


25. Office of Strategic Services, p. 42.

26. Ibid.

27. Ibid.


31. Chickering and Claxton, p. 11.


36. Yes, yes, I know that the issue is really one of cheating. Let the teacher “sign off” on the papers, then, certifying authorship, as schools have long done in Australia (and now in Vermont) where the school report is built, in part, out of local work submitted to external examiners.


42. Reliability problems abound in testing our youngest children this way (e.g., one must constantly make sure that the students are looking at the right line of pictures), which is why many districts and states have banned such testing.

43. For the definitive account of “knowledge” as intelligent performance, not as the “mental” application of declarative propositions to situations, see H. S. Berg, Knowledge and Inquiry (New York: Basic Books, 1965).

44. For the view that the idea of authenticity is a big fuss about a naive idea, see F. L. Finch, “Issues in Educational Performance Evaluation,” in idem, pp. 89-138.


49. Gardner, p. 93.

50. Millman and Greene, p. 348.

51. Gardner, p. 93.


53. This questionnaire appears on page 246 of Assessing Student Performance, the book from which this article is adapted.

