The Tests We Need and Why We Don't Quite Have Them

by E. D. Hirsch, Jr.

Originally published in *Common Knowledge*, Volume 13, #1, Winter 2000

I. Two Kinds of Tests

Statewide content standards are beginning to spawn high-stakes tests that have evoked furious opposition — not without cause. The greatest outcry (to reach my ears) has been occurring in Virginia and Massachusetts where the new tests are based on fairly specific content standards. In Kentucky and Maryland, where the high-stakes tests are based on vague general skills more than on specific curriculum content, the protests seem mild by contrast. Yet despite the louder outcry against curriculum-based tests, I believe they hold far more promise than skills-based tests to promote significant gains in achievement and equity.

It has to be conceded that, under present circumstances, the backlash against curriculum-based tests has been warranted. The policy makers who have instituted these high-stakes tests have made two strategic mistakes. First, they introduced content standards and tests before providing teachers and students with detailed outlines and teaching materials which define what the content standards really are. They have put in place no adequate system for training teachers in the subject matters identified by the content standards. They have failed to do the hard work of deciding which aspects of the content are the most essential to be included in textbooks, teacher seminars, and tests — a lack of specificity and selectivity which has made at least some of the tests less reasonable and fair than they should be.[1] Between the furious opponents of the curriculum-based tests and their determined advocates there seems to be no middle ground. Yet each party seems right in some respects — except for the anti-test extremists who want no objective statewide tests at all. The more reasonable critics of curriculum-based tests rightly object to inadequate guidelines and materials, and occasional flaws in the tests themselves, and they correctly observe that five decades of content-indifferent schooling and content-poor teacher preparation cannot be reversed overnight.

On the other side, determined test advocates are right that curriculum-based tests, are the fairest and most effective means of achieving the aims of democratic schooling. Test advocates should accept the reality that the needed improvements in teacher preparation, in teaching materials, and in the tests themselves cannot occur overnight.But I find myself squarely on the side of the test advocates in resisting any attempt to exploit the necessary slowness of the process as an excuse simply to call a moratorium on curriculum standards and curriculum-based tests. They are the most promising educational development in half a century.

How these curriculum-based tests should be phased in as criteria for student promotion and graduation is a practical and political question to be decided in a democracy by the representatives of the people. I want to shed light on a technical issue that can be useful in helping to make such policy decisions better informed — that is the differences and the connections between competency-based tests and curriculum-based tests.

Competency-based tests sample knowledge from a very broad range of domains, which enables the tests to exhibit a reliably high correlation between test scores and real-world competencies. Curriculum-based tests are narrower. They try to determine how well specific content standards in a particular domain for a particular age group have been learned. Whereas competency tests indicate overall achieved ability, curriculum tests indicate whether specific knowledge has been gained. The astute reader will perhaps see where I am going — that a well-devised curriculum, monitored by good curriculum-based tests should, over time, extend the breadth of a student's knowledge, and thus raise scores on broad-gauged competency-based tests. Since an indispensable aim of schooling is to increase student competency, the public has a right to demand that results on the two kinds of test should in due course show a positive correlation. A main purpose of this essay is to explain why good curriculum-based tests, based on good content-standards are the surest and most democratic means of raising scores on competency-based tests and achieving real-world competencies.

An excellent example of competency-based tests would be standardized reading tests such as the verbal portions of the Stanford 9, the ITBS, CTBS, the Nelson-Denny reading test, and so on. While these are norm-referenced instruments which rank students against each other in percentiles, they can also be scored to indicate a student's "grade-level" of reading comprehension. A score of 5.2 would mean that the student is reading at the level the average student has reached by the second month of grade 5. These grade-level calibrations (which have been criticized on various grounds) could also be translated into absolute scores which can be equated over many decades. All of the well-established reading tests are valid, reliable and highly correlated with one another.[2]

What sorts of questions are asked on a standardized reading test which cause it to indicate so reliably academic achievement and readiness? In the earliest-grade versions, there are of course questions about sounds and letters. Later versions include questions about vocabulary, the meanings of individual sentences, and the implications of passages from literature, the natural sciences, the social sciences, practical affairs and several other domains. How could such a test, disconnected from any specific curriculum, so reliably calibrate academic achievement, learning-readiness, and even real-world competency? One needs to offer not just the ample evidence that this claim is true, but to provide a credible theory which explains the strong correlation between reading and general competency.

To the extent that any clear theory at all underlies much educational research, it is often unspoken and may be inconsistent with widely accepted scientific opinion. In the sciences, a damaging criticism in peer review is to call a research finding "a-theoretical." Ever since Duhem, such indifference to theory has been understood as a scientific weakness, because empirical evidence can always be interpreted in multiple ways.[3]

For instance the claim that the planets go in complicated orbits round the earth rather than the sun fits the observational facts; it's just not a highly probable theory. The problem of a-theoreticity is especially severe in educational research, which is beset by a blizzard of uncontrolled variables such as teacher quality and outside-school influences whose effects cannot be estimated confidently without applying the most plausible theory.

So I think it will be useful to state some of the theoretical principles which explain why good competency tests in reading turn out to be powerfully indicative of achieved abilities that go far beyond reading. Such a theory has the additional benefit of explaining the potency of curriculum-based tests.

1. Reading has been shown to be a process of mentally phonicizing language, rather than being a separate linguistic process.[4] The interpretation of the written word is a re-enactment of the interpretation of the spoken word. Many of the conventions used in written language are used in speaking and listening. This mental re-enactment of speech explains why reading ability is correlated with general communicative competence — the ability to understand and make oneself understood in oral as well as written speech.

2. Such general communicative competence is required for effective social intercourse in modern society, and is especially critical in schooling, where it forms the basis for understanding the oral and written communications of other people, including teachers.

3. The level of one's reading ability (as reflected in the vocabulary items and passage-types on a reading test) predicts the level of one's ability to learn new things. A person learns new things by associating them with things already known. Scoring high on a reading test requires a broad vocabulary which represents broad knowledge that offers multiple points of association for gaining further knowledge. The more you know, the easier it is to learn still more — a principle well established in cognitive psychology.
This is the critical element of the theory. Breadth of knowledge is the single factor within human control that contributes most to academic achievement and general cognitive competence. Breadth of knowledge is a far greater factor, for instance, than socioeconomic status. The positive correlation between achieved ability and socioeconomic status is .422, whereas the correlation between achieved ability and general information is .811. This little-known and quite momentous fact means that imparting broad knowledge to all children is the single most effective means of narrowing the competence gap through schooling. [5]

4. A score on a test of reading ability shows the degree to which this broad knowledge is readily deployable. A merely passive vocabulary which cannot be marshaled and used critically for reading comprehension is inert knowledge. Psychologists use terms like "accessibility" and "availability" to describe such actively usable knowledge. Accessibility of knowledge is attested to by a person's ability to bring that knowledge to bear in comprehending and analyzing the diverse passages in the test.[6]

In sum, theory predicts that a good reading test will indicate students' level of communicative competence, their breadth of knowledge, and their ability actively to apply that knowledge to learning new things. Theory further predicts that these competencies will correlate well with job performance and the capacity to be an active citizen, because communicative competence and the ability to learn new things are highly important skills in meeting the duties and responsibilities of the modern world.

These predictions are confirmed by massive evidence.

1. Scores in Early Reading Tests Predict Scores in Later Reading Tests. The more one reads, the more automated becomes the process, and, through reading itself, the broader becomes one's knowledge and vocabulary, and consequently the more readily one understands ever more difficult matter.[7]

2. Scores on Reading Tests Predict Grades in School. There is a positive correlation between reading scores and academic achievement.[8]

3. Scores on Reading Tests Predict Job Performance. Obviously, reading scores do not predict whether somebody can fix your car's engine. But, according to studies conducted by the armed services, reading scores do predict how readily and well a person will learn to fix your car's engine.[9]

4. Scores in Reading Tests Predict Income. Given the causal connections between communicative ability, learning ability, and job performance, it is not surprising that superior job skill should be rewarded, on average, with superior pay.[10]

II. Both Kinds of Test Are Currently Needed

The scores on a reading test or other competency test may sometimes be relatively independent of the quality of schooling. One's reading score is better predicted by one's family environment and the amount of reading one has done than by the school one attends. This is a version of the finding by Coleman that influences outside the school are more determinative of academic achievement than influences inside the school. This is not an inevitable sociological law but rather a persistent feature of current American schooling, and it does not hold with the same force in France or Sweden.[11]

The gap-closing educational results in these countries remind us that an important purpose of democratic schooling is to help able people overcome accidents of birth and circumstance. I believe that educational policy in a democracy should aim to create a system of schooling in which scores on reading tests depend much more on school influences than they recently have in the United States.

Schools can accomplish this egalitarian purpose by making students better readers, i.e., causing them to score higher on competency tests, whether or not they come from educated homes. This goal can be reached only by an effective, cumulative curriculum which gradually builds up the knowledge and vocabulary that is being sampled in a reading test. This seems to me a criterion that should be met by state curriculum standards: Will teaching this content provide children with high communicative competence and the ability to learn new things, no matter what their home disadvantages may be?

This democratic criterion means putting in place the very policies that have created the current backlash — setting forth grade-by-grade knowledge standards, and monitoring whether that knowledge is being gained, an aim that has won strong support in low-income districts which recognize their democratic effect of this reform.

John Bishop of Cornell has shown that educational systems which require definite content standards and which use curriculum-based tests to determine whether the curriculum has been learned greatly improve achievement for all students, including those from less-advantaged backgrounds.[12] Additional evidence in support of curriculum-based testing comes from the recent finding that gains in reading are directly proportional to the completeness with which a school implements a coherent, content-rich curriculum.[13] Put starkly, a system of coherent standards, coupled with curriculum-based tests will in fact cause achievement on NON-curriculum-based tests to rise. It will result in higher achievement overall and a narrowing of the academic gap between rich and poor.

But this change must be instituted wisely, and the critical policy decisions must not be left to technical test makers. Testing companies are very good at creating instruments which have good "psychometric properties," that is, which rank-order students in a smooth, normal curve. Curriculum-based tests should not exhibit those statistical properties, at least not at first. The tests should mainly discriminate between the students who have gained essential knowledge and those who haven't, with maybe one further category for students who give an abundance of right answers. The earliest versions of the new tests shouldn't rank-order students beyond those three categories — fail, pass, superior. Later on, in a mature, content-based system, such as those Bishop studied, more refined scores might be appropriate.

To grasp the distinction between fancy test items, which aren't appropriate, and plain ones which are, consider the following examples:

The Civil War ended in

1. 1864
2. 1865
3. 1866
4. 1867

The Civil War ended in

1. 1812
2. 1830
3. 1865
4. 1880

Few will doubt that the first question will do a better job of inducing incorrect answers. By including plenty of hard items, test makers can ensure refined, neat rank orderings among students. But it should not be left up to test makers, or even to ad hoc advisory committees, to decide whether students at a particular grade level should have such exactitude of knowledge. That decision should be made and announced in advance by those officials who create the standards and the supporting materials. Curriculum-based exams best serve their purpose, at least at first, by being straightforward and unpedantic.

These considerations lead me to suggest that state education officials should

1. recognize that we are in a transition period after half a century of content-meagre schooling, and that state departments of education must provide the means for teaching and learning the required content standards before too much weight is placed on them. Low stakes before high stakes. Given the historical context, that's only fair.
2. make public in a very clear and detailed fashion the important aspects of the content standards that are to be emphasized in teacher training, textbooks, and curriculum-based tests.

3. use the tests as devices to focus effort on productive and important learning which yields centrally useful knowledge and high competence.

4. grade the straightforward tests generously on a pass-fail basis, (with perhaps a "superior" for answering a very high number of straightforward questions) during the transition period while teachers are being trained and appropriate textbooks are being created.

5. offer, apart from the official, secure tests, informal, no-stakes, year-by-year diagnostic tests which will enable schools to detect knowledge deficits and monitor student progress.

6. resist any call for a complete test moratorium, and give no ground on the basic principle of curriculum-based tests, which are in theory and, as Bishop has shown, also in fact, the best route to improved quality and equity.

7. keep at least a few competency tests in reading, writing, and math. They should carry high stakes (but not unreasonably exalted cut-off points) so long as society agrees that our citizens need these competencies. Well-verified competency-based tests are like those little birds who tell us whether the air in the mine is safe. They reflect the reality principle in education by showing whether competence is truly being achieved.

In short, those states brave enough to have started down this path should continue and improve the policy of using curriculum-based tests, with the stakes gradually getting higher. This is the only known way of achieving the democratic ideal of making the school as effective educationally as the home. That is the appropriate norm by which content standards and tests should be measured in a democracy.

Those state tests, on the other hand, which are based on no specific content standards, mainly increase anxiety without increasing learning. They are no better than commercially available competency tests; in fact they are generally less fair and accurate. For profound theoretical reasons, these skills-tests cannot help schools narrow the achievement gap between groups.[14]

In states where good curriculum-based tests are built upon good, specific content standards, the following can be predicted for current Kindergartners and first graders. By grade seven or eight, when the content-based curriculum has "diffused knowledge" (to use Jefferson's phrase), and has done much of its compensatory work, academic achievement will have risen for all groups. Higher scores on curriculum-based tests will be well-correlated with higher scores on competency-based tests, which will show a significant narrowing of the competency gap between groups. At that point, we shall have moved closer to the ideal of a truly democratic system of education.

Notes

1. Other shortcomings in the content standards are: arbitrariness, vagueness, enforcement of a particular pedagogy, and low content in K-3 where the greatest opportunity for equity exists. On the plus side, the very openness of content standards makes them subject to improvement through experience and democratic debate. For all their flaws, they are far better for equity and quality than statewide skills-standards with high-stakes tests that encourage wasting huge amounts of school time in practicing narrow test-taking activities at the expense of education.

2. The intercorrelation of reading tests with each other form part of the technical literature accompanying the tests. Different tests published by a large company are often "equated" to the other reading tests or test components sold by the company. Researchers have found strong intercorrelations between reading scores on the Armed Services Vocational Aptitude Battery (ASVAB) and the Armed Forces Qualification Test (AFQT) on the one side and the the various standardized reading tests such as Gates-Maginitie, Nelson-Denny, and The Stanford Tests of Academic Skills. The intercorrelations determined for reading-related skills range between .99 and .87 — at the very limits of the reliability of the tests! See: B.K. Waters, J.D. Barnes, P. Foley, S. Steinhaus, D. C. Brown, Estimating the Reading Skills of Military Applicants: Development of an ASVAB to RLG Conversion Table, Human Resources Research Organization, Alexandria, VA, 1988.

3. Pierre Duhem, The Aim and Structure of Physical Theory, 1905.Original: La theorie physique: Son objet et sa structure.

4. Conrad, R. and Hull, A.J. (1964) Information, acoustic confusion and memory span, British Journal of Psychology, 55, 429-432. Hintzman, D.L. (1967) Articulatory Coding in short-term memory, Journal of Verbal Learning

and Verbal Behavior, 6, 312-316, Naveh-Benjamin, M & Ayres, J.T. Digit span, reading rate, and linguistic relativity. Quarterly Journal of Experimental Psychology, 38, 379-51. A general discussion of the underlying "phonological loop" is to be found in Baddeley, A. Human Memory: Theory and Practice, Needham, MA, Allyn & Bacon, 1998, pp. 52-70, and passim. rm 60

5. For these precise correlations see Lubinski, D., and Humphreys, L.G., Incorporating general intelligence into epidemiology and the social sciences, Intelligence, 24 (1), pp. 159-201.In cognitive science the knowledge-competence principle has become so foundational that it has branched off into different specialties such as schema theory, and expert-novice studies. Experts learn new things faster than novices do because of the high accessibility of multiple points of reference and analogy. See, for instance, J. Larkin, et al., Models of Competence in Solving Physics Problems, Cognitive Science 4, (1980) 317-48.General discussions may be found in any textbook on cognitive psychology. See, for instance, pp. 125-143 in Baddeley, A. Human Memory: Theory and Practice, Needham, MA, Allyn & Bacon, 1998.

6. See, for instance, E. Tulving, The effects of presentation and recall of material in free-recall learning. Journal of Verbal Learning and Verbal Behavior, 5, 193-197.Baddeley, A. Human Memory: Theory and Practice, Needham, MA, Allyn & Bacon, 1998, pp. 193-194.

7. Cunningham, Anne E. Stanovich, Keith E., Early Reading Acquisition and Its Relation to Reading Experience and Ability 10 Years Later, Developmental Psychology v33 n6 p934-45 Nov 1997.

8. Lindblom-Ylanne, Sari And Others, Selecting Students for Medical School: What Predicts Success during Basic Science Studies? A Cognitive Approach. Higher Education v31 n4 p507-27 Jun 1996.Blai, Boris, Jr., The Nelson-Denny Reading Test and Harcum-earned Academic Averages., Harcum Junior Coll., Bryn Mawr, Pa.Jun 1971.Gudan, Sirkka, The Nelson-Denny Reading Test as a Predictor of Academic Success in Selected Classes in a Specific Community College. Schoolcraft Coll., Livonia, Mich. Jan 1983

9. Scribner, B.L.S., Smith, D.A., Baldwin, R.H., and Phillips, R.L., Are Smart Tankers Better? AFQT and Military Productivity, Armed Forces and Society, 12, 1986, pp.193-206;Horne, D., "The Impact of Soldier Quality on Army Performance," Armed Forces and Society, 13, 1987, pp. 443-445;Fernandez, J.C., "Soldier Quality and Job Performance in Team Tasks," Social Science Quarterly, 73, 1992, pp. 253-265, C. Jencks and M. Phillips, eds, The Black-White Test Score Gap, Brookings, Washington, DC, 1998, pp. 14-15, 75-76.

10. C. Jencks and M. Phillips, eds, The Black-White Test Score Gap, Brookings, Washington, DC, 1998, pp. 445, 489-94 passim; Hofstetter, C. Richard, Sticht, Thomas G., Hofstetter, Carolyn Huie, Knowledge Literacy, and Power, Communication Research v 26 (Feb. 1999) p. 58-80.

11. R. Erikson and J. Jonsson, eds, Can Education be Equalized? The Swedish Case in Comparative Perspective, Westview Press, Boulder, CO, 1996. For translated articles on France and data see: www.coreknowledge.org.

12. Bishop, John, Do Curriculum-Based External Exit Exam Systems Enhance Student Achievement?, Consortium for Policy Research in Education, Philadelphia, PA., 1998.Bishop, John H., The Effect of Curriculum-Based External Exit Systems on Student Achievement. Journal of Economic Education v29 n2 p171-82 Spr 1998. Bishop, John H., Impacts of School Organization and Signalling on Incentives To Learn in France, the Netherlands, England, Scotland and the United States. Working Paper 93-21., National Center on the Educational Quality of the Workforce, Philadelphia, PA., 9 Nov 93. Bishop, John, The Power of External Standards., American Educator v19 n3 p10-14,17-18,42-43 Fall 1995.

13. See the 3-year Johns Hopkins Study excerpted with graphs at www.coreknowledge.org. Level of curricular implementation predicts level of reading gain over three years at multiple sites.

14. Besides encouraging time-wasting skills-practice on narrow themes, state skills-tests offer no theoretical improvement whatever over ordinary competency based tests, which is what they essentially are. Although they do encourage everyone to work harder in a narrow range, they waste time on empty exercises which cause small gain in general competence and less in equity. They preserve the test-score gap between groups instead of narrowing it, because the biggest factor in the competency gap is a gap in general information, which can be narrowed only by a long-range, coherent focus on content. This is another illustration of the importance of basing policy on strong evidence and sound theory.