DOCUMENT RESUME

ED 373 114 TM 022 043

AUTHOR Davis, Alan
TITLE Using Tests To Evaluate the Impact of Curricular Reform on Higher Order Thinking.
INSTITUTION Colorado Univ., Boulder.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 31 Aug 92
CONTRACT RR91182001
NOTE 16p.; Paper commissioned by the Curriculum Reform Project.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; *Curriculum Development; Educational Assessment; *Educational Change; Elementary Secondary Education; Matrices; Multiple Choice Tests; Program Evaluation; Sampling; Test Content; *Test Use; *Thinking Skills
IDENTIFIERS *Performance Based Evaluation; *Reform Efforts; Subject Content Knowledge

ABSTRACT

The dominant issues in considering the use of tests developed outside the classroom to measure the impact of curriculum reform on higher order thinking are reviewed by a panel interviewed for this discussion. Panel members are: (1) Stuart Kahl, (2) Robert Linn, (3) Senta A. Raizen, (4) Lauren Resnick, and (5) Thomas A. Romberg. It is conceded that, in the past, most tests used for program evaluation and accountability have not been good measures of thinking. To measure thinking, it is necessary to think in terms of systems of assessment, in which good tests may include a mixture of performance tasks, open-ended items, and multiple-choice items. Testing must occur on more than one occasion, and matrix sampling of tasks and occasions may allow inclusion of extended performance tasks. Test items should include appropriate novelty in order to test thinking, although it may not be entirely appropriate to report higher order thinking as something apart from subject content knowledge. To compare the impact of two different curricula, tests should include instructional content common to both (the intersection), or the combined content of both (the union). (Contains 5 references.) (SLD)

Using Tests to Evaluate the Impact of Curricular Reform on Higher Order Thinking

Alan Davis

August 31, 1992

The principal thrust of curriculum reform in the United States today is aimed at increasing the ability of all students to think, a departure from the emphasis on mastery of basic skills of literacy and rote mathematics of a decade ago. This thrust has drawn on an economic argument -- the perception that our manufacturing-based economy has evolved into an information-based economy in which good-paying jobs require high levels of problem solving, critical thinking, and technical expertise. It is supported by the ascendancy of cognitive learning theory over behaviorism as our dominant theory of learning -- a theory that emphasizes the active "construction" of understanding by learners. And it is fueled by evidence from the National Assessment of Educational Progress, international mathematics and science comparisons, and SAT scores that while the basic competencies of American students in mathematics and reading have improved, more advanced skills of written expression, reasoning, and problem solving have declined.

Curriculum reform is closely linked to changes in assessment. Classroom teachers require ways to assess student progress that provide more insight into the processes and products of thinking. Educational agencies, from individual schools to state departments of education, seek ways to assess the impact of the reforms on student learning, and to drive the implementation of reforms through accountability pressures. As a result, millions of dollars are currently being spent in the development of new tests

and assessment programs. A new generation of tests is rapidly emerging that require higher order thinking in content area domains and model good instruction. For example, California, Connecticut, Kentucky, Maine, Massachusetts, North Carolina, and Vermont are among a growing set of states with new assessments drawing explicitly on The National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics* with increased emphasis on applied problem solving, communication in mathematics, and use of constructed-response formats. In science, state tests in Connecticut, New York, California, and Illinois require students to measure, classify, and test hypotheses using materials and equipment. Commercial test publishers have developed tests along similar lines not tailored to any particular state's specifications (e.g., Psychological Corporation's Mathematics Performance Assessment), and tests of critical thinking are available. Countless school districts have developed new district-wide measures to monitor performance in higher order skills.

This paper is aimed at policy makers, evaluators, and educators concerned with assessing the impact of curriculum reform, especially in the context of particular programs and schools. Its purpose is to identify and clarify the dominant issues in considering the use of tests developed outside the classroom to measure the impact of curriculum reform on higher order thinking. The paper will not guide the reader in selecting among existing tests -- for that purpose, the reader is directed to comparative guides such as Northwest Regional Education Laboratory's Guide to Tests of Higher Order Thinking Skills (Arter & Davis, 1987/1991) and comparative descriptions of state testing programs (Kahl, 1992; Flexer, 1992; Davis & Armstrong, 1991).

In the process of preparing this paper, interviews were conducted with five noted authorities on using tests to measure higher order thinking in school subjects, especially science and mathematics. In the interviews, the members of the panel were asked to discuss the use of new tests to evaluate the local impacts of curriculum reform on students' thinking, particularly in mathematics and science. They were not asked to evaluate or compare particular tests, but rather to identify central issues in the selection and interpretation of existing tests for local purposes. The author has paraphrased the salient remarks of the panel and summarized some areas of disagreement for a general audience -- a process fraught with danger, since many of the ideas expressed can only fully be understood in the context of each panelist's larger, complex view of assessment. The author accepts responsibility for errors of misinterpretation and over-simplification that may have resulted.

The Panel

Stuart Kahl is vice-president of Advanced Systems in Measurement and Evaluation, a firm specializing in large scale assessment of learning. He has played a direct role in the design, development, and administration of new assessments in several states.

Robert Linn is Professor of Education at the University of Colorado at Boulder and co-director of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). He has written or edited several books on educational measurement.

Senta A. Raizen is director of the National Center for Improving Science Education in Washington, D.C. She is the author of several books and articles on science education and assessment, including Assessment of Elementary School Science and Assessment in Science Education: The Middle Years.

Lauren Resnick is Professor of Psychology and Education, and Director of the Learning Research and Development Center at the University of Pittsburgh. Her research and writing focus on the learning of mathematics and science, and she has written extensively on assessment.

Thomas A. Romberg is Sears Roebuck Foundation Bascom Professor in Education at the University of Wisconsin at Madison and director of the National Center for Research in Mathematical Sciences Education. He chaired the commission on Curriculum and Evaluation Standards for School Mathematics for the National Council of Teachers of Mathematics.

- **In the past, most tests used for program evaluation and accountability have not been good measures of thinking.**

The multiple-choice achievement tests that have dominated the external evaluation of student learning for several decades have had serious limitations and negative side effects. While it is not true that multiple-choice items can only measure recall-level knowledge, the most widely used achievement tests have not provided adequate measures of thinking. Perhaps reading comprehension tests, which do measure students' ability to select correct inferences about such thing as word meaning, author's intent, and character motivation, have been among the best of these. But multiple-choice tests give no direct information about students' ability to express ideas in writing, to design experiments, or to evaluate problems that have no single correct answer, and they reveal little about the thought processes that underlie the selection of answers. The widely-used multiple-choice tests in mathematics problem solving and science achievement barely scratch the surface of what thinking in these domains should involve.

Most important, multiple-choice tests can negatively influence instruction because they test isolated skills rather than integrated applications. A fundamental rule when tests are used for evaluation or accountability is, "What you test is what you get." When teachers see that the learning of their students is measured by how well they can select the right answer to short questions touching on discrete concepts, operations, or facts, they will teach (and select textbooks) accordingly. This phenomenon has

contributed greatly to discouraging instruction that can truly develop thinking.

- **Local evaluators should not think in terms of "selecting the right test" to measure the local impacts of curricular reform. Instead, we need to think in terms of systems of assessment.**

Education is a public enterprise, and the public and its elected representatives continue to demand evidence of how well students are performing in schools. Most states and school districts have responded by adopting tests and administering them once a year in selected grades. This approach has always been seriously flawed, but as the curriculum changes to emphasize thinking, the flaws have become more apparent, in part because appropriate assessment tasks do not easily lend themselves to short, point-in-time administration. The most interesting new tests are not tests of the sort you can take off the shelf. They consist of tasks and scoring systems that may involve different students performing different tasks, administration of tasks at various times throughout the year, and collections of samples of work and performance in portfolios. The evaluator who seeks the "best test" to assess the impact of a particular curricular reform is operating from a set of assumptions that need to be challenged.

- **A good test may include a mixture of applied performance tasks, short open-ended items, and multiple-choice items.**

If good tests model good instruction and provide insight into thought processes, they should not consist primarily of multiple choice items. In our current educational environment, teachers know how to teach to multiple-choice items, and we need to avoid reinforcing that behavior. On the other hand, the inclusion of some items in a multiple choice format does not render

a test bad. If we consider a test to be a sample of performance from a larger domain of knowledge and tasks, then there is an inevitable trade-off between the adequacy of the sample in respect to breadth and adequacy in respect to authenticity and depth. One way to minimize this trade-off is to employ an ongoing system of assessment that is tied to curriculum and operates over time. If we are limited to testing once or twice in a year, however, it is advisable to balance the depth provided by a few applied performance tasks with the breadth provided by several short-answer or multiple-choice items.

- **To evaluate the impact of a program or curriculum at the local level, you should test students at more than one point in time.**

You cannot infer very much about the local impact of curriculum reform on student learning from a single test administration, regardless of how well the test matches instruction. All student learning builds on previous experience and background knowledge. There is evidence that thinking skills, such as the ability to recognize patterns and evaluate evidence, develop more gradually than more discrete knowledge, so scores on measures of such skills are particularly sensitive to the accumulated experience of the student, in addition to the impact of recent instruction. There is a general consensus that we should take a "value added" approach to evaluating the impact of educational programs. By testing the same students at at least two points in time, one can attend to growth, which is a better indicator of curriculum impact than point-in-time testing.

- **One way to include extended performance tasks in evaluations of growth is to employ matrix sampling of tasks and occasions.**

Extended performance tasks, which require students to spend more than a few minutes in routine work and to draw on diverse concepts and strategies in their solution, are more "authentic" indicators of thinking than short-answer problems in the sense that they better approach situations students may encounter outside of school. Such tasks have significant drawbacks, however. They are time consuming, difficult to administer, and expensive to score. From the standpoint of interpreting growth, they may have another drawback: they represent memorable learning experiences in themselves.

The first time a student is asked to use water and measuring devices to determine which of three paper towels is most absorbent, thinking is tested as the student considers how to proceed. Seven months later, a repeat performance on the same task can be expected to reflect learning from the pretest itself. Students are likely to do better the second time simply because they have already performed the task, and not because they have learned science in school.

If the primary purpose of testing is to evaluate the impact of curriculum reform locally, then matrix sampling of students, tasks, and occasions is likely to be a better solution than having students repeat the same tasks. Using this approach, assessment tasks would be randomly assigned to students on the pretest. Later, the same tasks would be given again, but they would be assigned to different students so that no one student performed the same task twice. This approach allows the evaluation to

include a far greater sample of tasks, and allows accurate reporting of growth for a sufficiently large group of students.

Lauren Resnick points out two serious problems with matrix sampling of assessment tasks and occasions. The first is that matrix sampling does not allow for reporting of individual scores (since different individuals take different tests), and without individual accountability we implicitly reward aptitude over effort. Individual people need to see what their effort produces, and teachers and students want individual score reporting.

Second, there is the problem of mobility. A great deal of movement of students can take place over the course of a year, and this is especially true in schools that serve a lot of poor children. In a particular local system, the effect of high mobility may be that not enough students remain to measure the growth of the group accurately, and those who do remain may under-represent children of low income families. This problem will affect any attempt to measure growth locally, but it particularly complicates matrix sampling of pre and post tasks.

- **To test thinking, test items must present an appropriate degree of novelty.**

It was pointed out that the first time a student is asked to use water and measuring devices to determine which of three paper towels is most absorbent, thinking is tested. If the same student is presented the same task again, the performance may tell us more about the student's ability to remember the procedure than to invent it. If we propose that there is a general skill of designing an experiment to test a hypothesis using systematic observation and measurement, then this skill can only be tested by posing

situations which are not exactly like ones the student has encountered previously. Similarly, when a fifth grader gives the answer to 9 times 6, the most likely interpretation is that she has memorized the multiplication tables. When a student who has not been taught the multiplication tables solves the same problem, the correct answer may mean that the child has learned procedure for "figuring out" multiplication by employing successive addition, or has cleverly come upon such a procedure herself. To interpret the score, it is necessary to know something about the previous experience of the student and to uncover the thought process used by the student to solve the problem.

These examples illustrate the problem of defining the sub-domains of higher order thinking and developing or selecting measures to assess them by analyzing the subject content without regard for the previous experience of students. Consider, for example, the domain of problem solving in mathematics. Conventionally, we have measured "math problem solving" by counting the number of specific types of tasks a student can perform. Arguing from a different perspective, what we are really interested in is improving the ability of students to solve non-routine problems. In a problem solving situation, the student may be confronted with a "real world" application requiring her to identify appropriate procedures, concepts, and symbolic systems and employ them to arrive at a solution. If the problem situation requires the use of procedures or symbol systems the student has not previously learned, its intellectual demands may be great indeed. On the other hand, if the student has rehearsed very similar problems in the past, the task may be routine. The demands of the task depend upon the nature of the task but also on the developmental level, experience, and conceptual

understanding of the student. One cannot make inferences about thinking from knowledge of the test items alone; one must have knowledge of the student as well, and some evidence of how the student approached the task.

- **There is no "thinking" apart from some content: all thinking is about something. But there is some disagreement among experts about whether it is advisable to report higher order thinking skills (such as critical thinking) as outcomes apart from particular school subjects.**

Terms to describe types of thinking, such as critical thinking, analytical thinking, and problem solving occur frequently in statements of goals of curriculum reform. While there is widespread agreement that it is important to develop the skills of children to think critically and to solve problems in a variety of contexts, more information is needed about the extent to which such skills learned in one context will enable students to perform better in another. Tests of critical thinking have been developed (R. Ennis & S. Norris, 1989), but unless there is more evidence that critical thinking is a general skill that students can actually employ in a variety of situations, it may not be meaningful to measure and report it as a score. Senta Raizen argues that problem solving and hypothesis testing similarly can be measured meaningfully only in the context of particular knowledge structures; Lauren Resnick counters that the issue of labeling thinking is less important than the selection of tasks for measuring it.

Senta Raizen: Real problem solving is heavily content-dependent and knowledge-structure dependent. We can think of a student bringing a knowledge structure to a problem situation consisting of content knowledge or declarative knowledge, procedural knowledge, and strategic knowledge. If

a problem requires a student to reach outside of the declarative, procedural, and strategic knowledge he has been taught, he is likely to fail. You cannot expect to find transference of reasoning skills such as hypothesis-testing to domains outside of the learned curriculum. In science, there is little consensus about the essential curriculum before the high school level, so we cannot safely make assumptions about the knowledge structure students bring to any tested task at a particular grade. Consequently, it does not make sense to create a test of, say, "hypothesis testing" by combining students' scores on a set of diverse tasks that bear no known relationship to their previous instruction.

Lauren Resnick: School subject divisions do not define everything we want students to learn about in school. There is considerable need to break out of common school subjects. But the alternative to this isn't to create classes on "thinking." It is to assume that we have to teach thinking everywhere, in all parts of the curriculum. When we report on student performance, I don't think the labels really matter. If a school wants to know about thinking, and you score tasks using a primary trait system focused on an aspect of thinking, then it is fine to report it as such. What really matters is not the label given to the score, but the set of tasks that make up the assessment. If you measure thinking through separate little tasks, what you are going to have is teaching to those little tasks, and there is zero evidence that that transfers to anything.

- **We cannot assume that "thinking skills" employed in one context transfer readily to thinking in a different context. Consequently, no external test can be assumed to match any local curriculum appropriately unless it is truly curriculum-embedded.**

Students who have practiced writing essays may write a good persuasive essay one week and a mediocre essay on a different topic the next. Students who have been taught about electronic circuits may be able to design a procedure to determine information about circuitry inside black boxes, but their performance on this task will not correlate highly with their ability to design a procedure to determine whether sowbugs prefer light or dark. Students who have been taught about fractions in the sense of cutting up a piece of pie may not be able to solve problems involving fractions as operators, such as drawing a map to scale.

In short, once students can perform well on a task that draws directly on topics, concepts and procedures that have been instructed, we cannot assume that they can transfer this skill to tasks involving different topics, concepts and procedures in the short run, even though our long-term goal is to produce students who can think critically and solve problems involving a wide range of topics.

Other things being equal, students will perform better on tests that match closely the instructional tasks of their particular classroom. Assessing the match between classroom instruction and an external test (i.e., a test developed by anyone other than a teacher for use in her class) is not a simple task. Certainly it cannot be accomplished by comparing lists of instructional objectives and published test content analyses, especially when skills of thinking are the central concern. To infer anything about the local impact of curriculum reform from scores on an external test, one must first examine

closely the actual curriculum that was implemented and its relationship to the test.

When a test consists of only a few performance tasks, the relationship between those tasks and instruction is particularly critical: a low mean score on the assessment may mean that students have not developed sufficient familiarity with the particular contexts presented by those tasks; they might perform quite well on different tasks. Administrators might be tempted to infer that a low scoring classroom had not done an effective job of developing students' "math problem solving skills" or "critical thinking," when in fact the problem lay with the test. We cannot assume that tests claiming to measure the broad domains we classify under the rubric of higher order thinking skills are interchangeable.

- **If we use growth on a common measure to compare the impact of two different curricula, the test should include either the instructional content common to both curricula (the intersection), or the combined content of both curricula (the union).**

Just as it is important to understand in some detail the relationship between actual instruction and the content of a test in order to conclude anything about curriculum impact from the scores, it is important to understand how a single test relates to two different curricula in order to interpret differences in scores. If the test matches one curriculum better than the other, the discovery that scores are higher may provide political support but tells us nothing of interest. One solution, albeit one easier said than done, is to restrict the test to those tasks or items common to the content of both curricula. There are serious limitations to this solution, however. One is that both curricula may include a given topic, but differ greatly in the amount of

time and emphasis given to it. Another is that the comparison will tell nothing about possible trade-offs in adopting one approach over the other.

A better solution may be to compare growth using a test which combines the content of both curricula. The advantage of this approach is that it allows for a discussion of trade-offs that almost inevitably occur, and informs the discussion of conflicting values that may underlie the choice of approaches. Again, however, there are disadvantages. One is the need for a longer test. Another is the inevitable objection of teachers and students to including items on a test that do not relate closely to instructional activities. A third is the likelihood that analysts will lack the time, and audiences the patience, for this kind of analysis and discussion ever to occur. If the comparison involves only total scores, interpretation is problematic; the outcome is likely to favor the curriculum that favors breadth over depth, especially if short-answer items predominate. A thoughtful analysis must involve comparisons of subtest performance to demonstrate the relative merits of each approach.

References

Arter, J. & Davis, A. (1987, 1991). Guide to Tests of Higher Order Thinking Skills. Portland, OR: Northwest Regional Education Laboratory.

Davis, A. & Armstrong, J. (1991). State Initiatives in Assessing Science Education. In A. E. Champagne, B. Lovitts, & B. Calinger (Eds.), Assessment in the Service of Instruction, pp. 81-102. Washington, DC: American Association for the Advancement of Science.

Ennis, R. H. & Norris, S. P. (1989). Critical Thinking Testing and Other Critical Thinking Evaluation: Status, Issues, Needs. In J. Algina (Ed.) Cognitive Assessment of Language and Math Outcomes. Norwood, NJ: Abalex.

Flexer, R. (April, 1992). Alternative Assessment in Mathematics -- The Action in the States: Who's Doing What? Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Kahl, S. (April, 1992). Alternative Assessment in Mathematics: Insights from Massachusetts, Maine, Vermont, and Kentucky. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.