A True Test: Toward More Authentic and Equitable Assessment

As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching national intellectual standards, Mr. Wiggins warns, student performance, teaching, and our thinking and discussion about assessment will remain flaccid and uninspired.

BY GRANT WIGGINS

WHEN AN educational problem persists despite the well-intentioned efforts of many people to solve it, it's a safe bet that the problem hasn't been properly framed. Assessment in education has clearly become such a problem, since every state reports above-average scores on norm-referenced achievement tests and since everyone agrees (paradoxically) that such tests shouldn't drive instruction but that their number and influence should nevertheless increase.[1] More ominously, we seem unable to see any moral harm in bypassing context-sensitive human judgments of human abilities in the name of statistical accuracy and economy.

We haven't moved beyond lamenting these problems, because we have failed to stop and ask some essential questions: Just what are tests meant to do? Whose purposes do they (and should they) serve? Are large-scale testing programs necessary? When are tests that are designed to monitor accountability harmful to the educational process? Need they be so intrusive? Is there an approach to upholding and examining a school's standards that might actually aid learning?

But we won't get far in answering these questions until we ask the most basic one: What *is* a true test? I propose a *radical* answer, in the sense of a return to the roots; we have lost sight of the fact that a true test of intellectual ability requires *the performance of exemplary tasks*. First, authentic assessments replicate the challenges and standards of performance that typically face writers, businesspeople, sci-

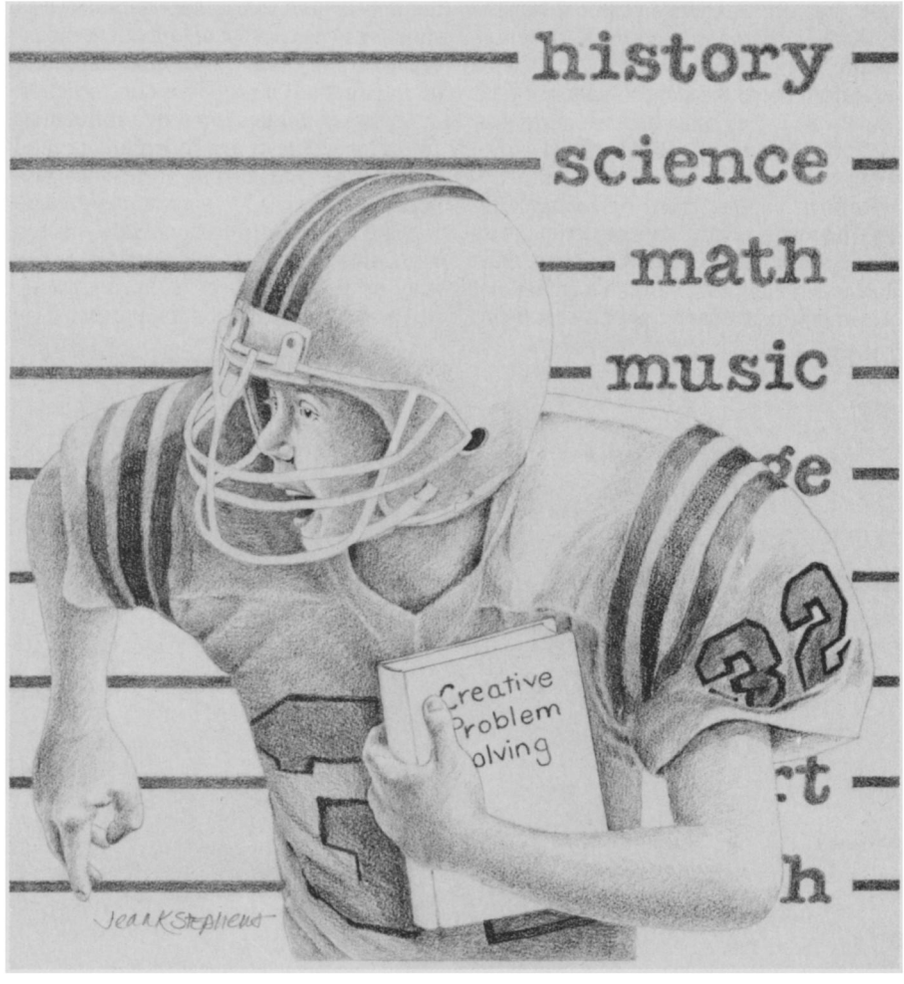


Illustration by Jean K. Stephens

GRANT WIGGINS is a senior associate with the National Center on Education and the Economy, Rochester, N.Y., and a special consultant on assessment for the Coalition of Essential Schools.

entists, community leaders, designers, or historians. These include writing essays and reports, conducting individual and group research, designing proposals and mock-ups, assembling portfolios, and so on. Second, legitimate assessments are responsive to individual students and to school contexts. Evaluation is most accurate and equitable when it entails human judgment and dialogue, so that the person tested can ask for clarification of questions and explain his or her answers.

A genuine test of intellectual achievement doesn't merely check "standardized" work in a mechanical way. It reveals achievement on the essentials, even if they are not easily quantified. In other words, an authentic test not only reveals student achievement to the examiner, but also reveals to the test-taker the actual challenges and standards of the field.

To use a medical metaphor, our confusion over the uses of standardized tests is akin to mistaking pulse rate for the total effect of a healthful regimen. Standardized tests have no more effect on a student's intellectual health than taking a pulse has on a patient's physical health. If we want standardized tests to be authentic, to help students learn about themselves and about the subject matter or field being tested, they must become more than merely indicators of one superficial symptom.

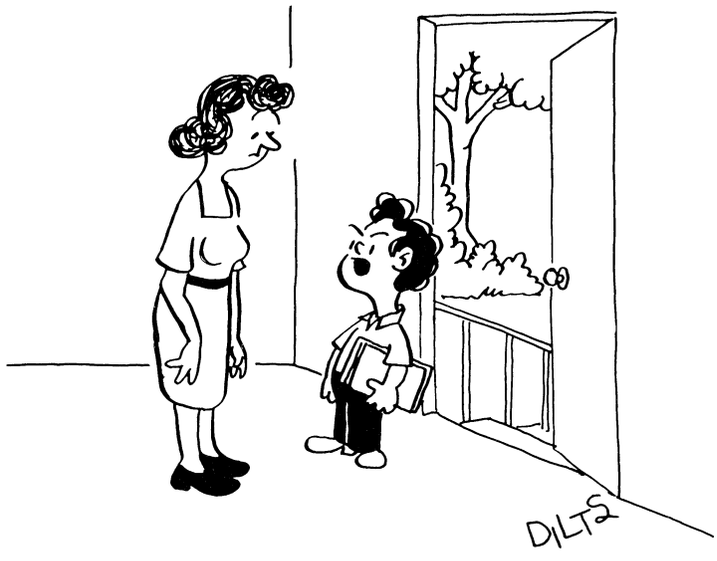
Reform begins, then, by recognizing that the test is central to instruction. Any tests and final exams *inevitably* cast their shadows on all prior work. Thus they not only monitor standards, but also set them. Students acknowledge this truth with their plaintive query, Is this going to be on the test? And their instincts are correct; we should not feel despair about such a view. The test always sets the de facto standards of a school despite whatever else is proclaimed. A school *should* "teach to the test." The catch is that the test must offer students a genuine intellectual challenge, and teachers must be involved in designing the test if it is to be an effective point of leverage.

SETTING STANDARDS

We need to recognize from the outset that the problems we face are more ecological (i.e., political, structural, and economic) than technical. For example, Norman Frederiksen, a senior researcher with the Educational Testing Service (ETS), notes that "situational tests are not widely used in testing programs because of considerations having to do with cost and efficiency."[2] In order to overcome the resistance to using such situational tests, we must make a powerful case to the public (and to teachers habituated to short-answer tests as an adequate measure of ability) that a standardized test of intellectual ability is a contradiction in terms. We must show that influential "monitoring" tests are so irrelevant (and even harmful) to genuine intellectual standards that their cost – to student learning and teacher professionalism – is too high, however financially efficient they may be as a means of gathering data.

The inescapable dilemma presented by mass testing is that using authentic standards and tasks to judge intellectual ability is labor-intensive and time-consuming. Examiners must be trained, and multiple, contextual tests of the students must be conducted. Genuine tests also make it more difficult to compare, rank, and sort because they rarely involve one simple, definitive test with an unambiguous result and a single residue number. Therefore, as long as tests are thought of only in terms of accountability, real reforms will be thwarted. After all, why do we need to devise more expensive tests if current data are reliable? When we factor in the self-interest of test companies and of colleges and school districts, we can see that resistance to reform is likely to be strong.

Using authentic standards and tasks to judge intellectual ability is labor-intensive and time-consuming.



"I don't understand all the fuss about my repeating third grade. Mr. Wilkins has been there for six years."

The psychometricians and the accountants are not the villains, however. As I have noted elsewhere, teachers fail to understand their own unwitting role in the growth of standardized testing.[3] Mass assessment resulted from legitimate concern about the failure of the schools to set clear, justifiable, and consistent standards to which it would hold its graduates and teachers accountable. But the problem is still with us: high school transcripts tell us nothing about what a student can actually do. Grades and Carnegie units hide vast differences between courses and schools. An A in 11th-grade English may mean merely that a student was dutiful and able to fill in blanks on worksheets about juvenile novels. And it remains possible for a student to pass all of his or her courses and still remain functionally and culturally illiterate.

But the solution of imposing an efficient and "standard" test has an uglier his-

tory. The tests grew out of the "school-efficiency" movement in the years between 1911 and 1916, a time oddly similar to our own. The movement, spearheaded by the work of Franklin Bobbitt, was driven by crude and harmful analogies drawn from Frederick Taylor's management principles, which were used to improve factory production. Raymond Callahan notes that the reformers, then as now, were far too anxious to satisfy external critics and to reduce complex intellectual standards and teacher behaviors to simple numbers and traits.[4] Implicitly, there were signs of hereditarian and social-class-based views of intelligence; the tests were used as sorting mechanisms at least partly in response to the increased heterogeneity of the school population as a result of the influx of immigrants.[5]

The "standards" were usually cast in terms of the increased amount of work to be demanded of teachers and students. As George Strayer, head of the National Education Association (NEA) Committee on Tests and Standards for School Efficiency, reported, "We may not hope to achieve progress except as such measuring sticks are available." A school superintendent put it more bluntly: "The results of a few well-planned tests would carry more weight with the businessman and parent than all the psychology in the world."[6]

Even with unionization and the insights gained from better education, modern teachers still fall prey to the insistent claims of noneducation interests. The wishes of college admissions officers, of employers, of budget makers, of schedulers, and even of the secretaries who enter grades on computers often take precedence over the needs of students to be properly examined and the needs of teachers to deliberate and confer about effective test design and grading.

Thus, when teachers regard tests as something to be done as quickly as possible after "teaching" has ended in order to shake out a final grade, they succumb to the same flawed logic employed by the test companies (with far less statistical justification). Such acquiescence is possible only when the essential ideas and priorities in education are unclear or have been lost. If tests serve only as administrative monitors, then short-answer, "objective" tests – an ironic misnomer[7] – will suffice (particularly if one teaches 128 students and has only a single day in which to grade final exams). However, if a test is seen as the heart and soul of the educational enterprise, such reductionist shortcuts, such high student/teacher ratios, and such dysfunctional allocation of time and resources will be seen as intolerable.

Schools and teachers do *not* tolerate the same kind of thinking in athletics, the arts, and clubs. The requirements of the game, recital, play, debate, or science fair are clear, and those requirements determine the use of time, the assignment of personnel, and the allocation of money. Far more time – often one's spare time – is devoted to insuring adequate practice and success. Even in the poorest schools, the ratio of players to interscholastic coaches is about 12 to 1.[8] The test demands such dedication of time; coaching requires one-to-one interaction. And no one complains about teaching to the test in athletic competition.

We need to begin anew, from the premise that a testing program must address questions about the inevitable impact of tests (and scoring methods) on students and their learning. We must ask different questions. What kinds of challenges would be of most educational value to students? What kinds of challenges would give teachers useful information about the abilities of their students? How will the results of a test help students know their strengths and weaknesses on essential tasks? How can a school adequately communicate its standards to interested outsiders and justify them, so that standardized tests become less necessary and less influential?

AUTHENTIC TESTS

Tests should be central experiences in learning. The problems of administration, scoring, and between-school comparisons should come only after an authentic test had been devised – a reversal of the current practice of test design.

If we wish to design an authentic test, we must first decide what are the actual performances that we want students to be good at. We must design those performances first and worry about a fair and thorough method of grading them later. Do we judge our students to be deficient in writing, speaking, listening, artistic creation, finding and citing evidence, and problem solving? Then let the tests ask them to write, speak, listen, create, do original research, and solve problems. Only then need we worry about scoring the performances, training the judges, and adapting the school calendar to insure thorough analysis and useful feedback to students about results.

To design an authentic test, we must first decide what are the actual performances that we want students to be good at.

This reversal in thinking will make us pay more attention to what we mean by *evidence of knowing*. Mastery is more than producing verbal answers on cue; it involves thoughtful understanding, as well. And thoughtful understanding implies being able to do something effective, transformative, or novel with a problem or complex situation. An authentic test enables us to watch a learner pose, tackle, and solve slightly ambiguous problems. It allows us to watch a student marshal evidence, arrange arguments, and take purposeful action to address the problems.[9] Understanding is often best seen in the ability to *criticize* or extend knowledge, to explain and explore the limits and assumptions on which a theory rests. Knowledge is thus displayed as thoughtful know-how – a blend of good judgment, sound habits, responsiveness to the problem at hand, and control over the appropriate information and context. Indeed, genuine mastery usually involves even more: doing something with grace and style.

To prove that an answer was not an accident or a thoughtless (if correct) response, multiple and varied tests are required. In performance-based areas we do not assess competence on the basis of one performance. We repeatedly assess a student's work – through a portfolio or a season of games. Over time and in the context of numerous performances, we observe the *patterns* of success and failure and the reasons behind them. Traditional tests – as arbitrarily timed, superficial exercises (more like drills on the practice field than like a game) that

are given only once or twice – leave us with no way of gauging a student's ability to make progress over time.

We typically learn too much about a student's short-term recall and too little about what is most important: a student's habits of mind. In talking about *habits of mind*, I mean something more substantive than "process" skills divorced from context – the formalism decried by E. D. Hirsch and others. For example, a new concept – say, irony or the formula F = ma – can be learned as a habit or disposition of mind for effortlessly handling information that had previously been confusing.[10] As the word *habit* implies, if we are serious about having students display thoughtful control over ideas, a single performance is inadequate. We need to observe students' *repertoires*, not rote catechisms coughed up in response to pat questions.

The problem is more serious than it first appears. The difficulty of learning lies in the breaking of natural but dysfunctional habits. The often-strange quality of new knowledge can cause us to unwittingly misunderstand new ideas by assimilating them into our old conceptions; this is particularly true when instruction is only verbal. That is why so many students who do well on school tests seem so thoughtless and incompetent in solving real-world problems. For example, the research done at Johns Hopkins University demonstrates how precarious and illusory "knowledge" of physics really is, when even well-trained students habitually invoke erroneous but plausible ideas about force on certain problems.[11]

The true test is so central to instruction that it is known from the start and repeatedly taken *because* it is both central and complex – equivalent to the game to be played or the musical piece to be performed. The true test of ability is to perform consistently well tasks whose criteria for success are known and valued. By contrast, questions on standardized tests are usually kept "secure," hidden from students and teachers, and they thus contradict the most basic conditions required for learning.[12] (Of course, statistical validity and reliability *depend* on the test being secret, and, when a test is kept secret, the questions can be used again.)

Designing authentic tests should involve knowledge use that is forward-looking. We need to view tests as "assessments of enablement," to borrow Robert Glaser's term. Rather than merely judging whether students have learned what was taught, we should "assess knowledge in terms of its constructive use for further learning. . . . [We should assess reading ability] in a way that takes into account that the purpose of learning to read is to enable [students] to learn from reading."[13] All tests should involve students in the actual challenges, standards, and habits needed for success in the academic disciplines or in the workplace: conducting original research, analyzing the research of others in the service of one's research, arguing critically, and synthesizing divergent viewpoints. Within reasonable and reachable limits, a real test replicates the authentic intellectual challenges facing a person in the field. (Such tests are usually also the most engaging.)

Most so-called "criterion-referenced" tests are inadequate because the problems are contrived, and the cues are artificial.

The practical problems of test design can best be overcome by thinking of academic tests as the intellectual equivalent of public "performances." To enable a student is to help him or her make progress in handling complex tasks. The novice athlete and the novice actor face the same challenges as the seasoned professional. But school tests make the complex simple by dividing it into isolated and simplistic chores – as if the student need not practice the true test of performance, the test of putting all the elements together. This apparently logical approach of breaking tasks down into their components leads to tests that assess only artificially isolated "outcomes" and provide no hope of stimulating genuine intellectual progress. As a result, teaching to such tests becomes mechanical, static, and disengaging. Coaches of musicians, actors, debaters, and athletes know better. They know that what one learns in drill is never adequate to produce mastery.

That is why most so-called "criterion-referenced" tests are inadequate: the problems are contrived, and the cues are artificial. Such tests remove what is central to intellectual competence: the use of judgment to recognize and pose complex problems as a prelude to using one's discrete knowledge to solve them. Authentic challenges – be they essays, original research, or artistic performances – are inherently ambiguous and open-ended. As Frederiksen has said:

> Most of the important problems one faces are ill-structured, as are all the really important social, political, and scientific problems. . . . But ill-structured problems are not found in standardized achievement tests. . . . Efficient tests tend to drive out less efficient tests, leaving many important abilities untested and untaught. . . . All this reveals a problem when we consider the influence of an accountability system in education. . . . We need a much broader conception of what a test is.[14]

Put simply, what the student needs is a test with more sophisticated criteria for judging performance. In a truly authentic and criterion-referenced education, far more time would be spent teaching and testing the student's ability to understand and internalize the criteria of genuine competence. What is so harmful about current teaching and testing is that they frequently reinforce – unwittingly – the lesson that mere right answers, put forth by going through the motions, are adequate signs of ability. Again, this is a mistake rarely made by coaches, who know that their hardest and most important job is to raise the standards and expectations of their students.

EXAMPLES OF AUTHENTIC TESTS

Let us examine some tests and criteria devised by teachers working to honor the ideas I've been discussing under the heading of "exhibition of mastery" – one of the nine "Common Principles" around which members of the Coalition of Essential Schools have organized their reform efforts.[15] Here are two examples of final exams that seem to replicate more accurately the challenges facing experts in the field.

An oral history project for ninth-grad-

ers.[16] You must complete an oral history based on interviews and written sources and present your findings orally in class. The choice of subject matter will be up to you. Some examples of possible topics include: your family, running a small business, substance abuse, a labor union, teenage parents, or recent immigrants. You are to create three workable hypotheses based on your preliminary investigations and come up with four questions you will ask to test each hypothesis.

To meet the criteria for evaluating the oral history project described above, you must:

- investigate three hypotheses;
- describe at least one change over time;
- demonstrate that you have done background research;
- interview four appropriate people as sources;
- prepare at least four questions related to each hypothesis;
- ask questions that are not leading or biased;
- ask follow-up questions when appropriate;
- note important differences between fact and opinion in answers that you receive;
- use evidence to support your choice of the best hypothesis; and
- organize your writing and your class presentation.

A course-ending simulation/exam in economics.[17] You are the chief executive officer of an established firm. Your firm has always captured a major share of the market, because of good use of technology, understanding of the natural laws of constraint, understanding of market systems, and the maintenance of a high standard for your product. However, in recent months your product has become part of a new trend in public tastes. Several new firms have entered the market and have captured part of your sales. Your product's proportional share of total aggregate demand is continuing to fall. When demand returns to normal, you will be controlling less of the market than before.

Your board of directors has given you less than a month to prepare a report that solves the problem in the short run and in the long run. In preparing the report, you should: 1) define the problem, 2) prepare data to illustrate the current situation, 3) prepare data to illustrate conditions one year in the future, 4) recommend action for today, 5) recommend action over the next year, and 6) discuss where your company will be in the market six months from today and one year from today.

The tasks that must be completed in the course of this project include:

- deriving formulas for supply, demand, elasticity, and equilibrium;
- preparing schedules for supply, demand, costs, and revenues;
- graphing all work;
- preparing a written evaluation of the current and future situation for the market in general and for your company in particular;
- preparing a written recommendation for your board of directors;
- showing aggregate demand today and predicting what it will be one year hence; and
- showing the demand for your firm's product today and predicting what it will be one year hence.

Connecticut has implemented a range of performance-based assessments in science, foreign languages, drafting, and small-engine repair, using experts in the field to help develop apt performance criteria and test protocols. Here is an excerpt from the Connecticut manual describing the performance criteria for foreign languages; these criteria have been derived from the guidelines of the American Council on the Teaching of Foreign Languages (ACTFL).[18] On the written test, students are asked to draft a letter to a pen pal. The four levels used for scoring are novice, intermediate, intermediate high, and advanced; they are differentiated as follows:

- *Novice.* Students use high-frequency words, memorized phrases, and formulaic sentences on familiar topics. Students show little or no creativity with the language beyond the memorized patterns.
- *Intermediate.* Students recombine the learned vocabulary and structures into simple sentences. Sentences are choppy, with frequent errors in grammar, vocabulary, and spelling. Sentences will be very simple at the low end of the intermediate range and will often read very much like a direct translation of English.
- *Intermediate high.* Students can write creative sentences, sometimes fairly complex ones, but not consistently. Structural forms reflecting time, tense, or aspect are attempted, but the result is not always successful. Student show an emerging ability to describe and narrate in paragraphs, but papers often read like academic exercises.
- *Advanced.* Students are able to join sentences in simple discourse and have sufficient writing vocabulary to express themselves simply, although the language may not be idiomatic. Students show good control of the most frequently used syntactic structures and a sense that they are comfortable with the target language and can go beyond the academic task.

Of course, using such an approach is time-consuming, but it is not impractical or inapplicable to all subject areas on a large scale. The MAP (Monitoring Achievement in Pittsburgh) testing program offers tests of critical thinking and writing that rely on essay questions and are specifically designed to provide diagnostic information to teachers and

"Ms. Kelsor says I'd do better with a team of teachers. She thinks six or eight would be about right."

students. Pittsburgh is also working, through its Syllabus-Driven Exam Program, to devise exemplary test items that are based more closely on the curriculum.[19]

On the state level, Vermont has recently announced that it will move toward a portfolio-based assessment in writing and mathematics, drawing on the work of the various affiliates of the National Writing Project and of the Assessment of Performance Unit (APU) in Great Britain. California has piloted performance-based tests in science and other subjects to go with its statewide essay-writing test.

RESPONSIVENESS AND EQUITY

Daniel Resnick and Lauren Resnick have proposed a different way of making many of these points. They have argued that American students are the "most tested" but the "least examined" youngsters in the world.[20] As their epigram suggests, we rarely honor the original meaning of the word *test*. Originally a *testum* was a porous cup for determining the purity of metal; later it came to stand for any procedures for determining the worth of a person's effort. To prove the value or ascertain the nature of a student's understanding implies that appearances can deceive. A correct answer can disguise thoughtless recall. A student might quickly correct an error or a slip that obscures thoughtful understanding; indeed, when a student's reasoning is heard, an error might not actually be an error at all.

The root of the word *assessment* reminds us that an assessor should "sit with" a learner in some sense to be sure that the student's answer *really* means what it seems to mean. Does a correct answer mask thoughtless recall? Does a wrong answer obscure thoughtful understanding? We can know for sure by asking further questions, by seeking explanation or substantiation, by requesting a self-assessment, or by soliciting the student's response to the assessment.

The problem can be cast in broader moral terms: the standardized test is disrespectful by design. Mass testing as we know it treats students as objects – as if their education and thought processes were similar and as if the reasons for their answers were irrelevant. Test-takers are not, therefore, treated as *human* subjects whose feedback is essential to the accuracy of the assessment. Pilot standardized tests catch many technical defects in test questions. However, responses to higher-order questions are inherently unpredictable.

TABLE 1.
An Item from the NAEP Science Test

Child's Name	Frisbee Toss (yds.)	Weight Lift (lbs.)	50-Yard Dash (secs.)
Joe	40	205	9.5
Jose	30	170	8.0
Kim	45	130	9.0
Sarah	28	120	7.6
Zabi	48	140	8.3

The standardized test is thus inherently inequitable. I am using the word *equity* in its original, philosophical meaning, as it is incorporated into the British and American legal systems. The concept is commonsensical but profound: blank laws and policies (or standardized tests) are inherently unable to encompass the inevitable idiosyncratic cases for which we ought always to make exceptions to the rule. Aristotle put it best: "The equitable is a correction of the law where it is defective owing to its universality."[21]

In the context of testing, equity requires us to insure that human judgment is not overrun or made obsolete by an efficient, mechanical scoring system. Externally designed and externally mandated tests are dangerously immune to the possibility that a student might legitimately need to have a question rephrased or might deserve the opportunity to defend an unexpected or "incorrect" answer, even when the test questions are well-structured and the answers are multiple choice. How many times do teachers, parents, or employers have to alter an evaluation after having an answer or action explained? Sometimes, students need only a hint or a slight rephrasing to recall and use what they know. We rely on human judges in law and in athletics because complex judgments cannot be reduced to rules if they are to be truly equitable. To gauge understanding, we must explore a student's answer; there must be some possibility of dialogue between the assessor and the assessed to insure that the student is fully examined.

This concern for equity and dialogue is not idle, romantic, or esoteric. Consider the following example from the National Assessment of Educational Progress (NAEP) science test, Learning by Doing, which was piloted a few years ago.[22] On one of the tasks, students were given three sets of statistics that supposedly derived from a mini-Olympics that some children had staged (see Table 1). The introductory text noted that the children "decided to make each event of the same importance." No other information that bears on the question was provided. The test presented the students with the results of three events from the competition.

The first question asked, Who would be the all-around winner? The scoring manual gives these instructions:

> Score 4 points for accurate ranking of the children's performance on each event and citing Zabi as the overall winner. Score 3 points for using a ranking approach . . . but misinterpreting performance on the dash event . . . and therefore, citing the wrong winner. Score 2 points for a response which cites an overall winner or a tie with an explanation that demonstrates some recognition that a quantitative means of comparison is needed. Score 1 point if the student makes a selection of an overall winner with an irrelevant or non-quantitative account or without providing an explanation. Score 0 for no response.

Makes sense, right? But now ask yourself how, using the given criteria, you would score the following response given by a third-grader:

> A. Who would be the all-around winner?
> No one.
> B. Explain how you decided who would be the all-around winner. Be sure to show your work.
> No one is the all-around winner.

The NAEP scorer gave the answer a score of 1. Given the criteria, we can see why. The student failed to give an explanation or any numerical calculations to support the answer.

But could that answer somehow be apt in the mind of the student? Could it be that the 9-year-old deliberately

and correctly answered "no one," since "all-around" could mean "winner of all events"? If looked at in this way, couldn't it be that the child was *more* thoughtful than most by deliberately *not* taking the bait of part B (which presumably would have caused the child to pause and consider his or her answer). The full sentence answer in part B – remember, this is a 9-year-old – is revealing to me. It is more emphatic than the answer to part A, as if to say, "Your question suggests I *should* have found one all-around winner, but I won't be fooled. I stick to my answer that no one was the all-around winner." (Note, by the way, that in the scorer's manual the word *all-around* has been changed to *overall*.) The student did not, of course, explain the answer, but it is conceivable that the instruction was confusing, given that there was no "work" needed to determine that "no one" was the all-around winner. One quick follow-up question could have settled the matter.

A moral question with intellectual ramifications is at issue here: Who is responsible for insuring that an answer has been fully explored or understood, the tester or the student? One reason to safeguard the teacher's role as primary assessor is that the most accurate and equitable evaluation depends on relationships that have developed over time between examiner and student. The teacher is the only one who knows what the student can or cannot do consistently, and the teacher can always follow up on confusing, glib, or ambiguous answers.

In this country we have been so enamored of efficient testing that we have overlooked feasible in-class alternatives to such impersonal testing, which are already in use around the world. The German *abitur* (containing essay and oral questions) is designed and scored by classroom teachers, who submit two possible tests to a state board for approval. The APU in Great Britain has for more than a decade developed tests that are designed for classroom use and that involve interaction between assessor and student.

What is so striking about many of the APU test protocols is that the assessor is meant to probe, prompt, and even teach, if necessary, to be sure of the student's actual ability and to enable the learner to learn from the assessment. In many of these tests the first answer (or lack of one) is not deemed a sufficient insight into the student's knowledge.[23] Consider, for example, the following sections from the assessor's manual for a mathematics test for British 15-year-olds covering the ideas of perimeter, area, and circumference.

1. Ask: "What is the perimeter of a rectangle?" [Write student answer.]
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. . . .*
10. "Estimate the length of the circumference of this circle."
11. Ask: "What would you do to check your estimate?" [String is on the table.] *If no response, prompt for string.*
13. Ask: "Is there any other method?" *If student does not suggest using* $C = \pi d$, *prompt with*, "Would it help to measure the diameter of the circle?"

Who is responsible for insuring that an answer has been fully explored or understood, the tester or the student?

The scoring system works as follows: 1) unaided success; 2) success following one prompt from the tester; 3) success following a series of prompts; 4) teaching by the tester, prompts unsuccessful; 5) an unsuccessful response, and tester did not prompt or teach; 6) an unsuccessful response despite prompting and teaching; 7) question not given; and 8) unaided success where student corrected an unsuccessful attempt without help. The "successful" responses were combined into two larger categories called "unaided success" and "aided success," with percentages given for each.[24]

The Australians for years have used similar tasks and similarly trained teachers to conduct district- and statewide assessments in academic subject areas (much as we do in this country with the Advanced Placement exams). Teachers give tests made up of questions drawn from banks of agreed-upon items and then mark them. Reliability is achieved through a process called "moderation," in which teachers of the same subjects gather to compare results and to set criteria for grading.

To insure that professionalization is aided, not undermined, by national testing, the process of "group moderation" has been made a central feature of the proposed new national assessment system in Great Britain. The tests will be both teacher-given and standardized. But what is so admirable – and equitable – is that

We must overcome the lazy habit of grading and scoring "on the curve" as a cheap way of setting and upholding standards.

the process of group moderation requires collective judgments about any discrepancies between grade patterns in different schools and between results in a given school and on the nationally standardized criterion-referenced test. Significantly, the process of moderation can, on occasion, override the results of the nationally standardized test:

> A first task of a moderation group would be to examine how well the patterns of the two matched for each group of pupils [comparing percentages of students assigned to each level]. . . . The meeting could then go on to explore discrepancies in the pattern of particular schools or groups, using samples of pupils' work *and knowledge of the circumstances of schools*. The group moderation would first explore any general lack of matching between the overall teacher rating distribution and the overall distribution of results on the national tests. The general aim would be to adjust the overall teacher rating results to match the overall results of the national tests; *if the group were to have clear and agreed reasons for not doing this, these should be reported . . . [and] departures could be approved if the group as a whole could be convinced that they were justified in particular cases.*[25] (Emphasis added)

At the school-site level in the U.S., we might consider the need for an oversight process akin to group moderation to insure that students are not subject to eccentric testing and grading – a committee on testing standards, for example. In short, what group moderation can provide is the kind of on-going professional development that teachers need and desire. Both equity in testing and reform of schooling ultimately depend on a more open and consensual process of establishing and upholding schoolwide standards.

A number of reasons are often cited for retaining "objective" tests (the design of which is usually quite "subjective"), among them: the unreliability of teacher-created tests and the subjectivity of human judgment. However, reliability is only a problem when judges operate in private and without shared criteria. In fact, multiple judges, when properly trained to assess actual student performance using agreed-upon criteria, display a high degree of inter-rater reliability. In the Connecticut foreign language test described above, on the thousands of student tests given, two judges using a four-point scoring system agreed on a student's score 85% of the time.[26] Criticisms of Advanced Placement exams that contain essay questions usually focus on the cost of scoring, not on problems of inter-rater reliability. Inadequate testing technology is a red herring. The real problem standing in the way of developing more authentic assessment with collaborative standard-setting is the lack of will to invest the necessary time and money.

True criterion-referenced tests and diploma requirements, though difficult to frame in performance standards, are essential for establishing an effective and just education system. We must overcome the lazy habit of grading and scoring "on the curve" as a cheap way of setting and upholding standards. Such a practice is unrelated to any agreed-upon intellectual standards and can reveal only where students stand in relation to one another. It tells us nothing about where they ought to be. Moreover, students are left with only a letter or number – with nothing to learn from.

Consider, too, that the bell-shaped curve is an *intended* result in designing a means of scoring a test, not some coincidental statistical result of a mass testing. Norm-referenced tests, be they locally or nationally normed, operate under the assumption that teachers have no effect – or only a random effect – on students.

> There is nothing sacred about the normal curve. It is the distribution most appropriate to chance and random activity. Education is a purposeful activity, and we seek to have the students learn what we have to teach. . . . [W]e may even insist that our efforts are unsuccessful to the extent that the distribution of achievement approximates the normal distribution.[27]

In addition, such scoring insures that, *by design*, at least half of the student population is always made to feel inept and discouraged about their work, while the other half often has a feeling of achievement that is illusory.

Grading on a curve in the classroom is even less justifiable. There is no statistical validity to the practice, and it allows teachers to continually bypass the harder but more fruitful work of setting and teaching performance criteria from which better learning would follow.

To let students show off what they

"I'll have the decaffeinated; 21 third-graders will be providing me with enough stimulation."

know and are able to do is a very different business from the fatalism induced by counting errors on contrived questions. Since standardized tests are designed to highlight differences, they often end up exaggerating them (e.g., by throwing out pilot questions that everyone answers correctly in order to gain a useful "spread" of scores).[28] And since the tasks are designed around hidden and often arbitrary questions, we should not be surprised if the test results end up too dependent on the native language ability or cultural background of the students, instead of on the fruit of their best efforts.

Tracking is the inevitable result of grading on a curve and thinking of standards only in terms of drawing exaggerated comparisons between students. Schools end up institutionalizing these differences, and, as the very word *track* implies, the standards for different tracks never converge. Students in the lower tracks are not taught and assessed in such a way that they become *better* enabled to close the gap between their current competence and ideal standards of performance.[29] Tracking simply enables students in the lower tracks to get higher grades.

In the performance areas, by contrast, high standards and the incentives for students are clear.[30] Musicians and athletes have expert performers constantly before them from which to learn. We set up different weight classes for wrestling competition, different rating classes for chess tournaments, and separate varsity and junior varsity athletic teams to nurture students' confidence as they slowly grow and develop their skills. We assume that progress toward higher levels is not only possible but is aided by such groupings.

The tangible sense of efficacy (aided by the desire to do well publicly and the power of positive peer pressure) that these extracurricular activities provide is a powerful incentive. Notice how often some students will try to sneak back into school after cutting class to submit themselves to the rigors of athletics, debate, or band practice – even when they are not the stars or when their team has an abysmal record.[31]

CRITERIA OF AUTHENTICITY

From the arguments and examples above, let me move to a consideration of a set of criteria by which we might distinguish authentic from inauthentic forms of testing.[32]

Structure and logistics. Authentic tests are more appropriately public, involving an actual audience, client, panel, and so on. The evaluation is typically based on judgment that involves multiple criteria (and sometimes multiple judges), and the judging is made reliable by agreed-upon standards and prior training.

Authentic tests do not rely on unrealistic and arbitrary time constraints, nor do they rely on secret questions or tasks. They tend to be like portfolios or a full season's schedule of games, and they emphasize student progress toward mastery.

Authentic tests require some collaboration with others. Most professional challenges faced by adults involve the capacity to balance individual and group achievement. Authentic tests recur, and they are worth practicing, rehearsing, and retaking. We become better educated by taking the test over and over. Feedback to students is central, and so authentic tests are more intimately connected with the aims, structures, schedules, and policies of schooling.

Intellectual design features. Authentic tests are not needlessly intrusive, arbitrary, or contrived merely for the sake of shaking out a single score or grade. Instead, they are "enabling" – constructed to point the student toward more sophisticated and effective ways to use knowledge. The characteristics of competent performance by which we might sort nonenabling from enabling tests might include: "The coherence of [the student's] knowledge, principled [as opposed to merely algorithmic] problem solving, usable knowledge, attention-free and efficient performance, and self-regulatory skills."[33]

Authentic tests are contextualized, complex intellectual challenges, not fragmented and static bits or tasks. They culminate in the student's own research or product, for which "content" is to be mastered as a *means*, not as an end. Authentic tests assess student habits and repertoires; they are not simply restricted to recall and do not reflect lucky or unlucky one-shot responses. The portfolio is the appropriate model; the general task is to assess longitudinal control over the essentials.[34]

Authentic tests are representative challenges within a given discipline. They are designed to emphasize realistic (but fair) complexity; they stress *depth* more than breadth. In doing so, they must necessarily involve somewhat ambiguous, ill-structured tasks or problems, and so they make student judgment central in posing, clarifying, and tackling problems.

Standards of grading and scoring. Authentic tests measure essentials, not easily counted (but relatively unimportant) errors. Thus the criteria for scoring them must be equally complex, as in the cases of the primary-trait scoring of essays or the scoring of ACTFL tests of foreign languages. Nor can authentic tests be scored on a curve. They must be scored with reference to authentic standards of performance, which students must understand to be inherent to successful performance.

Authentic tests are contextualized, complex intellectual challenges, not fragmented and static bits or tasks.

Authentic tests use multifaceted scoring systems instead of a single aggregate grade. The many variables of complex performance are disaggregated in judging. Moreover, self-assessment becomes more central.[35]

Authentic tests exist in harmony with schoolwide aims; they embody standards to which everyone in the school can aspire. This implies the need for schoolwide policy-making bodies (other than academic departments) that cross disciplinary boundaries and safeguard the essential aims of the school. At Alverno College in Milwaukee, all faculty members are both members of disciplinary departments and of "competency groups" that span all departments.

Fairness and equity. Rather than rely on right/wrong answers, unfair "distractors," and other statistical artifices to widen the spread of scores, authentic tests ferret out and identify (perhaps hidden) strengths. The aim is to enable the students to show off what they can do. Au-

thentic tests strike a constantly examined balance between honoring achievement, progress, native language skill, and prior fortunate training. In doing so, they can better reflect our intellectual values.

Authentic tests minimize needless, unfair, and demoralizing comparisons and do away with fatalistic thinking about results. They also allow appropriate room to accommodate students' learning styles, aptitudes, and interests. There is room for the quiet "techie" and the show-off prima donna in plays; there is room for the slow, heavy lineman and for the small, fleet pass receiver in football. In professional work, too, there is room for choice and style in tasks, topics, and methodologies. Why must all students be tested in the same way and at the same time? Why should speed of recall be so well-rewarded and slow answering be so heavily penalized in conventional testing?[36]

Authentic tests can be – indeed, should be – attempted by all students, with the tests "scaffolded up," not "dumbed down" as necessary to compensate for poor skill, inexperience, or weak training. Those who use authentic tests should welcome student input and feedback. The model here is the oral exam for graduate students, insuring that the student is given ample opportunity to explain his or her work and respond to criticism as integral parts of the assessment.

In authentic testing, typical procedures of test design are reversed, and accountability serves student learning. A model task is first specified. Then a fair and incentive-building plan for scoring is devised. Only then would reliability be considered. (Far greater attention is paid throughout to the test's "face" and "ecological" validity.)

"Boy, someday my kids will be able to learn an awful lot from my mistakes."

Only a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness.

As I said at the outset, we need a new philosophy of assessment in this country that never loses sight of the student. To build such an assessment, we need to return to the roots of authentic assessment, the assessment of *performance of exemplary tasks*. We might start by adopting the manifesto in the introduction of the new national assessment report in Great Britain, a plan that places the interests of students and teachers first:

> Any system of assessment should satisfy general criteria. For the purpose of national assessment we give priority to the following four criteria:
> - the assessment results should give direct information about pupils' achievement in relation to objectives: they should be criterion-referenced;
> - the results should provide a basis for decisions about pupils' further learning needs: they should be formative;
> - the grades should be capable of comparison across classes and schools . . . so the assessments should be calibrated or moderated;
> - the ways in which criteria are set up and used should relate to expected routes of educational development, giving some continuity to a pupil's assessment at different ages: the assessments should relate to progression.[37]

The task is to define *reliable assessment* in a different way, committing or reallocating the time and money needed to obtain more authentic and equitable tests within schools. As the British proposals imply, the professionalization of teaching begins with the freedom and responsibility to set and uphold clear, appropriate standards – a feat that is impossible when tests are seen as onerous add-ons for "accountability" and are designed externally (and in secret) or administered internally in the last few days of a semester or year.

The redesign of testing is thus linked to the restructuring of schools. The restructuring must be built around intellectual standards, however, not just around issues involving governance, as has too often been the case so far. Authentic restructuring depends on continually asking a series of questions: What new methods, materials, and schedules are required to test and teach habits of mind? What structures, incentives, and policies will insure that a school's standards will be known, reflected in teaching and test design, coherent schoolwide, and high enough but still reachable by most students? Who will monitor for teachers' failure to comply? And what response to such failure is appropriate? How schools frame diploma requirements, how the schedule supports a school's aims, how job descriptions are written, how hiring is carried out, how syllabi and exams are designed, how the grading system reinforces standards, and how teachers police themselves are all inseparable from the reform of assessment.

Authentic tests must come to be seen as so essential that they justify disrupting the habits and spending practices of conventional schoolkeeping. Otherwise standards will simply be idealized, not made tangible. Nor is it "soft-hearted" to worry primarily about the interests of students and teachers: reform has little to do with pandering and everything to do with the requirements for effective learning and self-betterment. There are, of course, legitimate reasons for taking the intellectual pulse of students, schools, or school systems through standardized tests, particularly when the results are used as an "anchor" for school-based assessment (as the British propose). But testing through matrix sampling and other less intrusive methods can and should be more often used.

Only such a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness. As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching our intellectual standards, student performance, teaching, and our thinking and discussion

about assessment will remain flaccid and uninspired.

1. For an explanation of the state reports of above-average test scores, see Daniel Koretz, "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?," *American Educator*, Summer 1988, pp. 8-15, 46-52; and Edward Fiske, "Questioning an American Rite of Passage: How Valuable Is the SAT?," *New York Times*, 1 January 1989.

2. Norman Frederiksen, "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, vol. 39, 1984, p. 200.

3. Grant Wiggins, "Rational Numbers: Scoring and Grading That Helps Rather Than Hurts Learning," *American Educator*, Winter 1988, pp. 20, 25, 45, 48.

4. Raymond Callahan, *Education and the Cult of Efficiency* (Chicago: University of Chicago Press, 1962), pp. 80-84.

5. David Tyack, *The One Best System: A History of American Urban Education* (Cambridge, Mass.: Harvard University Press, 1974), pp. 140-46.

6. Callahan, pp. 100-101.

7. Richard J. Stiggins, "Revitalizing Classroom Assessment: The Highest Instructional Priority," *Phi Delta Kappan*, January 1988, pp. 363-68.

8. Peter Elbow points out that in all performance-based education, the teacher goes from being the student's adversary to being the student's ally. See Peter Elbow, *Embracing Contraries: Explorations in Teaching and Learning* (New York: Oxford University Press, 1986).

9. For more on content as knowledge in use and on the design of curricula and tests around "essential questions," see Grant Wiggins, "Creating a Thought-Provoking Curriculum," *American Educator*, Winter 1987, pp. 10-17.

10. Gilbert Ryle, *The Concept of Mind* (London: Hutchinson Press, 1949).

11. M. McCloskey, A. Carramaza, and B. Green, "Naive Beliefs in 'Sophisticated' Subjects: Misconceptions About Trajectories of Objects," *Cognition*, vol. 9, 1981, pp. 117-23.

12. See also Walter Haney, "Making Testing More Educational," *Educational Leadership*, October 1985, pp. 4-13.

13. Robert Glaser, "Cognitive and Environmental Perspectives on Assessing Achievement," in Eileen Freeman, ed., *Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference* (Princeton, N.J.: Educational Testing Service, 1988), pp. 40-42; and idem, "The Integration of Instruction and Testing," in Eileen Freeman, ed., *The Redesign of Testing for the 21st Century: Proceedings of the 1985 ETS Invitational Conference* (Princeton, N.J.: Educational Testing Service, 1986).

14. Frederiksen, p. 199.

15. For a complete account of the nine "Common Principles," see Theodore R. Sizer, *Horace's Compromise: The Dilemma of the American High School*, updated ed. (Boston: Houghton Mifflin, 1984), Afterword. For a summary of the idea of "exhibitions," see Grant Wiggins, "Teaching to the (Authentic) Test," *Educational Leadership*, April 1989.

16. I wish to thank Albin Moser of Hope High School in Providence, R.I., for this example. For an account of a performance-based history course, including the lessons used and pitfalls encountered, write to David Kobrin, Department of Education, Brown University, Providence, RI 02912.

17. I wish to thank Dick Esner of Brighton High School in Rochester, N.Y., for this example. Details on the ground rules, the information supplied for the simulation, the logistics, and the evaluation can be obtained by writing to Esner.

18. Manuals are available from the Office of Research and Evaluation, Connecticut Department of Education, P.O. Box 2219, Hartford, CT 06115. For further information on the ACTFL guidelines and their use, see *ACTFL Provisional Proficiency Guidelines* (Hastings-on-Hudson, N.Y.: American Council on the Teaching of Foreign Languages, 1982); and Theodore Higgs, ed., *Teaching for Proficiency, the Organizing Principle* (Lincolnwood, Ill.: National Textbook Co. and ACTFL, 1984).

19. See Paul LeMahieu and Richard Wallace, "Up Against the Wall: Psychometrics Meets Praxis," *Educational Measurement: Issues and Practice*, vol. 5, 1986, pp. 12-16; and Richard Wallace, "Redirecting a School District Based on the Measurement of Learning Through Examination," in Freeman, *The Redesign of Testing . . .*, pp. 59-68.

20. Daniel P. Resnick and Lauren B. Resnick, "Standards, Curriculum, and Performance: A Historical and Comparative Perspective," *Educational Researcher*, vol. 14, 1985, pp. 5-21.

21. Aristotle *Nicomachean Ethics* 1137b25-30.

22. *Learning by Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics* (Princeton, N.J.: Educational Testing Service, Report No. 17-HOS-80, 1987).

23. Similar work on a research scale is being done in the U.S. as part of what is called "diagnostic achievement assessment." See Richard Snow, "Progress in Measurement, Cognitive Science, and Technology That Can Change the Relation Between Instruction and Assessment," in Freeman, *Assessment in the Service of Learning . . .*, pp. 9-25; and J. S. Brown and R. R. Burton, "Diagnostic Models for Procedural Bugs in Basic Mathematical Skills," *Cognitive Science*, vol. 2, 1978, pp. 155-92.

24. *Mathematical Development, Secondary Survey Report No. 1* (London: Assessment of Performance Unit, Department of Education and Science, 1980), pp. 98-108.

25. Task Group on Assessment and Testing, *(TGAT) Report* (London: Department of Education and Science, 1988), Paragraphs 73-75.

26. Personal communication from Joan Baron, director of the Connecticut Assessment of Educational Progress.

27. Benjamin Bloom, George Madaus, and J. Thomas Hastings, *Evaluation to Improve Learning* (New York: McGraw-Hill, 1981), pp. 52-53.

28. Jeannie Oakes, *Keeping Track: How Schools Structure Inequality* (New Haven, Conn.: Yale University Press, 1985), pp. 10-13.

29. Ibid.

30. On the engaging quality of "exhibitions" of mastery, see Sizer, pp. 62-68.

31. For various group testing and grading strategies, see Robert Slavin, *Using Student Team Learning*, 3rd ed. (Baltimore: Johns Hopkins Team Learning Project Press, 1986).

32. Credit for some of these criteria are due to Arthur Powell, Theodore Sizer, Fred Newmann, and Doug Archbald and to the writings of Peter Elbow and Robert Glaser.

33. Glaser, "Cognitive and Environmental . . . ," pp. 38-40.

34. See the work of the ARTS Propel project, headed by Howard Gardner, in which the portfolio idea is described as it is used in pilot schools in Pittsburgh. ARTS Propel is described in "Learning from the Arts," *Harvard Education Letter*, September/October 1988, p. 3. Many members of the Coalition of Essential Schools use portfolios to assess students' readiness to graduate.

35. Alverno College has prepared material on the hows and whys of self-assessment. See Faculty of Alverno College, *Assessment at Alverno*, rev. ed. (Milwaukee: Alverno College, 1985).

36. For a lively discussion of the research results on the special ETS testing conditions for dyslectics, who are given unlimited time, see "Testing, Equality, and Handicapped People," *ETS Focus*, no. 21, 1988.

37. Task Group on Assessment and Testing, *(TGAT) Report* (London: Department of Education and Science, 1988), Paragraph 5. K

"You changed an F to a B and a C to an A. Nice work, son!"