DOCUMENT RESUME

ED 348 400 TM 018 885

AUTHOR Wiggins, Grant
TITLE Toward One System of Education: Assessing To Improve, Not Merely Audit. State Policy and Assessment in Higher Education. ESC Working Paper.
INSTITUTION Education Commission of the States, Denver, Colo.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE 91
NOTE 41p.
AVAILABLE FROM Distribution Center, Education Commission of the States, 707 17th Street, Suite 2700, Denver, CO 80202-3427 (Order No. PA-91-2; $6 plus postage and handling charges).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; *Accountability; *Accreditation (Institutions); *Educational Assessment; Educational Change; Educational Policy; Elementary Secondary Education; Evaluation Methods; Guidelines; Higher Education; *Policy Formation; Standardized Tests; *Systems Development; Testing Problems
IDENTIFIERS Alternatives to Standardized Testing; *Authentic Assessment; *Performance Based Evaluation

ABSTRACT

Assessments should improve performance by providing usable feedback, and should not merely audit it. Problems with educational accountability policies stem from a flawed view of student assessment. Intellectual excellence cannot be obtained via one-time mandated tests composed of proxies for real challenges. Common standards should be developed for use in evaluating local standards and measures, not common tests. A more performance-based accreditation process is proposed, with policies that induce schools and colleges to explicitly benchmark local work, chart progress over time, and give incentives for meeting high performance standards. Authentic educational tests simulate problems of knowledge use found in professions and after formal education. Assessments must teach students that tasks, criteria, and standards found in schools and colleges are appropriate for all rational inquiry and fruitful intellectual life. Assessments with flexible and context-sensitive opportunities reveal student expertise. Two vignettes for focusing policy reform and 10 guidelines for developing a consistent system of assessment are given. Outcome-based education and site-based decision making ensure that all local testing from kindergarten through graduate school (K-GS) involves the worthiest tasks and best exit-level challenges, and adapts to all grades. A seamless K-12-graduate school system includes: authentic tasks and standards linking different system stages that are known to all students and teachers at lower levels and recur throughout their work; and authentic standards and measures that are thoroughly explained, taught, and practiced with constant opportunity for revision and improvement so that schools and students are genuinely culpable for substandard performance. The appendixes provide guidance about assessment practice from various sources in terms of general principles and recommendations, specific suggestions, scoring scales for writing activities, literacy profiles (reading), and "work requirements" in literature study and chemistry. (35 references) (RLC)

State Policy and Assessment in Higher Education

ecs working papers

Published by the Education Commission of the States



# TOWARD ONE SYSTEM OF EDUCATION:

# ASSESSING TO IMPROVE, NOT MERELY AUDIT

# TOWARD ONE SYSTEM OF EDUCATION: ASSESSING TO IMPROVE, NOT MERELY AUDIT

by

Grant Wiggins
Center on Learning, Assessment
and School Structure (CLASS)

Education Commission of the States
707 17th Street, Suite 2700
Denver, Colorado 80202-3427

This paper is part of an Education Commission of the States series on Assessment in Higher Education, written in 1990-91. The project was funded by the Fund for the Improvement of Postsecondary Education (FIPSE).

Copies of this book are available for $6.00 from the ECS Distribution Center, 707 17th Street, Suite 2700, Denver, Colorado 80202-3427, 303-299-3692. Ask for No. PA-91-2.



The Education Commission of the States is a nonprofit, nationwide interstate compact formed in 1965. The primary purpose of the commission is to help governors, state legislators, state education officials and others develop policies to improve the quality of education at all levels.

Forty-nine states, the District of Columbia, American Samoa, Puerto Rico and the Virgin Islands are members. The ECS central offices are at 707 17th Street, Suite 2700, Denver, Colorado 80202-3427.

It is the policy of the Education Commission of the States to take affirmative action to prevent discrimination in its policies, programs and employment practices.

| Postage and handling charges: | |
|---|---|
| Up to $10 | $ 1.90 |
| $10.01-$25.00 | $ 3.00 |
| $25.01-$50.00 | $ 5.50 |
| $50.01-$100.00 | $ 8.00 |
| Over $100.01 | $10.50 |

## TABLE OF CONTENTS

# TOWARD ONE SYSTEM OF EDUCATION: ASSESSING TO IMPROVE, NOT MERELY AUDIT[1]

Is it really proper to say that we have an education "system"? I believe we do not have one — and not because we lack a national curriculum. Rather, the long-standing incoherence in education stems from four failures: tests that are not built out of exemplary tasks; our penchant for impatiently mandating uniformity (instead of requiring quality performance from *appropriately* varied syllabi and tests); the use of one-shot tests in comparing different student cohorts instead of assessing the same cohort's progress over time, in reference to authentic standards; and our myopia about why traditional tests cannot, in principle, assess the ultimate outcomes of a good education.

The heart of the argument is the view that any assessment should improve performance by providing usable feedback, not merely audit it. Trying to obtain intellectual excellence through one-time mandated tests composed of proxies for real challenges is a contradiction. It would be more effective policy to develop common standards for use in the evaluation of *local* standards and measures, not common tests. The state *can* mandate high standards (and provide better incentives for meeting them) without imposing standardized tests, as many countries long ago realized — better, more performance-focused *accreditation* — not more superficial testing providing only the illusion of accountability.

Schools and universities then would be appropriately free to develop their own assessments, while oversight agencies would reserve the right to audit tests and performance results for technical soundness, fairness and effectiveness. When comparisons of schools are necessary, we can develop agreed-upon indicators and calibration procedures to provide data without imposing high-stakes, superficial tests that corrupt the school's aims and autonomy.[2]

We would thus be devising policies that induce schools and colleges to do what all professions and the best companies do in the quest for improvement: explicitly "benchmark" local work, chart progress over time and provide incentives for meeting high (and higher) standards of performance. In the case of education, this means ensuring that tests are "authentic" so that they mirror or simulate the problems of knowledge-use found in the professions and at the *end* of formal education. Those tasks involve research leading to a dissertation and its defense, the "test" of apprenticeship or effective grappling with realistic case studies (as found in most law, business and medical schools). We would have a *system* when we ensure that younger students are confronted with such maximally enabling tasks — not to be confused with tests designed to certify control over common knowledge and orthodox ideas, tests that present more of a hurdle and an easy sorting device rather than being standard-revealing.

Since life's tasks, our students and our school customers *properly* differ, we need assessments that provide flexible and context-sensitive opportunities for revealing student expertise. But we must also demand that local work be done to high standards. Such a view strikes many

policy makers as paradoxical — how can there be standards without mandated standardization of tests?

Two Vignettes for Focusing Policy Reform

Consider for a moment the characteristics of an interesting assessment strategy found now within our educational world. There, the challenges are not at all standardized; indeed, they are by design fully personalized, allowing the student free rein as to topic and direction. Nor are students subject to uniform tasks required of all; no one thinks this odd or "invalid." Further, contrary to common practice, the test is never secret: the assessment, in fact, is centered on the students' creativity and thoughtfulness, knowledge crafted into products and performances of their own design. No student must show a uniform mastery of orthodox knowledge to "earn" this right to create — even though his or her school is mainstream and not at all "alternative."

In these institutions, the schedule suits the learner's pace and talents, so that the student is only assessed when ready. Here, the relationship between teacher and student is not at all adversarial. The teacher is the student's ally and guide through the challenges of the assessment, not the enemy to be "psyched out." Here, the assessors *do* "sit with" the student, literally and figuratively, probing the student's ideas. The assessor is, in fact, obligated to understand the student's point of view to validly test the student's grasp of ideas — a far cry from the typical "gotcha!" test.

Perhaps by now you have seen through this ironic picture. In a conference designed to broaden our typically narrow perspective on assessment, it is worth noting that kindergartners and graduate students in our best schools and universities have much in common. All the previous conditions of assessment apply to the two extreme points of our educational world. At both extremes we standardize the quality of performance expected, not the tasks. One might even say that at both ends we focus far more on the student's intellectual virtues than the correctness of their words. Nor would we dream of glibly comparing these students in merely a normative way. Each piece of work, be it a drawing or a dissertation, is examined for what it reveals about the learner's habits of mind and ability to create meaning, not one's "knowledge" of "facts." Even the details of test administration are parallel: resources are not only allowed in the "test," but we often want to determine whether the student wisely uses all available resources. Test "security" thus is a foolish and counter-productive strategy in each case.

Indeed, it is only at the beginning and end of formal education that we acknowledge the truth of how intellectual accomplishment is best judged: through an evocative examination of the student's use of knowledge and by "subjective" interaction of mind and mind — dialogue. The essential question is not: Is the student correct? But rather: Do the student's ideas, arguments and products *work* — i.e., do they effectively and gracefully achieve the student's intention? Is the student making progress in meeting apt standards of craftmanship and rigor?

Somehow we have gone wrong in the vast middle of schooling, violating our sense of what "education" and "assessment" really mean. What must we do to recapture the radically common-sensical view that any test of intellectual ability must be interactive? Why do we not see that typical tests bear no relation to the challenges facing the student in later worlds and thus lack the kind of validity that should matter most? And what must a policy maker do to rectify the situation when such dysfunctional habits run so deep?

Yes, "habits." Habits, not rational policy choices, brought us to this point. Unthinking reliance on past practice sustains large-scale testing as a solution to performance problems, not evidence or a lack of "research" or money for alternatives. Real reform begins when we see that more imposed testing as a response to our educational problems is like any other addictive and rationalized behavior. The second vignette to keep in mind in contemplating new policy, therefore, is the tale of the "Emperor's New Clothes."

In the story, rascals pose as tailors "weaving" a suit of the "finest" cloth for the king, earning riches by the fashioning of an illusion — an illusion not only about the garment itself but also about their skill in serving the king. The king's nakedness remains unseen because of the tailors' warning: only boorish folk would fail to recognize the quality of the "incredibly fine" yarn. And so it happens that the townspeople rationalize the nakedness and their secret doubt; they, like the king's retinue who fear for their honor, praise the king as he parades in his "finery." The king, too, knows that *he* is not a commoner and is sucked into the self-deception. It is the innocent child, unpossessed of a need to appear refined, who exposes the hoax. "But he has nothing on!" exclaims the child.

Oddly enough, few people recall the story's ending — and it is the ending that shows how harmful unthinking habits can be. The elders initially dismiss the remark of the young "innocent." Eventually the truth of the child's words cut through the illusion. But, while thinking that the now-skeptical townspeople must be right, "the Emperor thought to himself, 'I must not stop or it will spoil the procession.' So he marched on even more proudly than before, and the courtiers continued to carry a train that was not there at all."[3]

The tale is instructive about current testing policy on many levels. We *still* do not want to spoil the procession. Testing increases while few useful results emerge from the investment. We are *still* dismissing the remarks of the "innocents." We do not look through the eyes of students as they prepare for and take the tests we buy to see how debilitating they are to intellectual engagement, courage and imagination. Nor do we look through the eyes of employers, teachers and administrators to see how rarely they study test results to *understand* an applicant's abilities or the meaning of their errors. We are so self-deceived that one invariably hears in conversations on test reform: "But we made it through the system, didn't we?" — as if these high-stakes tests were nothing more than the "harmless" indignities of a freshman initiation of years gone by.

Multiple-choice test makers literally profit from the illusion that, like the tailors' yarn, all "fine" tests must be built with a specialist's mysterious skill. Testing, rather than the very

common practice of assessing student performance on the tasks we value, becomes an arcane science that is entrusted — and apparently is entrustable — only to statisticians.[4] Critics of such tests fear looking like the crude folks the tailors warn their critics will be; wary practitioners are routinely made to feel ignorant of the true "finery" in test validity and reliability.

The unreal simplicity of the typical test item is much like that of the king's nakedness: so obvious as to make one feel that some complex set of standards must surely render the test substantive. Like the townspeople in the story, we end up talking as if the real achievements we value and ought to be measuring were being *directly* observed in detail on multiple-choice tests. Even the supposed necessity of secure tests becomes "obvious" to everyone — as if all the important challenges and criteria in life (getting employed, writing a thesis for a doctorate, obtaining a driver's license, winning the Super Bowl or submitting a winning engineering or graphics design bid) were routinely kept secret from prospective performers.

The inevitable then happens. Rather than having policy incentives that would improve classroom assessment in a manner appropriate to instructional aims, teachers and professors are encouraged to imitate the psychometric "tailors." Even while talking of the foolishness or harm of such tests, educators usually end up employing their own *inadequate* versions of them — the true sign of the tests' mythic rather than rational power. The call for more mandated, valid and securely administered standardized tests then naturally increases as local assessment (thus, performance) deteriorates; the vicious circle continues.[5] Bring in the tailors! Let the king march more proudly!

Unlike in the story, then, our assessment "garments" still seem sublime rather than illusory. The child's voice — common sense — still remains largely unheard or dismissed in policy circles. Thus, cramming takes precedence over pride in one's work and joy in being genuinely tested; schools and colleges worry more about test results than whether scores represent thoughtful or thoughtless mastery. We have all lost sight of the fact that education succeeds when we provide students with the joy of thinking deeply and well about important things — the joy that is one of the chief incentives for staying in school and *really* mastering the "basics."

## Toward a Consistent System of Assessment

With the point of this paper to provoke sharpened discussion and to carefully revisit the "obvious" policy answers, let me offer some brief propositions intended to advance thinking beyond the hackneyed. These are postulates for further discussion; I do not propose to justify all these notions here, but they do have value as policy foundations.

1. *Tests ultimately teach more than they measure what was taught. They should therefore be composed of the tasks and qualities we most value.* What we test is what gets taught. The tests we design are the de facto aims of education; they must therefore embody our standards.

Can students do research, fashion compelling arguments, bring closure to discussions, plan and execute a project, infer conclusions from confusing data and write engaging prose or poetry? Let our assessment systems be primarily built out of such tasks, modified for age and experience as necessary.

This is the argument for "authentic assessment." But the postulate further suggests that it is the obligation of the assessor to evoke from the students the full extent of their knowledge — even, perhaps, not resting content with the first student answer given. We are obligated, I believe, to base important decisions on an accurate and complete profile of the student's intellectual accomplishments — as opposed to a single score based on the student's success or failure in answering a small set of imposed test items.

Put simply, we should be standardizing the (high) standards for judging (often idiosyncratic) intellectual performances, not standardizing items for use on inherently superficial tests.

The policy implication is clear: any "audit" of performance should be derived from this primary record of (local) work and achievement, not through the imposition of a common, indirect test. If indirect measures must be used, they should be carefully scrutinized for obtrusive or counter-productive influence on learning and teaching.

The collection of student work can be selected and edited as necessary to serve policy questions and needs. Why have we consistently failed to use the basic statistical tool of effective sampling of student performances and productions, just as we do in the artist's portfolio or the doctor's residency? Why haven't policy makers asked educators to "agree to agree" locally on the kinds of tasks worth mastering and to report on the results over time? (California's matrix sampling of student writing, Vermont's sampling of student portfolios and Connecticut's sampling of high school seniors in performance-based tasks in math and science are notable recent examples that illustrate the point in the K-12 state policy world).

Operating under such a premise would get us beyond two utterly dysfunctional habits of formal education — the view that students must first learn and be tested upon "the basics," out of context, before earning the right(?) to tackle our most valued tasks; and the view that large-scale tests must be composed of proxies for authentic work on economic and psychometric grounds. But at what cost in dropouts and to the integrity of our schools and universities? At what cost to the capacity of schools and colleges to produce scholars — i.e., students who can do their own research well? And where has the debate been over *what kind of margin of error is tolerable* in assessment? Other countries have historically been content to use human judges in scoring, even in high-stakes examinations. Are we using a micrometer when a ruler would do? Is the demand for such reliability causing our intellectual values to be co-opted in the name of "precision" and "objectivity"?

2. *The measure of a successful assessment system is the degree to which it improves student and school performance while validly and reliably measuring performance.* Assessments should be assessed, in other words, not merely for their validity but also for their

effectiveness in improving performance. Credible and worthy measures and standards in the "real" world do more than monitor performance. They improve performance; standards and expectations are both raised, whether we are talking about Nintendo, engineering or musical productions.

Why are school standards and expectations falling? In part because of our tests. Large-scale tests often have to be "low-ceiling" to fully capture and discriminate the full range of normative, not exemplary, performance. If school and college faculties are increasingly asked to "teach to the (low-level) test," they become like the myopic patient in fear of pain who at the yearly physical thinks only of psyching out the doctor on the few tests given in the office visit. Like the patient, our faculties come to confuse cause and effect because of our policies and pressures; they, too, fixate on short-cuts to get the indicators up — instead of making a long-term commitment to the daily regimens that produce (intellectual) healthfulness and better long-term results.

A solution is for states to formulate and support assessment policies that make *systemic effectiveness* part of the essential parameters of assessment design and use.[6] A good assessment system assists in improving local performance because it is in complete harmony with the intended outcomes. Put a different way, we need assessment policies at the local and state levels to ensure that the design, purchase and use of assessment strategies support the aims of the organization.

At the very least, this view suggests a policy like those found in many European countries where local educators are free to design their own assessments, but the instruments have to meet standards set by regional boards. Similarly in scoring: through the "moderation" process (frequently used in Great Britain and Australia and proposed for the national exams of the New Standards Project in this country), local scores are re-calibrated as necessary to ensure that common standards are upheld. Such a procedure is the only way to obtain both high standards and ownership of the assessment policy by local educators.

"Improvement" can only be measured if we assess the same cohort of students many times, in reference to stable standards. At the very least, we should be using pre-test/post-test measures for holding schools and colleges accountable. To compare one year's cohort to another's, given the rates of mobility and changes in demographics that effect schools and colleges, is to forfeit insight into the determinants of accountability.

3. *Standards are set by establishing benchmarks and models: quality performances at exemplary tasks.* We must assess both "input" as well as the quality of student "output" — i.e., the qualities of tasks assigned and products received. But as important as it is to choose more authentic tasks, we must begin to see that choosing *tasks* is not sufficient to set *standards*. The two senses of "standards" — worthiness of task undertaken versus adequacy of result — are independent, captured in the differentiation made in music, diving and gymnastics — the *difficulty* of a task and the *quality* of a performance. The current discussion of "standards" is thus hopelessly confused because it conflates the two meanings

and often collapses them into one statistical artifact. (See the recent NAEP math standards task force report for an example of this mistake at the national policy level; NAEP has selected test items to represent standards).

Standards would require us to design tasks that make the student perform or produce a product. Quality can only be observed when the student must use knowledge. Have we completely forgotten Bloom's taxonomy — never mind common sense? Bloom and his colleagues argued that the upper-level capacities could by and large only be assessed by the student being asked to fashion "unique products, performances or discourse." Simply because someone can point to the right answer on an upper-level math item on the NAEP test reveals nothing about the thoughtfulness and effectiveness whereby they habitually do their work.

What we need in both assessment and curriculum design is agreement on the important set of rich tasks to be mastered and samples of work that embody the standards set for those tasks. This would be the intellectual equivalent of designing the Decathlon, and fixing appropriate qualifying scores for each course and level of education grounded in some sound "benchmarking" or equating process. (This is the true spirit of portfolio assessment, whereby we collect samples of work from students on all the important genres and tasks over time to assess for habitual achievement; it is also the only way to ensure that testing is valid since we cannot assess control over all important tasks in one test sitting).

In Kentucky's new state assessment system (through the work of the Council on Performance Standards), there would be broad faculty consensus on and extensive use of an "exemplary-task bank" from which assessments might be constructed. The tasks would be performance and production challenges, developmentally modified as necessary, deemed worthy by both scholars and professionals.

Such a view means thinking of "tests" in the same sense in which rock-climbing tests the climber — a revealing measure of both training and the essential habits of mind. Such perspective is not the luxury of the elite. Empathy with odd, unfamiliar or alien views; ability to construct plausible cases in ambiguous situations; self-adjustment; sustained, effective and responsive analysis; capacity to argue effectively but tactfully, etc., are the hallmarks of any thoughtful person.[7] Until our assessment system both evokes and requires these virtues for success on the tests, we will be vitiating liberal education.

4. *Performance standards are not test norms, nor are they arbitrary cut scores. They are anchored by "benchmarks" — specific, apt, exemplary performances or products.* Quality is not an abstraction or a statistical artifact. Standards for rigor, precision, creativity, implied persistence, etc., are set by examples — samples of actual products/performances that exemplify the qualities we seek. Performance standards are empirically induced, in other words, by a combination of research, observation and wise judgment. We "set the standard" at the top of our scoring scale through the wise choice of "anchors" — samples of work that we believe to be genuinely excellent and apt models for emulation. Like the bullseye or prize-winning essay, a real standard supplies the essential element of effective self-assessment

— usable feedback — where I see for myself how my work compares with the standard-setting work. Thus, a real standard empowers the student: it enables one to effectively self-adjust performance.

The key questions, then, for policy become: Who chooses the benchmarks and by what process? What is a *justifiable* standard, one mindful of our highest expectations, matters of equity and sensitivity to the goals and contexts of schools and colleges? What are appropriate expectations on the developmental path toward those (stable) standards versus arbitrary grade- or "exit-level" expectations that are set in a vacuum?

What we now refer to as "standards" in testing are mere cut scores — useless for a meaningful examination and improvement of performance. If there is no qualitative difference between a 59 and a 61, what does it mean to say that 60 is "passing"? This is not a problem limited to local testing and grading. The Advanced Placement exams are scored with reference to historical norms of the distribution of past results. We really have little idea whether this year's percentage of Advanced Placement students who get 5s, 4s and 3s are writing essays as good as those of their equals of 20 years ago because the Head Reader refers to patterns of old scores (linked to the standard curve) when turning raw scores into final scores. This is a critical issue if we seek to increase the number (and, presumably, alter the general calibre) of the students taking and passing the exams in a way that doesn't lower standards.

To set stable and evocative standards:

5. *We should routinely assess student work from the viewpoint of stable exit-level criteria and standards.* This is the only conceivable wise way to chart genuine progress over time. The British and Australians have tried such a system (see the appendix for two samples) and it warrants our immediate attention. Only when we routinely score current work against exit-level standards can we be sure that local grades are sufficiently reliable and correlated with the ends of education; only then can we know what constitutes not merely "normal" growth but effective progress — i.e., the "slope" of improvement, in reference to authentic standards. (Think of how athletes and musicians — and video game players — fix their sights on exemplary performances to gauge and direct their own progress).

This is also the only way to get beyond the sham of age-grade equivalents that are really arbitrary timetables for charting norms. At present, our testing rewards native talent and speed in learning instead of providing a revealing and flexible record of (necessarily varying) student progress in meeting standards.

It follows that:

6. *An important problem in our educational assessment system is the absence of standards for faculty grading of student work.* There are now neither shared criteria nor stable performance standards for ensuring adequate inter-rater reliability and consistency across

faculty members, nor are there policies that require faculties to "agree to agree" on grading standards within a "tolerable" margin of error. The British and others have addressed this problem through the "moderation" process, and we need to emulate it (as Vermont plans to do in its portfolio process).

While faculties at all levels, and especially at the college level, may balk at such a requirement on grounds of professional autonomy, such an argument is specious: no teachers have the right to design invalid and/or unreliable assessments, nor do they have the right to design tests that do not finely articulate with the stated mission, goals and outcomes of the institution.

7. *Quality work is incompatible with one-event, predominantly external, testing.* If our aim is not merely to hope for uniformly excellent work from all students but evoke and receive it on a consistent basis, we must provide the student with multiple opportunities to be re-tested on the same essential tasks — just as we do in all performance-based teaching and learning. Or is our aim to set standards through invidious comparisons only, whereby we fail large numbers of students in the name of high standards, i.e. steep "curves"? (Note that this dysfunctional view of assessment is not found at the kindergarten or Ph.D levels).

Because so much of both local grading and large-scale testing is typically norm-referenced, we have trouble imagining "high standards" being met by all students — without a corruption of standards. That a wide diversity of performance should be *expected* after an education is a view that is not found, thank God, in flight school or in medical school where uniform success is the only standard. This is similar in vocational programs. A vocational teacher told me that he required all his students to get 100% on the tests for use of a radial saw before allowing unsupervised use. Think of a typical 60% being an acceptable performance.

If we had a genuinely standard-referenced system, we could and would happily have all students earn A's (as long as we audit to be sure that the grades genuinely reflect what the judges claim they mean). This is what I think the job of state departments of education and accreditation agencies should be: *setting standards for local standard-setting* and conducting audits to ensure that local criteria and standards are honored in local assessment within "tolerable" margins of error across faculty members. This is not implausible — the New York State Regents Exams have operated this way for 100 years.

Demanding and getting quality work is a local, daily affair. Excellence is only obtained by successive approximations toward (the known) "standard" performances. This means that a demand for excellence and one-shot tests, with no opportunity for effective feedback and revision, are inconsistent. Quality control is the avoiding of sub-standard performance — impossible if the only assessing that counts occurs once, late in the year and relies on a relatively unpredictable sample of some de-contextualized body of knowledge, through the device of a "secure" test.

In fact:

8. *The demand for glib comparability of all schools and colleges, using a few scores on indirect tests, is not accountability at all.* Of all the issues discussed at the conference for which this paper was written, nothing generated more immediate objections and gasps of surprise than the question: "Who really wants such comparability?"

Comparisons are inevitable in judging the quality of any performance. They also aid the student in improving — but only if the comparisons offer useful feedback about how to improve. I know of no walk of life except education where the sole measure of one's achievement is a one-shot superficial test, yielding one aggregate number that *cannot* assist the learner in improving. Whether we consider apprentice athletes, doctors, pilots or soldiers, we find that an array of statistics and usable feedback are presented to profile the capacities and accomplishments of the student.

Consider a simple example to show how blinded we are on this matter. It is mathematically indefensible to aggregate a baseball player's many statistics into one aggregate score; "hits" and "runs batted in" and the others are incommensurable. So, too, in the intellectual world — "intellectual initiative," "persistence," "powers of analysis" and "consistency and precision of results" are different, valued capacities. Why, then, is it tolerable in testing and reporting to reduce a year's worth of complex work to one score on a proxy test? Why don't we see that a score for de-contextualized "knowledge" and "skill" is fundamentally misleading if we don't assess the judgment and care in the employment of knowledge and skill or value the task that the test item demanded? A report of aggregate student scores, used to compare schools glibly, tells us little when curricula are diverse and differently organized from school to school.

Comparisons in education have almost always been invidious and of little value for stimulating improvement. The reasons are obvious — we test what is easy to test and not what is essential; large-scale tests tend to be immune to local curricula, hence local gains; rankings rarely deviate much from year-to-year results. The scores usually say more about lucky gene pools than value-added achievement by the school or university.[8] Worse, norm-referenced tests are designed to exaggerate the differences between students. In test pilots, questions are thrown out if everyone gets them right or wrong; the aim is to maximally discriminate. Such a mechanism ensures that an institution or student is extremely unlikely to change position in the ranking. What does that do to incentive, never mind fair accountability? The built-in stability of conventional test scores and their "curves" tend thus to yield self-fulfilling prophecies about test results, hardening our prejudices and fatalism about change.

The solution to this mess was proposed 80 years ago: a test of performance in which the performers compete against known standards, not other performers only. Alas, so-called "criterion-referenced" tests have been a sham up to this point because of their proxy and secure nature, *and because the "criteria" were applied to item selection and not student*

*performances*. Let us enter the modern world of performance assessment by devising tests in which the tasks, scoring criteria and standards are benchmarked and enticing. But that means challenging the most basic aspect of traditional testing — test "security."

9. *The aim of getting quality work from all students is utterly incompatible with "secure" tests and mysterious criteria and standards*. Quality work in all domains always depends upon accurate self-assessment and self-adjustment in reference to specific, known standards. How can we expect students or schools to improve if the assessments rely on a fairly unpredictable sample of unknown, generic test questions? Only deep-seated and unthinking habits blind us to an obvious fact: pervasive secrecy in assessment is counter-productive — *and immoral*. A few fanciful vignettes may be useful in jolting us to see the harm:

A. Consider what our response as adults would be to a job evaluation process in which the employer could do what test-makers routinely do: pick a few tasks from the hundreds we had learned and performed over the years, without our knowledge or consent, and assess our performance on a one-hour "secure" paper-and-pencil test. (Worse, imagine your employer relying on a testing company to assess your performance through the use of generic multiple-choice tests.) It is telling that, for adults, the practice would be regarded as unfair, inappropriate and likely illegal. Why does it not seem so when dealing with students?

B. Imagine if student musicians had to wait until test day to know the music they would be playing in concert. Assume, too, that students play their instruments through microphones connected to other rooms where judges could listen but students could not hear themselves play. Weeks later the student would receive a single score telling them where they stood relative to all the clarinet or trumpet players in the state, and a computer print-out summarizing the stylistic and technical areas they should work on.

C. What if baseball were played all season long, but the pennant races were decided using one-shot tests with one aggregate score designed by statisticians? Thus, on the last day of the season, specially constructed, secure tests would be given to each player, composed of a sample of drills and game situations. The pennant races would be decided by each year's (new) test and its results.

Performance and performance improvement is impossible when the information received is so limited — a limit imposed by prior secrecy and the vagaries of test-maker sampling. But more than the first two vignettes, this last one reveals how unwittingly *obtrusive* the designers of one-shot, high-stakes secret tests can be — even if their aim is to be merely helpful statisticians. The test designer seeks to design a valid assessment of all the important sub-skills as specified by others; secrecy is required to enable simple parts of a complex game to be efficiently used to draw valid inferences. Yet, we easily see how such a system would corrupt coaching and the game itself. Not only the student-players, but the teacher-coaches would be robbed of the capacity to concentrate on excellent play in such an assessment system.[9]

10. *To meet the most demanding exit-level standards requires that we practice meeting them throughout our career — whether we are thinking of a year's worth of schoolwork or a K-graduate school career.* We must recapture what every coach knows: drill at the simple and constituent parts of a complex task is necessary but not sufficient. We must continually practice the ultimate performance as well as the constituent parts to achieve mastery. All the more so if the ultimate performance is the creative, rigorous and personalized work of developing one's own ideas into knowledge.

## Assessment and Opportunity

This last proposition goes to the heart of any common-sense view of a "seamless" education system. The point of organizing schooling around exit-level intended outcomes couldn't be simpler: design curricula backwards around the achievements you wish all students to master. It follows that our assessments must do more than reveal whether the learner has mastered what one isolated teacher happened to just finish teaching. The examining of students at *each* stage must embody and point toward the tasks, criteria and standards that represent the future goals. Assessment must always be constructed out of "enabling" tasks.

The ultimate intellectual challenge of formal schooling is the dissertation and its defense (or, more generally, the rigorous development and full explication of one's own ideas and arguments). We should therefore anchor our K-graduate school system by this "standard-setting" task — not because all or even most students will earn Ph.Ds, but because ensuring maximal mastery of the ultimate task requires that students continually practice and be tested on it (even if in simplified form) from the beginning of learning. No novelty here: look at Little League, ballet or chess.

This is not as far-fetched as it might seem. In the Advanced Placement Art Portfolio Exam, students submit evidence of their choosing that reveals the breadth of control over important genres and studies and an in-depth project focused on one problem, theme or style. They also supply written explanations of their intentions with the pieces chosen for inclusion. In effect, they are judged on the effectiveness of the realization of their intentions (as with the dissertation), not someone else's view of what subject matter they should "know." The instructions to students and judges makes this clear: the assessors look for pieces that "work in their own way." Similarly, I have seen high school English teachers anchor their portfolio assessment in the student's choice of "major pieces" and their self-assessment — a contract with the self, as it were, to produce quality.

We can put this talk of "seamlessness" and building "ownership" of the assessment process in a very different way — in the language of equity. If we are ever to qualify students for the upper-level task of quality production (and thus maximally qualify them for a fruitful adulthood), we must require that all tests throughout the system reveal and point the student and teacher in those directions. To cast the point in the common-sense language of quality control: *"Quality is not what our tests say it is; quality is what the customer wants and*

*requires from us.*"[10] Educators are quick to dismiss this kind of remark with a wave of the hand and some derisive comments about businesspeople and glass houses, etc. But the point is well taken, particularly if we remember that each division of schooling has an internal customer — the next level(s) of schooling. How rare it is for middle school teachers to see our best high school papers and tests to know how to equip their students for success! Worse, how rare it is for isolated high school teachers who are prone to covering content to see how dysfunctional their own view of teaching, learning and assessing often is — if the point is to maximally qualify students for complicated, interesting, intellectual work.

The following example, a freshman European history exam at Harvard, illustrates "what the good-college-as-customer wants":

1. (30 minutes) The student must choose eight of 16 sets of items and explain why one of the items in each set does not belong.

   a. Waterloo, Trafalgar, Austerlitz

   b. Night of August 4, General Will, terror

   c. Montesquieu, Madison, Calhoun

2. (45 minutes) Choose one question.

   a. Imagine yourself Jean-Jacques Rousseau, but living in the early 20th century. Write a brief review and evaluation of Freud's work in light of your own theories of society.

   b. Imagine yourself Karl Marx, living half a century later. Write a brief evaluation of the programs of the Fabian socialists and the American reformers such as T. Roosevelt to present to the Socialist International.

   c. "Women's history can only be understood as part of general historical development." Do you agree? Why or why not?

3. (45 minutes) Choose one.

   a. "Both Germany and the U.S. fought decisive wars of unification in the 1860s, but in Germany, the landlords retained great power after the war, while in America, the landlord classes lost decisively." Discuss.

   b. Compare and contrast the causes of the two world wars.

   c. Would the European economies have developed differently without the role of the non-European peoples?

4. (One hour) Choose one.

   a. Is the history of Western society in the last 350 years or so a history of progress? (Make sure you define "progress.")

   b. "Until 1945, the history of Europe is a history of warfare: preparing for it, conducting it, concluding it. This was the price of a continent of nation states, and democracy did not improve the situation." Discuss.

Clearly, something more than mastery of dates and names is wanted here. Rigorous and creative analysis is required — and a good amount of style in answering is no mere frill. Observe, too, that students have significant choice, true of most good college exams. But of most importance in this assessment are the implicit and unspecified standards and criteria of performance expected — a paradox of high-quality, upper-level education. We (by-and-large correctly) assume that students in good colleges ought to understand the kind and quality of the answers that are required here. In an excellent college, where students have long practiced the construction and refinement of historical analyses, such vague instructions as "Discuss" or "Compare" are (or ought to be) sufficient. A good pre-collegiate education prepares you for such exams. But how prepared are students who graduated from the average system — the "content-coverage school district"?

They clearly are not well-prepared, given our non-system of educational tribalism, where each test is built out of parochial myths or eccentric tastes about what really matters. The graduates of Recall and Regurgitate High School are in for a rude shock — an immoral shock — that results from our "system" being no system at all. Why aren't most high school teachers obligated to know the kinds of tasks and grading standards that are required for success in good colleges? What policies might induce them to anchor their work in exemplary upper-level work — as successful coaches routinely do in all schools? What untold harm is done to students who find out *too late* that they are unprepared for their future aspirations because their schools felt no need to go beyond mandated tests to scout out and reveal the standards in force at the higher levels of education and employment? Why are these matters so often left to test companies who are more interested in the most generic and cost effective (i.e., marketable) test rather than exemplary and enabling assessment?

I am talking about *maximally qualifying* students for admission to the most worthy programs, not ensuring their admittance to any particular place.[11] Nor should we be dissuaded from this task by looking at current admissions tests. The issue is better protection of students and schools from the inevitable harm of using exclusively secure and indirect tests to rank and sort efficiently. (It is also unconscionable that students and schools pay the entire fee for colleges' admissions practices.)

Suppose we then think of our job in the K-12 arena as maximizing the likelihood that all students can handle the tasks found in the best universities and places of employment. What

will all students need from our assessment system? For one, our examining must involve recursive practices more routinely. We will want to know whether students are making progress in performing at important tasks. Such a view means breaking another bad habit: viewing assessment as the testing of what was taught — after teaching and learning are over. Assessment must be both educative and enabling, effectively communicating the kinds of tasks and the quality of performance we eventually expect students to handle when they are adults.

It follows that we should routinely assess student performance against exit-level tasks and standards as opposed to age-cohort expectations and norms only, as we do now. We should do more routinely what Illinois does in its state writing assessment: compare student papers from three different grades against the same standards and criteria. We should do what many vocational programs and some high schools now do: assess their students against entry-level job or college performance standards. We then ask: what kinds of "scaffolding" or "training wheels" would have to be provided to younger, less experienced students to give them guided practice in handling such tasks? This is one question we should be addressing regularly in our assessment and syllabus designs but rarely do. Do we, then, have a *system*?

Other countries have done a bit better in this matter than we have — ironically, through external examinations. Leaving aside the wisdom of our having national exams (in a country with no national curriculum???) we ought at least to appreciate the informative and standardizing impact of high-stakes, syllabus-linked exams that directly relate to entrance requirements in college, as occurs in Canada and most European countries. In Alberta, for example, half a student's grade for the senior year is his or her exam score; the other half is the teacher's grade for the work of the year. One noteworthy result: Teacher grades are very reliable, and there is extremely high correlation between local grades and exam scores. One can quickly see why. A teacher would be pilloried who gave consistent 90s to students who ended up getting 40s on the exam, given the 50/50 grading system. Known, shared and locally used standards are in everyone's interest, in other words — all the more so since the exam scores in most cases count toward admission and student "majors" at all Alberta's colleges and trade schools. We severely underestimate the power of students to meet high standards in this country because we so rarely inform them well ahead of time about those (unbending) standards, and we so rarely make them real in their day-to-day work as they do in Alberta.

One ironic virtue of living in a world of external examinations is that the teacher becomes the student's ally, not the judge, jury and executioner and "guess-what-I-want" tester (we see this same relationship in good Advanced Placement courses in America).[12] Students are then far better equipped with knowledge of standards, criteria and tasks. All teachers in Alberta are required to teach students about prior exam questions, show anchor papers and share the scoring criteria for the essays. Consider this example from a Canadian provincial exam in history for use in scoring the required essays (in addition, the student receives the four scoring rubrics for each set of traits scored):

**Marks are awarded depending upon how well a student meets the following requirements:**

1. **Defense of Position**

    a. Evidence of a Position
        - Is the writer's position evident?

    b. Logic and Persuasiveness
        - How well-chosen are the examples?
        - How well does the writer make the examples serve the position?
        - Are the arguments based on scholarship and reason rather than unsupported assertions?
        - Are the arguments based on valid assumptions?

2. **Discussion of Value Positions**

    a. Identification of Value Positions
        - Are two or more value positions indicated?

    b. Thoughtfulness
        - How adequately developed is the discussion of alternative values?
        - What depth of understanding of the issue is demonstrated?

3. **Presentation of Examples or Case Studies**

    a. Relevance
    b. Accuracy
    c. Comprehensiveness

4. **Quality of Language and Expression**

    a. Organization
    b. Convention
    c. Syntax and Vocabulary

**Please ensure that all students have prior access to this scoring guide.**

Conclusion: Honoring the Purposes of Liberal Education

The problems with all educational policy concerning accountability grow out of a flawed view of student assessment. Shouldn't assessment, even large-scale assessment, be designed to assist the student by embodying standards and offering usable feedback, not merely "measuring" performance through proxies? By assuming that mandated tests should primarily serve overseers, not teachers and learners, our students are routinely "tested" but never challenged, understood or inspired.

To "assess" is to "sit with" the learner, as the word's roots reveal. Our task is to find out whether students can demonstrate the ability to employ knowledge effectively and elegantly. A multiple-choice test is a simplistic audit, and, as in business, an audit of the books bears little relation to the quality of the company's performance. We have become blind to the harm of substituting these machine-scored items for authentic intellectual performance.

The harm stems from the form of the instruments and their effect on teaching. The format of current tests is a structure at odds with modern views of knowledge.[13] Genuinely "good" answers are not "correct" (as multiple-choice tests require — and ultimately teach) but justified or well supported. Real problems grow out of idiosyncratic courses and contextual concerns; they are not generic "items." Answers to real questions are not self-evidently right or wrong, but require analysis — maybe even dialogue — for their soundness or plausibility to be established. By not evoking or requiring student production and by not using judges to determine the adequacy of student reasoning or depth of understanding, standardized tests prevent us from assessing the cardinal virtues of importance to both educators and employers: craftsmanship, thoughtfulness, initiative, persistence and rigor in intellectual work.

An education should provide all students with an intellectual voice, empathy and autonomy as a thinker. Traditional assessment points in the opposite direction. It never asks us to produce our own products. It never asks us to use our judgment or justify our conclusions. It never enables us to explain our seemingly "wrong" answers. It never asks us to respond to the arguments of those who disagree with us. Yet, aren't these the intellectual challenges and achievements we value? Whether or not one ever attends higher education, all schooling — hence, all assessment — should be infused with these core values. Only then will all students understand what our ultimate expectations really are — and why they matter.

The impatient among us will argue that there is no viable alternative in the large-scale policy arena to efficient, indirect tests. Yet, basic questions about this urge to mandate superficial measurement go perpetually unasked. Just what is being measured, at a penny per student, so that a school is genuinely assessed and accountable for what is in its control? Most such tests are less sensitive to value-added achievement than to the demographics and aptitude of the population. Why, for example, isn't the "marketplace" (of employer hirings, upper-level educational admissions and vocational and civic success) a better mechanism for comparing schools and offering incentive for local change? With no national curriculum and an appropriately diverse set of programs and schools, what can possibly be worth comparing on a single test?

There is a painful irony in the call for more mandated tests as the answer to our woes. The move toward a more centralized "planned" education via state and national policy is occurring just as the rest of the Western political world embraces the wiser view that renewal depends upon liberated, local enterprise.[14] Such mandates pose unseen threats to the most sought-after graduate and undergraduate programs in the world since higher education is about the freedom of students and faculty to do <u>their own</u> work, but to do it well — quality, without either imposed orthodoxy or uniformity. The thoughtless imposition of tests will thus

do more than impoverish the will and capacity of school faculties to raise expectations and hold high(er) standards. It will undermine the very idea of the university as a place for promoting and rewarding sound, free thinking.

I thus seek the ultimate in what is now called "outcome-based" education and "site-based decision-making" — ensure that all *local* testing, across the K-graduate school continuum, be composed of the worthiest tasks, leading toward our best exit-level challenges, adapted to all grades as necessary. "But few students go on to upper-level college and graduate work!" Yes, but how many students are prevented from doing so by K-12 syllabi and tests thoroughly at odds with thoughtful inquiry and effective production? Coherent and authentic assessment is an *equity* and *empowerment* issue just as much as it is an *accountability* issue.

The well-trained upper-level student knows what far too few younger students ever get to know: experts disagree for good reasons; important questions are not so much answered as explored effectively; knowledge is not so much "imparted" as effectively induced, constructed, used, tested, extended, criticized or transformed into unfamiliar truths. All assessment in education should reveal that good learning is not revealed by what "right" answers" one knows but by whether one can provide well-reasoned and appropriate results. Students must learn through assessment that all real-world "testing" of ideas occurs through interaction — a "dialogue" between people, ideas, specific situations and contextual constraints that compel the creative adaptation of the more general truths and skills learned in schooling.

We then teach a lesson that is morally as well as intellectually empowering: the judge, too, is subservient to standards and criteria. Grades and scores must not represent the apparent or mysterious tastes of judges. Otherwise, we teach students to passively end up trying to figure out "what they want." A system built out of secure multiple-choice questions, with no recourse to dialogue with assessors, is inquisitional. It violates the spirit of liberal education and ensures that many students never make it to the end — except those who already trust adults.

But it is not sufficient to design assessment practices that enable students to understand the standards by which they will be judged as adult thinkers. Our assessment practices must teach students that the tasks, criteria and standards they encounter in schools and colleges are the appropriate ones for all rational inquiry and a fruitful, intellectual life. This means undoing the dogmatic, anti-intellectual effect of standardized, multiple-choice testing: the acquired, cynical view that there is an orthodox set of propositional truths that everyone "just has to know," that all questions have fixed, clear, official and unquestionable answers, and that machine-scored questions are "objective" and human-judged questions are "subjective" — i.e., the scoring is too soft and unreliable to use as solid evidence.

Students must learn from assessment the truth about human judgment: the grounds of sound assessment are indeed objective — even if and when judges disagree — because the criteria are neither arbitrary nor ineffable. We have succeeded in educating students, in fact, when

they have so internalized the standards and criteria of evidence and argument that they can effectively dispute an inappropriate score or grade. (Secure tests usually prevent students from even seeing an account of their specific errors). Excellence is not about satisfying elders and their apparent tastes, but is found in meeting the objective specifications of tasks, situations or problems.

This is old news in vocational, athletic, musical and dramatic settings — in performance-based learning and assessing, in other words. There, students see that "quality" is not an abstraction or a matter of teacher taste. The performance either worked or it didn't; intentions were either realized or they weren't. Self-assessment is taught and learned because the criteria and standards are clear and sensible. The students grasp the objectivity of criteria and standards by constantly having before them *models that embody them* — examples of excellence by which they might learn and toward which they can aspire. By contrast, the student in the classroom rarely sees the masterful historian, scientist or the mathematician at work or the products of their thinking so as to see that judgments of intellectual merit need not be arbitrary. The student thus rarely gains either adequate insight or the incentive that follows from experiencing what the "test" of real performance in a field is really like.

A genuinely "seamless" K-12-graduate school system would therefore have a simple motto: *No surprises, no excuses.* No surprises, because the authentic tasks and standards that link the different stages of the system together would be known to all students and teachers at lower levels and recur throughout their work. No excuses, because if the standards and measures were authentic, thoroughly explained, taught, and practiced with constant opportunity for revision and improvement, schools and students would be genuinely culpable for sub-standard performance. That is genuine accountability.

## APPENDIX A

### 1. Principles of Assessment for Better Learning

a. The interests of the students shall be paramount. Assessment shall be planned and implemented in ways that maximize benefits for students, while minimizing any negative effects on them.

b. The primary purpose of assessment shall be to provide information which can be used to identify strengths and to guide improvement. In other words, it should suggest actions which may be taken to improve the educational development of students and the quality of educational programs.

c. Assessment information should not be used for judgmental or political purposes if such use would be likely to cause harm to students or to the effectiveness of teachers or schools.

d. Every effort should be made to ensure that assessment and evaluation procedures are fair to all.

e. Community involvement is essential to the credibility and impact of assessment and evaluation processes. All parties with a direct interest should have an opportunity to contribute fully. Self-assessment is the appropriate starting point.

f. Careful consideration should be given to the motivational effects of assessment and evaluation practices.

g. In the assessment of intellectual outcomes, substantial attention should be devoted to more sophisticated skills such as understanding of principles, applying skill and knowledge to new tasks, and investigating, analyzing and discussing complex issues and problems.

h. Emphasis should be given to identifying and reporting educational progress and growth, rather than to comparisons of individuals or schools.

i. The choices made in reporting assessment information largely determine the benefit or harm resulting from the information. For this reason, the selection, presentation and distribution of information must be controlled by the principles outlined previously.

From Assessment for Better Learning: A Public Discussion Document, (1989), New Zealand Department of Education.

## 2. Recommendations from the National Commission on Testing and Public Policy

a. Testing policies and practices must be reoriented to promote the development of all human talent.

b. Testing programs should be re-directed from over-reliance on multiple-choice tests toward alternative forms of assessment.

c. Test scores should be used only when they differentiate on the basis of characteristics relevant to the opportunities being allocated.

d. The more test scores disproportionately deny opportunities to minorities, the greater the need to show that the tests measure characteristics relevant to the opportunities being allocated.

e. Test scores are imperfect measures and should not be used alone to make important decisions about individuals, groups or institutions.

f. More efficient and effective assessment strategies are required to hold institutions accountable.

g. The enterprise of testing must be subjected to greater public accountability.

h. Research and development must be expanded to create assessments that promote the development of the talents of all our peoples.

From Gatekeeper to Gateway: Transforming Testing in America, (1990), National Commission on Testing and Public Policy, Boston College, Chestnut Hill, MA.

## 3. Assessment Blue-printing and Task Design

The following pages contain design suggestions for direct assessment by performance, product, project, exhibition or portfolio. These "templates" suggest the types of situations, simulations, roles and problems that can be used to make tasks authentic and engaging for any subject matter or age group.

Common to all the ideas are three essential principles:

(1) "Higher-order" thinking and acting requires that students produce "unique products or performances" (in the words of Bloom)

(2) Task ideas can come from the modification of existing high-quality instructional activities — including such non-scholastic activities as Scouting, Odyssey of the Mind, vocational simulations and competitions, etc.

(3) Assessment tasks should reveal the types of challenges actually encountered in the field when professionals are called upon to use knowledge effectively, imaginatively and in context — i.e., the products and performances are sensitive to audience, purpose, particular constraints of the setting, cost/benefit considerations, etc.

These ideas are meant to be more than interesting, optional provocations. The assumption is that districts and schools would develop **blue-printing policies** for how all assessments should be constructed to ensure that they are maximally authentic, "higher-order" and articulated with system standards and performance targets.

a. **"Higher-order performance verbs" for use in assessment blue-printing**

Discern a Pattern

Grasp Purpose & Reach Audience

Empathize with the Odd

Pursue Alternative Answers

Achieve an Intended Aesthetic Effect

Exhibit Findings Effectively

Polish a Performance

Lead a Group to Closure

Develop and Effectively Implement a Plan

Design, Execute and De-bug an Experiment

Make a Novice Understand What You Deeply Know

Induce a Theorem or Principle

Explore and Report Fairly on a Controversy

Lay Out "Cost-Benefit" Options

Assess the Quality of a Product

Graphically Display and Effectively Illuminate Complex Ideas

Rate Proposals or Candidates

Establish Principles

Make the Familiar Strange

Infer a Relationship

Facilitate a Process and Result

Create an Insightful Model

Disprove a Common Notion

Reveal the Limits of an Important Theory

Successfully Mediate a Dispute

Thoroughly Rethink an Issue

Shift Perspective

Imaginatively and Persuasively Simulate a Condition or Event

Thoughtfully Evaluate and Accurately Analyze a Performance

Judge the Adequacy of an Appealing Idea

Accurately Self-Assess and Self-Correct

Communicate in an Appropriate Variety of Media or Languages

Complete a Cost-Benefit Analysis

Question the Obvious or Familiar

Analyze Common Elements of Diverse Products

Test for Accuracy

Negotiate a Dilemma

Make the Strange Familiar

**b. Authentic assessment —**

1. Involves tasks we value, and at which we want students to excel — tasks worth learning and "teaching to."

2. Simulates the challenges facing adults or workers in a field of study, or the real-life "tests" of civic and personal life in which our educational knowledge is required.

3. Is composed of "ill-structured" challenges that require (a) problem-clarification and knowledge in use, (b) effective use of a repertoire of knowledge, (c) good judgment in solving the problem and (d) overcoming realistic constraints to fashion an effective and appropriate response in context.

4. Focuses on the students' ability to produce a quality product and/or performance.

5. Involves de-mystified and non-secret tasks, criteria and standards; allows for thorough preparation and accurate self-assessment by the student.

6. Relies on trained assessor judgment, in reference to clear and appropriate criteria (as opposed to those most easily observed or scored).

7. Is typically composed of interactions between assessor and student. Focuses on the student's ability to justify answers and respond to follow-up or probing questions.

8. Involves patterns of response and behavior, consistency of performance: emphasis is on consistency of quality, habits of mind.

9. Calls upon different forms of communicating and means of displaying mastery — in an integrative "performance" or set of products, e.g., an oral report, supported by a paper.

## APPENDIX B:
## LONGITUDINAL SCORING SCALES FROM AUSTRALIA AND GREAT BRITAIN

### 1. Proposed Scoring System for K-10 Assessment (U. K.) — Writing

| Level | Description |
|---|---|
| | *Pupils should be able to:* |
| 1 | • Use pictures, symbols or isolated letters, words or phrases to communicate meaning. |
| 2 | • Produce, independently, pieces of writing using complete sentences, some of them demarcated with capital letters, periods or question marks.<br>• Structure sequences of real or imagined events coherently in chronological accounts.<br>• Write stories showing an understanding of the rudiments of story structure by establishing an opening, characters and one or more events.<br>• Produce simple, coherent non-chronological writing. |
| 3 | • Produce, independently, pieces of writing using complete sentences, mainly demarcated with capitals, periods and question marks.<br>• Shape chronological writing by beginning to use a wider range of sentence connectives than "and" and "then."<br>• Write more complex stories with detail beyond simple events and with a defined ending.<br>• Begin to revise and re-draft in consultation with the teacher or other children in the class, paying attention to meaning and clarity as well as checking for things such as correct use of tenses and pronouns. |
| 4 | • Produce pieces of writing in which there is a rudimentary attempt to present subject matter in a structured way (e.g., title, paragraphs, verses); in which punctuation is generally accurate; and where evidence exists of ability to make meaning clear to readers.<br>• Write stories that have an opening, a setting, characters, a series of events and a resolution.<br>• Organize non-chronological writings in orderly ways.<br>• Begin to use some sentence structures different from those most characteristic of speech (e.g., subordinate clauses).<br>• Attempt independent revising of their own writing and talk about the changes made. |

| | |
|---|---|
| 5 | • Write in a variety of forms, e.g., notes, letters, instructions, stories, poems, for a range of purposes, e.g., to plan, inform, explain, entertain, express attitudes or emotions.<br>• Produce pieces of writing in which there is a more successful attempt to present simple subject matter in a structured way, e.g., by lay-out, headings, paragraphing; in which sentence punctuation is almost always accurately used, and in which simple uses of the comma are handled successfully.<br>• Write in standard English (except in contexts where non-standard forms are appropriate) and show an increasing differentiation between speech and writing, e.g., by using constructions which decrease repetition.<br>• Assemble ideas on paper and show some ability to produce a draft from them and to redraft or revise as necessary. |
| 6 | • Write in a variety of forms for a range of purposes, showing some ability to present subject matter differently for different specified audiences.<br>• Make use of literary stylistic features, such as alteration of word order for emphasis or the deliberate repetition of words or sentence patterns.<br>• Show some ability to recognize when planning, drafting, redrafting and revising are appropriate and to carry these processes out. |
| 7 | • Produce well-structured pieces of writing, some of which handle more demanding subject-matter, e.g., going beyond first-hand experience.<br>• Make a more assured and selective use of a wider range of grammatical and lexical features appropriate for topic and audience.<br>• Show an increased awareness that a first draft is malleable, e.g., by changing form in which writing is cast (from story to play), or by altering sentence structure and placement. |
| 10 | • Write, at appropriate length, in a wide variety of forms, with assured sense of purpose and audience.<br>• Organize complex subject matter clearly and effectively. Produce well-structured pieces in which relationships between successive paragraphs are helpfully signalled.<br>• Make an assured, selective and appropriate use of a wide range of grammatical constructions and of an extensive vocabulary. Sustain the chosen style consistently. Achieve felicitous or striking effects, showing evidence of a personal style. |

## 2. From the Victoria, Australia, Literacy Profiles: Reading

- Student's behavior charted over time using nine progressive "bands" for grades K-10
- Teachers keep running records, based on observation and tasks assigned

**Reading Band A**

Holds book the right way up
On request, indicates the beginning and end of sentences
Refers to letters by name
Responds to literature (smiles, claps, listens intently)
Turns pages front to back
Locates words, lines, spaces, letters
Identifies known, familiar words
Joins in familiar stories
Shows preference for particular books

**Reading Band B**

Takes risks when reading
Asks others for help with meaning and pronunciation
Recognizes root words within other words
Creates ending when text is left unfinished
Uses pictures for clues to meaning of text
Predicts words
Makes a second attempt at a word if it doesn't sound right
Retells with approximate sequence

**Reading Band D**

Selects books to fulfill own purposes
Substitutes words with similar meanings when reading aloud
Themes from reading appear in art work
Reads materials with a wide variety of styles and topics
States main idea in a passage
Self-corrects, using knowledge of language structure or sound-symbol to make sense of a word or phrase
Uses vocabulary and sentence structure from reading in written work as well as talk

**Reading Band F**

Selects relevant passages to answer questions
Maps out plots and character developments in novels
Makes connections between texts
Discusses styles used by different authors
Offers reasons for response provoked by text
Justifies own appraisal of text
Formulates questions and finds relevant information from reading
Varies reading strategies according to purposes of reading and nature of text
Discusses author's intent
Forms generalizations about a range of genres including myth, short story
Offers critical opinion or analysis of reading passages in discussion

**Reading Band I**

Explains textual innuendo and undertone
Identifies and explains deeper significances of text
Discusses and writes about the author's bias
Interprets analogy, allegory and parable in text
Defends each interpretation of text
Analyzes the cohesiveness of text as a whole

## APPENDIX C
## VICTORIA, AUSTRALIA "WORK REQUIREMENTS"

Work Requirements — Literature

Keeping a Reading Journal

- Personal responses to texts
- Notes from group and class discussions
- Short responses to textual issues and questions
- List of texts read, with comments

Developing a Portfolio

- Three "finished responses" included for each of two units; one must be "discursive-analytical," another "creative"
- At least one response in oral form

Presentation of a Review of Reading

- Based on student's independent reading
- Presentation to class, as well as written; mindful of audience

Writing a Text

- Finished piece should be read by an audience other than the teacher

Exploring Fiction in Television, Film or Radio

Producing an Extended Response

Interpreting a Text for Performance

Comparing Readings

Investigating Contexts

Analyzing a Review

Presenting a Written Analysis

Work Requirements — Chemistry (built around four units: materials, chemistry in everyday life, chemistry and the marketplace, and energy and matter)

**Modeling Structures**

- Construct models to represent continuous lattices, discrete molecules
- Inspect and evaluate models of nuclear atom, polymers, ceramics, alloys
- Discuss strengths and weaknesses of models

**Investigation of Waste Materials**

- Experiment on properties of waste materials
- Design and perform experiment on how to treat a waste sample, e.g., contaminated water
- Identify waste materials generated during production of a useful material, strategies used for dealing with waste
- Discuss advantages and disadvantages of methods used in waste treatment and their implications for continued use of material

**Investigation of Oxidation-Reduction Reactions**

- Perform a range of experiments to observe oxidation reactions, demonstrate electron transfer nature by constructing simple galvanic cells
- Design and perform an experiment which relates to metal reactivity and corrosion protection, evaluate the experimental design

**Other Work Requirements:**

Media File
Record of Reactions
Investigation of an Instrument
Changing Models of the Atom
Food: Annotated Flow Charts
Investigation of Useful Materials
Investigation of a Chemical of Local Importance
Product Analysis
Investigation of Equilibrium
Investigation of Periodic Table
Concept Map
Investigation of Properties of Water and Atmosphere

**Laboratory activities should occupy at least 25% of each of the 4 units.** Students should record accurate details of lab activities in a log book. Such records should be used to prepare reports; in each unit students should prepare two full reports.

## REFERENCES


Archbald, D. and Newmann, F. "The Functions of Assessment and the Nature of Authentic Academic Achievement," in *Assessing Achievement: Toward the Development of a New Science of Educational Testing*, Berlak (ed.). Buffalo, NY: SUNY Press, 1989.

Archbald, D. and Newmann, F. *Beyond Standardized Testing: Authentic Academic Achievement in the Secondary School.* Reston, VA: NASSP Publications, 1988.

Astin, Alexander. *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education.* New York: American Council on Education/Macmillan Publishing Co., 1991.

Bloom, B., Madaus, G. and Hastings, J. T. *Evaluation to Improve Learning.* New York: McGraw-Hill, 1981.

California State Department of Education. *A Question of Thinking: A First Look at Students' Performance on Open-Ended Questions in Mathematics.* Sacramento, CA, 1989.

California State Department of Education. *Writing Achievement of California Eighth Graders: Year Two.* Sacramento, CA, 1989.

Chubb, John E. and Moe, Terry M. *Politics, Markets & America's Schools.* Washington, D.C.: Brookings Institution, 1990.

College Board. *Evaluating the AP Portfolio in Studio Art* and "General Portfolio Guidelines" in *Advanced Placement in Art.* Princeton NJ: ETS/CEEB, 1986.

Connecticut Department of Education. *Toward A New Generation of Student Outcome Measures: Connecticut's Common Core of Learning Assessment.* Hartford, CT: State Department of Education, Research and Evaluation Division, 1990.

Department of Education [New Zealand]. *Assessment for Better Learning: A Public Discussion Document.* Wellington, NZ, 1989.

Department of Education [New Zealand]. *Learning and Achieving: Second Report of the Committee of Inquiry into Curriculum, Assessment and Qualifications in Forms 5 to 7.* Wellington, NZ, 1986.

Department of Education and Science and the Welsh Office [U.K.]. *English for Ages 5 to 16: Proposals of the Secretary of State for Education and Science.* London, England: Department of Education and Science and the Welsh Office, 1989.

Department of Education and Science and the Welsh Office [U.K.]. *National Curriculum: Task Group on Assessment and Testing: A Report*. London, England: Department of Education and Science, England and Wales, 1988. [A brief "Digest for Schools" is also available.]

Educational Testing Service. *Learning By Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics*. A Report on the NAEP pilot of performance-based assessment, 1987. (The full report: "A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics," ETS Report #17-HOS-80).

Educational Testing Service. "A Guide for Assessing Prior Experience Through Portfolios," working paper #6 of the Committee for the Assessment of Experiential Learning. ETS, 1975.

Elbow, P. "Trying to Teach While Thinking About the End" and "Evaluating Students More Accurately" in *Embracing Contraries: Explorations In Teaching And Learning*. New York: Oxford University Press, 1986. [The former chapter originally published in Grant, Elbow et al. (1979)].

Forrest, A. and Steele, J. "Defining and Measuring General Education Knowledge and Skills," Technical Report #1976-81. College Outcome Measures Project, American College Testing Program, 1982.

Frederiksen, J. and Collins, A. "A Systems Approach to Educational Testing," *Educational Researcher*, vol. 18, no. 9, December 1989.

Frederiksen, N. "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, vol. 39, no. 3, March 1984, p. 193-202.

Gardner, H. "Assessment in Context: The Alternative to Standardized Testing" in *Report to the Commission on Testing and Public Policy*, B. Gifford, ed. Boston: Kluwer Academic Press, 1989.

Glaser, R. "Cognitive and Environmental Perspectives on Assessing Achievement" in *Assessment in the Service of Learning*. 1987 ETS Invitational Conference Proceedings.

Glaser, R. "The Integration of Instruction and Testing" in *The Redesign of Testing for the 21st Century*. 1985 ETS Invitational Conference Proceedings.

Grant, G., Elbow, P., et al. *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education*. San Francisco: Jossey-Bass, 1979.

Gronlund, N. E. and Linn, R. L. *Measurement and Evaluation in Teaching*, 6th ed. New York: Macmillan and Co., 1990.

Koretz, D. "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" *American Educator*, vol. 12, no. 2, Summer 1988.

Madaus, G. "The Influence of Testing on the Curriculum" in *Critical Issues in Curriculum.* 87th Yearbook of the National Society for the Study of Education, Part I, 1985.

Ministry of Education, Victoria, Australia. *Literacy Profiles Handbook: Assessing and Reporting Literacy Development.* Melbourne, Victoria: Education Shop, 1990.

National Commission on Testing and Public Policy. *From Gatekeeper to Gateway: Transforming Testing in America.* Chestnut Hill, MA: NCTPP, Boston College, 1990.

Resnick, D. P. and Resnick, L. B. "Standards, Curriculum and Performance: A Historical and Comparative Perspective," *Educational Researcher*, vol. 14, no. 4, April 1985, pp. 5-21.

Vermont Department of Education. *Vermont Writing Assessment: The Portfolio* (draft: October 1989).

Wiggins, G. "Standards, Not Standardization: Evoking Quality Student Work," *Educational Leadership*, vol. 48, no. 5, February 1991, pp. 18-25.

Wiggins, G. "Secure Tests, Insecure Test Takers" in *The Prices of Secrecy: The Social, Intellectual and Psychological Costs of Testing in America.* A Report to the Ford Foundation, Judah Schwarz, editor, 1990.

Wiggins, G. "The Futility of Trying to Teach Everything of Importance," *Educational Leadership*, vol. 47, no. 3, November 1989a.

Wiggins, G. "A True Test: Toward More Authentic and Equitable Assessment," *Phi Delta Kappan*, vol. 70, no. 9, May 1989b.

Wiggins, G. "Teaching to the (Authentic) Test," *Educational Leadership*, vol. 46, no. 7, April 1989c, pp. 41-47.

Note: The April 1989 issue of *Educational Leadership*, the May 1989 issue of *Phi Delta Kappan* and the December 1989 issue of *Educational Researcher* were all devoted to assessment and testing reform issues.

## ENDNOTES

1. This is a revised version of a paper presented for discussion at a conference on assessment policy across the educational system, sponsored and hosted by the Education Commission of the States in June 1991 in Breckenridge, Colorado.

2. The recently announced New Standards Project, headed by Lauren and Daniel Resnick and Marc Tucker, would use such a moderation procedure for developing a variety of regional assessments as part of a new national examination system for grades 4, 8 and 12. And the development of comparable longitudinal data bases, established through partnerships by higher education institutions, has been well documented and analyzed by Astin (1991).

3. As translated by Naomi Lewis (1981).

4. Consider, by contrast, the recent British manifesto underlying their new national assessment design (which will rely heavily on classroom-based, teacher-overseen assessment): "The national system should employ tests [of] a wide range of modes of presentation, operation and response....a mixture of tests, practical tasks and observations should be used in order to minimize curricular distortion....The group has no doubt that a successful system depends upon teachers' confidence in it....the report recommends that teachers be given more support in assessment, especially by providing them with a wider range of diagnostic tests....the tests should be so unobtrusive as to seem to students no different than typical classroom activity." (from the TGAT Report)

5. Educators in other countries derisively refer to our fetish for increasingly relying on multiple-choice tests as "the American solution" to educational problems.

6. See the appendix for some sample policy statements designed to ensure that the purpose of assessment is honored by policy and practice.

7. See Ewen in Grant, Elbow et al. (1979) on the competencies of the liberal arts.

8. See Astin (1991) for a detailed account of "value-added" assessment of higher education.

9. For an excellent discussion of authentic assessment and how to maximize the beneficial impact of tests on schooling — "systemic" validity — see Frederiksen & Collins (1989).

   What Messick and others in the psychometric community have called "consequential validity" — an idea so obvious that one is stunned to realize that few within this community saw the need until recently to factor the effect of testing into test design — is usually applied to test and teaching practice but too rarely to the

views of what a curriculum is. A critical need is to rethink what we mean by "writing curriculum" so that we don't think of either instruction or assessment as measuring whether what was "covered" was "learned" (i.e., recalled or used in a low-level way).

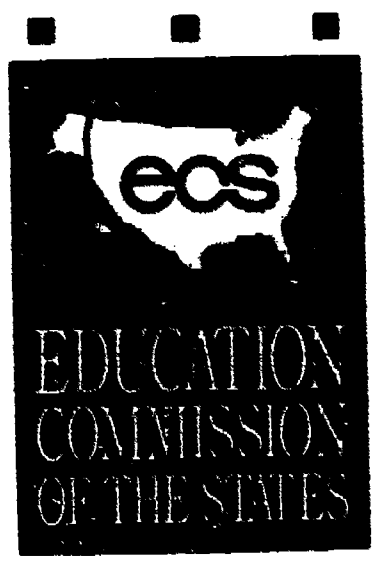
10. As reported in Peters (1989), spoken by a Dow Chemical quality control engineer. Cf. Wiggins (1991).

11. We know the bad news: Harvard, Stanford, Howard, Earlham and others are running "offices of rejection." The 8:1 ratio of application to acceptance or worse is irrelevant, however, to the moral obligation of all lower-schooling faculties to equip students to be maximally prepared, should the "right fat letter" come on April 16th.

12. See Elbow (1986) on this point; Cf. Astin (1991) for his distinction between tests as incentives and tests as providing feedback, where the same point is made.

13. While such tests have long been criticized as steeped in out-dated learning theory, the epistemological implications are more troublesome. It is a mistaken view to suggest that "knowledge" consists of facts and unequivocal right answers, as opposed to well-reasoned or supported claims and arguments. The theory of knowledge and understanding embedded in traditional tests is at odds with the modern arts and sciences; it harkens back to a medieval view of knowledge. See "The Futility of Teaching Everything of Importance" (Wiggins 1989a).

14. This is the core of the argument made in the widely discussed Chubb and Moe book (1990). Most of the press attention focuses on their call for "choice" but that call is a proposed solution. The problem, as they see it, is that American education is incapable of reforming itself as long as mandates from external governing bodies, not market forces working on autonomous schools, determine policy.


ecs
EDUCATION
COMMISSION
OF THE STATES