DOCUMENT RESUME

ED 378 243 TM 022 628

AUTHOR	Walberg, Herbert J.; And Others
TITLE	Assessment Reform: Challenges and Opportunities. Fastback 377.
INSTITUTION	Phi Delta Kappa Educational Foundation, Bloomington, Ind.
REPORT NO	ISBN-0-87367-377-8
PUB DATE	94
NOTE	43p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Accountability; *Cognitive Psychology; Cost Effectiveness; *Educational Assessment; Educational Change; Educational Innovation; Educational Policy; Educational Technology; Elementary Secondary Education; Intelligence Tests; School Restructuring; Standards; Test Construction; *Test Use
IDENTIFIERS	High Stakes Tests; *Reform Efforts

ABSTRACT

This fastback reference analyzes contrasting opinions about educational assessment and testing in the light of available evidence. The reform of student assessment is an essential component of the revitalization of American schools. Accountability issues relate to the proliferation of testing and the increasing use of high-stakes tests for policy decisions. A new focus on cognitive psychology has stimulated innovations in assessment practices. While cognitivists may attempt to go beyond behaviorally developed tests, they have yet to produce convincing and practical methods that can be easily used in classrooms. Technological developments are making tests easier to develop, administer, and score, but critical economic and technological barriers must be overcome before technology fulfills its promise in assessment. As the adequacy of current assessments is considered, three areas of debate arise: purposes of assessment, standards of technical quality, and cost. These considerations are equally important in the development of alternative assessments. Alternative assessments promise a great deal yet require sober evaluation. One figure illustrates a developed test item. (Contains 25 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

377 FASTBACK®

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

□ Minor changes have been made to improve reproduction quality

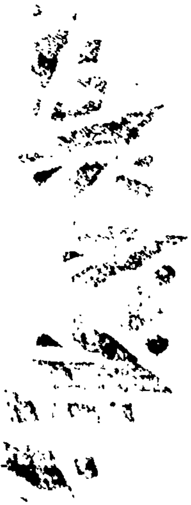
● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. BURLESON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 378 243

TM022628



Assessment Reform: Challenges and Opportunities

Herbert J. Walberg
Geneva D. Haertel
Suzanne Gerlach-Downie

PHI DELTA KAPPA EDUCATIONAL FOUNDATION

HERBERT J. WALBERG

GENEVA D. HAERTEL

SUZANNE GERLACH-DOWNIE

Herbert J. Walberg, Research Professor of Education at the University of Illinois at Chicago, serves as chairman of an advisory committee on education indicators for the Paris-based Organization for Economic Cooperation and Development and served as a founding member and chairman of the Design and Analysis Committee of the National Assessment Governing Board. He has taught educational measurement since 1961 and has served as an advisor on testing and education research in the United States and abroad.

Geneva D. Haertel is a research associate at the Temple University Center for Research in Human Development and Education and at the Western Michigan University Center for Research and Educational Accountability and Teacher Education. She is a co-editor with Herbert J. Walberg of *The International Encyclopedia of Educational Evaluation*.

Suzanne Gerlach-Downie is associate director of the Bing Laboratory Nursery School at Stanford University in California. She also provides technical assistance on research and assessment, particularly for projects involving young children.

Series Editor, Donovan R. Walling

Assessment Reform: Challenges and Opportunities

by

Herbert J. Walberg,
Geneva D. Haertel,
and
Suzanne Gerlach-Downie

Library of Congress Catalog Card Number 94-67115
ISBN 0-87367-377-8
Copyright © 1994 by the Phi Delta Kappa Educational Foundation
Bloomington, Indiana

This fastback is sponsored by the University of Ottawa and Montreal Quebec Chapters of Phi Delta Kappa, which made a generous contribution toward publication costs.

Table of Contents

Traditional vs. Alternative Assessments

New forms of assessment result from education reforms, developments in psychology, and advances in testing technology. Much of the debate about assessment reform has relied on opinion, not fact. This fastback analyzes contrasting opinions in the light of available evidence.

In this fastback the term "traditional tests" refers to standardized, norm-referenced, multiple-choice achievement tests administered using a paper-pencil format under standardized conditions. These tests are used to measure individual student performance so that students' scores can be compared. During the past decade such traditional tests have been challenged.

The term "alternative assessments" means all assessments other than traditional tests. Alternative assessments include essays, portfolios, interviews, simulations, projects, and performances. Many alternative assessments, such as assigned written compositions and portfolios of artwork, have been used for decades; thus some of them are neither new nor untested.

The term "authentic assessments" refers to assessments, especially performance assessments, that purportedly measure valuable, real-world, complex tasks. Authentic assessments often are contrasted with traditional tests and promoted as being a significant improvement over the limitations of traditional tests. In this fastback, the term "authentic assessments" is rarely used. The term "authentic" is a rhetorical device that suggests that traditional assessments are inauthentic or do not measure important knowledge or skills. The claims of the authentic assessment advocates have yet to be proved.

Influence of Education Reform

During the past decade, American public schools have been in the throes of reform. Apparent poor student performance on basic skills and knowledge tests, low levels of achievement for U.S. students compared to their international counterparts, and low rates of adult literacy have caused educators, as well as the general public, to call for reforms. The reform of student assessment is an essential component of the revitalization of American schools.

Accountability

The public values assessment data as a means to evaluate students, school systems, reform efforts, and the standing of U.S. students compared to students in other nations. Since accountability places responsibility for the success of the students on their teachers, it has become a central feature of education reform. Some reformers believe that the education system will improve only if teachers are held accountable for their students' test performance, because assessment data are the best evidence that schools are reforming. Adequate levels of achievement should be defined in terms of national standards, as well as comparative standards of progress among students. Thus some reformers argue that the quality of education will be improved only by establishing high standards of achievement and holding teachers responsible for ensuring that their students meet those standards.

ERIC
Full Text Provided by ERIC

Proliferation of Testing

Elementary and secondary students take an estimated 127 million separate standardized tests each year as a result of district and state mandates (National Commission on Testing and Public Policy 1990). During 1986-1987, approximately 105 million standardized tests were administered to 39.8 million public school students. Of these, more than 55 million were tests of achievement, competency, and basic skills administered to students in compensatory and special education programs. Some two million tests were used to screen prekindergarten and kindergarten students, and 41 million tests were administered in regular classrooms in grades 1 to 12. The General Education Development testing program, the National Assessment of Educational Progress (NAEP), and admission requirements for a variety of colleges and secondary schools accounted for an additional six million to seven million tests (Neill and Medina 1989). The National Commission on Testing and Public Policy (1990) reported that test revenues doubled between 1960 and 1966, and increased fivefold between 1967 and 1980. The revenues increased from approximately $40 million in 1960 to $100 million in 1989.

High-Stakes Testing

Whenever important consequences are attached to test results, it is considered high-stakes testing. The Scholastic Aptitude Test (SAT) and the American College Testing program (ACT) have always been high-stakes tests for college-bound students, because receiving a poor score may result in the test taker being denied admission to the college of choice School systems may suffer enrollment drops because of the importance given to test scores by some community members. Even the real estate market may be affected by the newspaper reports of local test scores and the ranking of schools and districts according to their test scores.

ERIC
Full Text Provided by ERIC

Media reporting of test scores has raised the stakes for schools and students. Teachers feel pressured to improve test scores and to cover tested material. Some districts use assessment scores to determine merit pay and dismissal decisions. Increasing the stakes of tests for teachers and administrators can exacerbate problems of overzealous test preparation and teaching to the test. Darling-Hammond (1991) listed the negative consequences of using test scores to make decisions about rewards or sanctions for schools and teachers including,

> . . . designating large numbers of low-scoring students for placement in special education so that their scores won't "count" in school reports, retaining students in grade so that their relative standing will look better on grade-equivalent scores, excluding low-scoring students from admission to "open enrollment" schools, and encouraging low-scoring students to drop out. (p. 223)

In many states test scores have risen in the first few years following the introduction of a high-stakes testing program. Whether these increased scores reflect real improvement in student achievement or only gains specific to a particular test remains to be determined. Some studies show that dropout rates increase in schools with competency tests as a graduation requirement and test-based retention policies (Madaus 1991). Students usually are motivated to do well on tests if they see a relationship between their performance on these tests and their grades or college and job prospects.

Emphasis on tests can produce desirable effects on curriculum, teaching, and learning. High-stakes tests may serve to focus instruction and highlight students' and teachers' goals. Some researchers assert that a better match between what is taught and what is tested may revitalize an obsolete curriculum

ERIC
Full Text Provided by ERIC

Influence of Psychology

Cognitive psychology challenges common views of learning, teaching, and assessment. The shift from behaviorism to cognitive psychology in the late 1950s initiated a new focus on how individuals learn, think, and acquire and apply knowledge. This new focus stimulated innovations in assessment practices.

Cognitive psychologists see learners as actively constructing knowledge structures that learners modify as their level of expertise rises. Behaviorists emphasize that higher-order understandings are the result of mastering discrete skills and prerequisite learnings. Thus behaviorists see teachers as knowledge transmitters who directly influence student learning, while cognitivists believe that teachers indirectly enhance student thought by asking questions, providing examples, giving instructions, and creating learning environments.

Behaviorists believe that complex processes, such as reading comprehension, can be broken down into a series of discrete skills. For behaviorists, tests are constructed by specifying behavioral outcomes that must be mastered for each instructional goal. In contrast, cognitivists believe that assessments should measure a wide range of thought, including knowledge, metacognitive processes, learning errors, and affective thought processes. Cognitivists measure knowledge by assessing the relationships among facts, principles, procedures, and beliefs. They measure metacognitive skills that an individual uses to appraise his or her own thinking, including the abil-

ERIC Full Text Provided by ERIC

ity to plan, activate, monitor, and evaluate actions. Affective thoughts are measured through coping and self-regulatory skills.

Merlin Wittrock (1991) expresses the concerns of cognitive psychologists:

> Many standardized intelligence tests, achievement tests, and ability tests . . . were not designed to measure diagnostically useful cognitive and affective thought processes. . . . [T]hese tests do not measure student conceptions, learning strategies, or metacognition or affective thought processes relevant to instruction. (p. 3)

Despite these claims from cognitive psychologists, many educators and other psychologists adhere to behaviorism or adopt an eclectic approach. Behaviorism is evident in mastery learning, computer-assisted instruction, and criterion-referenced testing. Cognitivists may strive to go beyond behaviorally developed tests, but they have yet to produce convincing, practical methods that can be used easily in classrooms.

Measurement of School Achievement

Four broad understandings have emerged from cognitive psychology: 1) the description of subject matter in terms of declarative, procedural, and prior knowledge; 2) the characterization of increases in knowledge along a continuum from novice to expert performance; 3) the cataloguing of learning errors specific to subject areas; and 4) the identification of metacognitive processes and learning strategies that individuals use to manage their own learning.

Declarative, Procedural, and Prior Knowledge. Students organize knowledge into schemas that are unique to the subject matter. Declarative knowledge is a network of facts and ideas. A student's ability to retrieve information efficiently is directly related to the organization of his or her declarative knowledge. According to psychologists, achievement testing in a subject area should include both estimates

ERIC
Full Text Provided by ERIC

of the amount of declarative knowledge a student possesses and how that knowledge is organized.

Traditional tests can provide a partial measurement of declarative knowledge. Items that require the student to recognize the correct answer can be used to assess the student's command of facts, principles, and vocabulary. Such tests may be less able to measure the way the student organizes information. Alternative item types, such as word associations or semantic maps, are more suited to measuring the organization of knowledge.

Procedural knowledge is knowledge of the processes and routines used in thinking. As knowledge becomes proceduralized, it becomes automatic and requires little attention by learners. The more quickly a student completes a task, the more proceduralized the knowledge and skills have become. There are no practical methods for teachers to test procedural knowledge, except through experienced observation.

Prior knowledge refers to the knowledge and skills that a student brings to the instructional setting. A student's idiosyncratic knowledge structures include not only the knowledge and skills they have acquired, but also their preconceptions, misconceptions, and beliefs. Information about a student's prior knowledge is useful when planning instruction.

The diagnosis of preconceptions, misconceptions, and beliefs can be accomplished through the use of constructed-response, alternative test items, as well as with multiple-choice items. Although facts and skills frequently are assessed, assessment of preconceptions, misconceptions, and beliefs is rarely used.

Novice and Expert Performance. Experts possess more complex knowledge structures than novices and efficiently organize their knowledge. They pay little attention to the surface characteristics of problems and carefully monitor their own problem solving. Experts generate rich problem representations as a guide for selecting solutions. Assessing expertise is easiest in subjects such as mathematics, in which the content is explicit and problem solving is well understood.

Several techniques used to document novice and expert differences include transcripts of students' solutions to problems, semantic or conceptual maps, and word associations. Semantic maps show relationships among the words and concepts that students use. Word associations involve generating word responses to a stimulus word. These assessment methods are less suitable for classroom use than for research because they require training the test administrators, administering individual assessments, transcribing transcripts, and detailed analyzing and scoring of responses.

Studies of expert performance can identify milestones that students need to master enroute to expert performance. These milestones can serve as a blueprint for test specifications. Assessments of students' subject matter expertise should consider: 1) the level of detail used to represent a problem, 2) the characteristics of the problem, 3) the conceptual skills and principles used, 4) the degree of organization and flexibility in reasoning, and 5) the selection and execution of solution strategies.

Learning Errors. Individuals make a variety of errors when solving problems in specific subjects. Some psychologists believe that errors are rule-governed. Rule-governed errors are exemplified by the systematic mistakes of elementary-age students when applying subtraction algorithms or doing place-value arithmetic. Other types of learning errors include naive theories and misconceptions. Naive theories are common prescientific beliefs that individuals hold about natural phenomena. For example, in astronomy, some students believe the sun rotates around the earth. As individuals mature and increase their knowledge of these phenomena, they shift toward more scientific conceptions.

Researchers stress error identification because it can be helpful in diagnosing learning difficulties and in developing remediation. Learning errors are more easily identified in mathematics or the sciences. In less-defined subject areas, such as the arts, compiling an inventory of learning errors is difficult.

ERIC
Full Text Provided by ERIC

Learning errors cannot be diagnosed using traditional tests. Researchers have developed alternative methods, including individual clinical interviews, semantic maps, and verbal transcripts of explanations. These individually administered assessments require large expenditures of time and money; thus there has been some interest in developing group measures of learning errors.

Metacognitive Processes and Learning Strategies. Metacognitive processes involve the self-management of thinking. These processes include planning, activating, monitoring, and evaluating one's actions. Metacognitive skills can be specific to a subject area or they can be general. Knowing that a particular strategy will enhance performance and knowing how and under what conditions to apply the strategy are metacognitive skills. Many reading programs, for example, now are designed to teach metacognitive skills.

Weinstein and Meyer (1991) identified several types of learning strategies, such as rehearsal, elaboration, and organization. Rehearsal requires the simple repetition of items in order to secure them in memory. Elaboration involves the addition of symbolic content, such as mental imagery, to increase the meaningfulness of the information to be learned. Elaboration facilitates the integration of knowledge by increasing the relationships among information in a student's knowledge structure. Organization transforms information into a format that is easier to understand. The construction of a timeline is an example of organization.

Comprehension monitoring is another metacognitive skill, which involves establishing learning goals, assessing their accomplishment, and modifying ineffective strategies. For instance, students may ask themselves questions about information in order to discover knowledge gaps. Affective strategies allow students to persist longer at difficult learning tasks and feel more effective. For example, when students schedule study sessions before an examination as a way to relieve anxiety, they are using an affective strategy

The assessment of metacognitive learning strategies cannot rely on traditional tests. In research studies, students' metacognitive learn-

ERIC Full Text Provided by ERIC

ing is exposed through structured interviews, self-report measures, observations, and occasional paper-pencil tests. In performance assessments, students provide extended oral or written responses that may reveal the metacognitive learning being used. In performance assessments, teachers can examine an essay, a science experiment, or a detailed written justification of a mathematics solution to gather evidence of the metacognitive processes a student is applying.

Because learning and study strategies affect achievement, they must be assessed separately from achievement itself. Traditional tests fail to provide information on metacognition or study practices. The Learning and Study Strategies Inventory developed by Weinstein, Schulte, and Palmer (1987) attempts to remedy these deficiencies. It contains 10 subscales, including attitude, motivation, time management, anxiety, concentration, information processing, selecting main ideas, study aides, self-testing, and test strategies. This inventory can help teachers to design optimally effective teaching and learning strategies for all students.

ERIC
Full Text Provided by ERIC

Technological Developments

Technological developments have lightened the work of psychometris' and educators by making assessments easier to develop, administer, and score. Computers can make assessment more efficient as well as create new learning environments. However, for technology to fulfill its promise, critical economic and technological barriers must be surmounted.

Test Development and Scoring

As computer capacity and speed have increased, computers have become more widely used in all aspects of testing, including managing, storing, and updating item banks. Item banks make it possible to develop customized assessments. Test items can be stored electronically by instructional objective, technical characteristics, and other categories. CD-ROM technology is being developed to store longer items for which students must construct or produce a response. Computers also have been used to generate tests using laser printers, which allow complicated drawings to be included.

The technology of "mark-sense" (or scannable) answer sheets made large-scale assessment much more feasible and made the printing of thousands of answer booklets obsolete. Optical mark-reading equipment can score more than 6,000 answer sheets in an hour.

Currently, computers can score free-response items by comparing students' responses with keyword lists or previous answers that have been sorted into correct, partially correct, and incorrect categories. Computers also have been used to score students' writing. They can provide general essay evaluations and specific suggestions for word use and sentence construction. But one difficulty in the computer scoring of essays is that the written works cannot be easily converted into machine-readable form.

Computer software can select, order, and administer test items to individual students at their convenience. These administrations usually require microcomputers and may include the use of televisions, slides, or audio recordings. Computer-based administrations do not require students to record their answers on test booklets or answer sheets. Rather, students use a keyboard, mouse, or touch-sensitive screen. In the future, students should be able to write their answers on a computer screen.

Computerized testing affords greater standardization of conditions, because the computer can present identical screens to all students. Students can take computerized tests at their own convenience and pace in public libraries or at home, using modems with "dumb terminals" or inexpensive personal computers. Computer based assessment makes it possible to administer individual mastery tests or criterion-referenced examinations to a classroom of students, each of whom is at a different level of competence. The computer selects the appropriate items and the point at which to discontinue testing for a given diagnosis, thus reducing the amount of time that the student or the teacher needs to devote to classroom testing.

Moreover, test security can be enhanced. Since passwords and encryptions are employed, no paper version of an assessment need exist. Test items can be sequenced randomly among computers to reduce the chance for students at adjacent computer stations to cheat. A final advantage is the wide range of stimuli that can be employed in the computerized presentation, including audio and video material. Videodises can store up to 54,000 still images or 30 minutes of full-motion, color, video images (Blando and Ryan 1992).

ERIC Full Text Provided by ERIC

Adaptive Testing

In computerized adaptive testing, the computer uses a student's previous answers to select subsequent items that are most suitable in terms of optimal measurement, motivation, and time savings. Computerized adaptive testing can cut testing time in half because fewer items are required for reliable assessment.

Adaptive testing can be used for diagnosis and instructional feedback; selection, placement, and certification; and accountability or system monitoring. For example, the Portland Achievement Level Testing program is a combined norm- and criterion-referenced battery employing computerized adaptive testing. The testing program serves three purposes: 1) to test students when they enter the district in order to place them in appropriate instructional programs, 2) to provide continuous assessment of the students throughout the school year, and 3) to select students for placement in special programs at any point during their enrollment. In addition, a version of the computerized adaptive test has been used for such accountability functions as the evaluation of compensatory education programs (U.S. Congress, Office of Technology Assessment 1992).

Integrated Learning Systems

Four decades ago, Ralph Tyler pointed out that "Measurement [should be] conceived, not as a process quite apart from instruction, but rather as an integral part of it" (1951, p. 47). Integrated learning systems (ILS) are computer systems that permit an individual student's test results to guide instruction. Because computers can store large numbers of items and rapidly calculate estimates of a student's ability following the administration of each item, shorter tests that are individually suited to the student and provide nearly instantaneous feedback also may enhance motivation and utility.

ILS technology is guided by two aspects of curriculum. One is instructional experiences that move students through the domain of

ERIC
Full Text Provided by ERIC

content to accomplish educational goals. The other is the set of course standards that serve as milestones of beginning, intermediate, or terminal accomplishments. ILS make use of instructional activities that move students along a path of expertise marked by testing milestones. Thus ILS are able to provide continuous analysis, diagnosis, and monitoring of student learning.

ILS items can be presented as part of the instructional process. The successive screens displayed on the computer provide presentations, checks for understanding, practice, coaching, and feedback. At the request of the student or teacher, a "mastery map" can be displayed on the monitor that shows what the student has accomplished and what standards are yet to be completed. The standards can reflect student, teacher, or district goals. ILS may include the following features: displays of student progress and options, directed practice on tasks, on-line prototype answers to assessments, cumulative archives of individual student data, computer-guided coaching, predictions of learning rates to guide review, and minute-by-minute analyses of classroom and group performance to aid in classroom, school, and district instructional management. Over the past 30 years, ILS developed by vendors such as WICAT, Computer Curriculum Corporation, and Jostens have been implemented in a variety of school districts.

Conventional computer-assisted instruction programs, an early application of ILS, have been studied carefully. Research syntheses show that their effects on specific, short-term learning outcomes are greater than those of conventional teaching (Niemiec and Walberg 1987). However, the data on more advanced ILS applications using general, long term educational outcomes have been less convincing.

Intelligent Measurement

The use of knowledge bases and inferencing procedures permits computer systems to produce "intelligent measurement" (Bunderson, Inouye, and Olsen 1989). Intelligent measurement requires a knowledge base that contains expertise specific to a subject area. Three types of intelligent measurement are applicable to education. The first type provides prescriptive advice, or intelligent interpretations. An expert system can model the knowledge of a teacher who is familiar with the subject area, the instructional management system, and the curriculum. The expert system can model good pedagogy, match instruction with characteristics of learners, and generate trajectories of student progress.

A second educational application is automatic holistic scoring. Knowledge bases used in automatic holistic scoring represent mastery standards for the assessment tasks and the scoring knowledge of experts. In automatic holistic scoring, an expert system performs the complex scoring of assessments, replicating the judgments made by human scorers.

A third application of intelligent measurement is the automation of individual profile interpretations. The output of the computer includes: 1) questions for counselors to ask students in order to clarify the students' performance and 2) interpretative commentaries. The automation of individual profile interpretations reduces the need for each teacher to be an expert in interpreting assessment results.

Adequacy of Current Assessments

As assessment has increased in importance, the technical merit of individual assessments has become more critical. Technical merit may be evaluated using, for example, the *Standards for Educational and Psychological Tests* (1985) or the *ETS Standards for Quality and Fairness* (1987). Because there is no consensus about what constitutes appropriate technical standards for alternative assessments, determining their merit is more difficult than determining the merit of traditional assessments. For example, how does one determine the merit of a geometry assessment that promotes enthusiasm by allowing students to use their artistic talents in answering questions but that does not meet traditional technical standards of validity and reliability? Three areas of debate have emerged: purposes of assessment, standards of technical quality, and costs.

Purposes of Assessment

Assessments should fulfill at least one of three purposes: 1) the monitoring of individual student progress and the diagnosis of learning difficulties, 2) the placement and certification of individual students, and 3) the evaluation and/comparison of groups to ensure the accountability of the education system (Resnick and Resnick 1989).

Monitoring and Diagnosis of Student Learning. Monitoring individual student progress and diagnosing learning difficulties aid teachers

ERIC Full Text Provided by ERIC

in classroom management. Monitoring student progress usually is accomplished with teacher-made tests that help teachers determine instruction. Some teachers criticize standardized achievement tests as lacking the capacity to provide diagnostic information for the enhancement of day-to-day instruction.

Traditional tests have been criticized for their dependence on recognition items, their limited coverage of domains of knowledge, and their alleged failure to elicit a range of higher-order thought processes. Traditional multiple-choice achievement tests, it is argued, place too much emphasis on facts and procedures for solving well-structured problems that are presented without context. In addition, these traditional tests are limited in their utility to identify the characteristics of a student's learning. Of course, traditional tests were never intended to do all these things. They were intended as complements to such teacher assessments as essays and laboratory assignments, which can accomplish these things.

Critics of multiple-choice items argue that such items are easier because they require only that the student recognize a correct answer. This is in contrast to fill-in-the-blank or essay items, which require a student to recall appropriate information and to generate a response. In addition, multiple-choice items sometimes can be answered correctly by guessing.

According to Ward, Rock, and La Hart (1990), traditional and alternative item formats can be arranged along a continuum based on several dimensions: 1) selection/identification, 2) reordering/rearrangement, 3) substitution/correction, 4) completion, 5) construction, and 6) presentation/performance. For example, multiple-choice items are considered the most constrained, because they do not require a student to generate a response. Rather, the student selects an answer from those presented. Proponents of alternative assessments assert that the items used in alternative assessments are less constrained and can be used to measure more realistic and complex problem-solving than can multiple-choice test items. An example of such an alternative is shown in Figure 1.

ERIC Full Text Provided by ERIC

Draw a line connecting the sun, the cat, the plants, and the mice to show the direction in which energy travels through the food web.

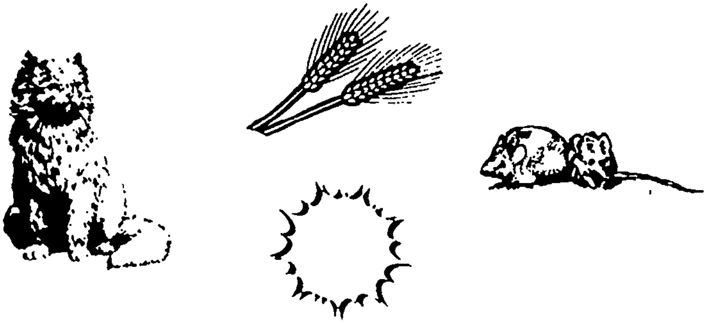


Figure 1. A less-constrained item from an alternative assessment.

Proponents of performance-based alternative assessments emphasize the complexity and authenticity of less-constrained items. They argue that performance assessments derive their value from examining actual performances, rather than from examining indicators of potential performances, as occur with traditional tests. Following is a form used to score a complex performance-based assessment that currently is being used on a small scale. In this assessment, "Students used a laboratory setup to determine which of three paper towels held the most and least water" (Shavelson and Baxter 1992, p. 21).

Paper Towels Investigation: Hands-on Score Form*

1. **Method**

A. Container	B. Tray (surface)
Pour water in/put towel in	Towel on tray/pour water on
Put towel in/pour water in	Pour water on tray/wipe up

*Adapted from Shavelson, R.J., and Baxter, G.P. "What We've Learned About Assessing Hands-on Science." *Educational Leadership* 49, no. 8 (1992): 21.

ERIC Full Text Provided by ERIC

2. **Saturation** A. Yes B. No

3. **Determine Result**
   A. Weigh towel
   B. Squeeze towel/measure water (weight or volume)
   C. Measure water in/out
   D. Time to soak up water
   E. No measurement
   F. Count # drops until saturated
   G. See how far drops spread out
   H. Other ____________

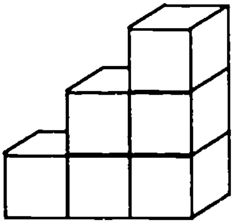
4. **Care in Measuring** Yes No

5. **Correct Result** Yes No

Many researchers and educators are wary of claims that such alternative assessments can replace multiple-choice tests for monitoring student learning. Many valued educational tasks require simple recognition, and some skills (including higher-order thinking) and subject areas can be competently assessed using a multiple-choice format. Figure 2 shows a multiple-choice item that tests higher thought processes.

You are building a staircase out of cubes:

1 step = 1 cube
2 steps = 3 cubes
3 steps = 6 cubes



How many cubes does it take to build a staircase that is 6 steps high?

a. 36 cubes
b. 28 cubes
c. 21 cubes
d. 15 cubes

From *A Sampler of Mathematics Assessment*, California Department of Education, 1991, p. 46. Reprinted with permission.

Figure 2. An example of an enhanced multiple-choice item that measures higher-order thought processes.

ERIC
Full Text Provided by ERIC

Some advocates argue that alternative assessments provide a multidimensional view of a particular skill or content area. Yet breadth of coverage often is traded for depth of coverage. Performance assessments are based on one or a small number of tasks and thus may assess only a limited sample of what a student knows compared to the dozens of facts and ideas than can be assessed using multiple-choice items.

Researchers have examined the equivalence of multiple-choice and alternative assessments in various subject areas. In this research, the knowledge measured and the scores assigned using multiple-choice and alternative items – in particular, open-ended items – are very similar. If the knowledge and scores assigned are the same, then the capacity of these two types of assessments to diagnose learning difficulties is equivalent. Thus alternative assessments, which may be costly and may lack technical standards, have yet to demonstrate more value than teacher-made and traditional multiple-choice tests.

Certification and Placement of Students. A second purpose of testing is to make decisions about placement in instructional programs and certification of mastery of a content area. Tests used for this purpose do not inform the management of daily classroom activities but are used to make administrative decisions about a student's progress through the school system.

Tests used for making high-stakes decisions, including placement and certification, must meet high technical standards that warrant their use as the single piece of evidence for making decisions about a student's future. Of course, better decisions are made when multiple sources of evidence are used.

The use of traditional and alternative assessments for the placement and certification of a diverse student body has been challenged with evidence of racial, ethnic, and gender differences in performance on these tests. Some critics contend that traditional tests are inherently biased and produce an adverse impact on some groups of students, and thus should play a minor role, if any, in high-stakes decisions.

But cautions regarding technical quality, equity, bias, and adverse impact pertain to all types of tests and assessments used for high-stakes decisions, including alternative assessments.

Some researchers indicate that some alternative forms of assessment, at least initially, widen the performance gap between males and females and between socioeconomic and ethnic groups. In contrast, others cite evidence that the written essay part of advanced placement exams in various subject areas produce smaller gender differences than do the multiple-choice parts of these exams. Less information is available about the reliability and validity of alternative forms of assessment. Again, there is the question of proof. Simply criticizing multiple-choice items or misuse of traditional tests hardly makes an affirmative case for alternative assessments.

Comparison of Groups for Accountability. The third purpose of assessment is the comparison of groups of students in order to evaluate schools, programs, states, and nations, and thus maintain accountability. Typically, multiple-choice tests have been used for this purpose. Recently, alternative assessments using constructed-response items have been considered for use in large-scale testing programs.

Improved accountability requires not only accurate comparisons among groups but also measures of student performance on content in which students have received instruction. School district goals, instructional materials, methods, curricula, and assessments often are poorly aligned or integrated. School districts develop local goals and curricula but depend on commercial textbooks and standardized tests for instruction and assessment. Because these commercial textbooks and tests are prepared for national use, they rarely reflect all the local educational priorities.

This mismatch may lead to inefficiencies and morale problems. Teachers may use instructional materials that fail to reflect district goals, even though they will be evaluated in part on their students' attainment of those goals. Students may be examined on knowledge and skills they have not studied, or they may study content that is not considered a priority in their community but on which they will be tested. These mismatches have been identified as a cause of poor educational productivity in the United States. Poor alignment of instruction, materials, and tests cannot be eliminated simply by using alternative assessments, although the local development of assessments can ensure more agreement among the elements of instruction and their assessment. However, local assessments cannot serve the purposes of comparing districts, states, or nations.

The National Assessment of Educational Progress (NAEP), a congressionally initiated survey of educational achievement, is a testing program used for monitoring student performance at the national and state level. Since 1969, NAEP has collected assessment data in reading, mathematics, science, writing, history/geography, and other fields. NAEP draws on a representative sample of schools that participate in the assessments. In 1990, for the first time, NAEP conducted state-by-state comparisons as part of the mathematics assessment. Advocates of NAEP believe that these comparisons hold educators accountable for their students' performance over time and also for the level of performance their students display in comparison to similar students nationwide.

Recently, both major political parties advocated the establishment of a national examination system to monitor the nation's schools. Referred to as America 2000 during the Bush Administration and Goals 2000 in the Clinton Administration, this examination system proposes "world-class standards" in English, mathematics, history, science, and geography (U.S. Department of Education 1991).

Standards of Technical Quality

Reliability and validity are two key psychometric concepts that are applied to the evaluation of tests and assessments.

Reliability is the consistency with which a test measures content. One type estimates the consistency of the test to measure the same individual's performance on several occasions (test-retest reliability).

Other types of reliability establish the equivalence of several forms of a test (parallel and alternate-forms reliability). Still other types of reliability establish whether each item on a test measures the same content (split-half or alpha reliability). Finally, inter-rater reliability estimates the consistency with which raters assign scores to an individual's performance.

The reliability of traditional multiple-choice tests is well documented. In contrast, little information is available on the reliability of many alterpative assessments. The reliability information that is available tends to focus on inter-rater reliability of performance-based alternative assessments. Preliminary evidence shows that expert raters often lack consensus on their assessments of written essays, laboratory exercises, and other alternative tasks.

Vermont initiated the first statewide assessment to measure student achievement using portfolios, one of the more popular forms of alternative assessment. This statewide assessment is one of the few alternative assessment projects being dispassionately evaluated by an external evaluator. A recent article describing the assessment (Viadero 1993) revealed the difficulty of es.ablishing adequate reliability when using alternative test formats: "A 1992 report by the RAND Corporation . . . finds that the 'rater reliability' in scoring the portfolios . . . was very low" (p. 18). Because of low rater reliability, the results were not reported at the school or district level.

Validity refers to whether a test measures what it is claimed to measure. Recently, psychometricians Linn, Baker, and Dunbar (1991) developed an expanded definition of validity that applies to alternative and traditional assessments. In their view, the evaluation of validity for all forms of assessment should, at the minimum, include evidence regarding directness and transparency, consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, and meaningfulness.

Directness refers to the extent to which the assessment task matches the instructional goals. An example of the direct assessment of writing is when students are asked to produce a writing sample. In contrast, an indirect assessment of writing skill requires students to answer multiple-choice questions about correct punctuation or stylistic considerations. *Transparency* refers to the clarity of the criteria used in judging performances. Assessments with high transparency have high acceptability and are viewed as legitimate measures. Directness and transparency can be viewed as components of *face validity*.

Fairness requires the identification of potential sources of bias, such as rater effects or insensitive or irrelevant materials. Bias sometimes can be detected using statistics that identify items on which groups of test takers perfor.n differently. These differences may not reflect true differences in test takers' knowledge but, rather, differences in their cultural experiences. Adverse impact on identified groups of students must be considered when judging the fairness of an assessment. An assessment should not result in members of any racial, gender, or ethnic group being evaluated differentially, assuming all groups of students are equally qualified. In addition, some students may need to be taught how to take tests more effectively in order to ensure a fair evaluation of their knowledge.

Critics of traditional assessments question the degree to which successful performance on traditional assessments transfers to real-world activities. Performance assessments are believed to have increased transferability to non-academic tasks.

However, performance assessments present special problems in terms of their generalizability. Because developers of performance assessments create tasks that are held to be realistic, complex, and contextualized, the assessment tasks require more time than traditional tests. As a result, fewer tasks can be administered. Thus such assessments provide fewer incidences of student behavior and a limited sample of student knowledge and skills.

When an assessment requires the test taker to use several abilities, as opposed to a simple, less developmentally advanced way to solve problems, it is considered cognitively complex. Complexity should

ERIC Full Text Provided by ERIC

be determined by analyzing the types of skills and processes students use to answer questions. Students can correctly answer items using processes and strategies other than those expected by the test developers. Thus items that were designed to assess students' higher thought processes may be solved using less-advanced approaches, or vice versa.

Judging the quality of an assessment should include a review of its content. Adequate content coverage should express the breadth and depth of the subject. Subject matter experts should systematically determine whether the assessment adequately covers current ideas and material of long-standing importance. This type of review is particularly important in the case of performance assessments that sample only a limited aspect of a subject area.

Whether students and teachers perceive assessment problems as meaningful affects their motivation and performance. When assessments are meaningful to students, their content is relevant to the students' experiences. Advocates of performance assessment believe that assessment can be meaningful learning. By this criterion, however, life in classrooms can never be as "authentic" as that outside.

Costs of Assessments

Beyond consideration of the money spent on all types of assessments, educators increasingly are concerned about the time students spend preparing for and taking tests and the time teachers spend preparing, administering, and scoring tests.

Student Time Spent on Testing. Based on a national survey, researchers Dorre-Bremme and Herman (1986) concluded that only modest amounts of student time are devoted to testing. At the elementary level, total testing time, in all subjects, averaged 76 hours a year, or 8.6% of the total class time of students. Elementary students took a test in reading and a test in math about once every eight days. High school students spent about 12% of their time taking tests in English and mathematics classes. A typical 10th-grader spent nearly 26½ hours

ERIC Full Text Provided by ERIC

annually completing tests in English and 24 hours annually completing tests in mathematics. A high school student took an English test and a mathematics test every three or four days.

Dorre-Bremme and Herman also found that both high school and elementary students spent the largest percentage of their testing time on teacher-developed tests and the next-largest percentage on tests included with curriculum materials. In contrast, minimum competency testing, on average, consumed a very small percentage of testing time. State- and district-mandated tests took about 25% of high school students' total testing time.

Teacher Time Spent on Testing. According to Dorre-Bremme and Herman, for each hour a student spent taking a test, a teacher spent two to three hours preparing for the test, grading the test, and recording students' scores. Interviews with elementary teachers indicated that they spent about 12% to 15% of their work time, both in and out of school, on achievement testing in all subject areas. This averages to about 200 to 250 hours throughout a school year. Similar figures were not available for high school teachers, but the researchers claimed that high school teachers spent about two hours outside the class for every hour of student testing.

The amount of time teachers devote to alternative assessments also has been a subject of debate but has not been well researched. According to one report, teachers in Great Britain, who have heavily relied on alternative assessments in the past few years, are displeased with the time commitment that such tests require.

Although alternative assessments may require more teacher time for development of the assessments and the training in their use, such time can be viewed as a benefit rather than a cost. For example, teachers involved in developing and scoring the California Assessment Program report that these processes are the most effective staff development activity in which they have participated (Carlson 1991).

Costs. The three basic costs incurred when conducting traditional or alternative assessments are: 1) money costs, 2) non-money costs, and 3) estimated opportunity costs. Money costs are the dollars spent on development, administration, scoring, and reporting results. Although estimates vary on the exact money costs of traditional tests and performance assessments, experts estimate performance items to be much more expensive. According to the U.S. Congress, Office of Technology Assessment:

> The costs of performance assessment represent a substantial barrier to expanded use. Performance assessment is a labor-intensive and therefore costly alternative unless it is integrated in the instructional process. Essays and other performance tasks may cost less to develop than do multiple choice items, but are very costly to score. One estimate puts scoring a writing assessment as 5 to 10 times more expensive as scoring a multiple choice examination, while another estimate based on a review of several testing programs administered by ETS . . . suggests that the cost of assessment via one 20- to 40-minute essay is between 3 to 5 times higher than assessment by means of a test of 150 to 200 machine scored, multiple choice items. Among the factors that influence scoring costs are the length of time students are given to complete the essay, the number of readers scoring each essay, qualifications and location of readers (which affect how much they are paid, and travel and lodging costs for the scoring process), and the amount of pretesting conducted on each prompt or question. The higher these factors, the higher the ratio of essay to multiple choice costs. (1992, p. 243)

Non money costs for traditional and alternative assessments include expenditures of employee time, materials, equipment, space, and energy. Other non money costs may be stress and a decrease in morale for students, teachers, and administrators. The enthusiasm produced by some of the hands on activities used in alternative assessments must be weighed against the expenditures of time required to administer and score performance assessments. Traditional tests often are met with less enthusiasm by teachers and students but require a more modest expenditure of time, materials, and space.

ERIC
Full Text Provided by ERIC

Opportunity costs require educators to consider what is displaced for students and teachers when a testing program is implemented. When resources of time, money, and energy are invested in an assessment program, they are unavailable for other uses. For example, the time spent by teachers on administering and scoring assessments should be weighed against the time that could have been used for lesson planning, tailoring instruction to individual students, or upgrading teachers' content knowledge and pedagogical skills.

A second example of opportunity costs is provided by comparing the costs and benefits of using different types of assessments. Some educators argue that alternative assessments provide better data for diagnosing and remediating learning difficulties than do traditional tests. However, the opportunity costs of improved diagnostic information may include a loss of instructional time for students or planning time for teachers, and a reduction in the budget due to the expense of the alternative assessment. Alternative assessments may require more resources on the part of the education system than their promised benefits warrant.

Conclusion

Assessment is integral to the educational process. It serves three fundamental purposes: the day-to-day management of instruction, the classification and placement of students, and the maintenance of accountability for educators and students. Because of these fundamental uses, assessment has become a primary tool for the reform of education. In the past decade, educators have argued over the purposes, format, technical adequacy, and costs of assessment. New assessments are emerging from these debates. Some employ new item formats; others make use of computer-based technologies.

Psychological research is another influence on assessment. Psychologists have argued for assessments that measure students' knowledge schemas, pathways to expertise, and metacognitive learning and study strategies. However, leading education researchers have cautioned against hasty applications of cognitive psychology. Richard Snow and David Lohman assert, "Cognitive psychology has no ready answers for the measurement problems of yesterday, today or tomorrow" (1989, p. 320).

Alternative assessments, particularly those that are referred to as "authentic" because of their reliance on complex, real life tasks, are viewed by some as a remedy for the misuse of traditional testing. Alternative assessments are regarded as having high face validity and close curricular and test alignment. Other advocates of alternative assessments see these tests as the best way to measure subject matter expertise. They believe that expertise is better demonstrated in assessments that require extended performances and go beyond recognition items. Despite the supposed benefits attached to alternative assessments, there is little evidence of their wide-scale feasibility, practicality, and utility.

When the purpose of assessment is monitoring the educational standing of school districts, then traditional tests may be the assessment method of choice. Standardization and norming are necessary when comparisons among groups of students are to be made. The larger the pools of students being compared, the more important it is that the assessment procedure be affordable, objective, standardized, and easy to administer and score. These criteria are not easily met by many alternative assessments. Multiple-choice tests can serve the purpose of accountability and, with enhancement, can measure higher thought processes. When selecting an assessment, educators must be attentive to the trade-offs in cognitive sensitivity, technical adequacy, costs, and ability to fulfill the assessment purposes.

Alternative assessments promise much, but they require sober evaluation. There is little information about the technical characteristics of many new forms of assessment. Evidence of difficulties in the use of performance assessment – one form of alternative assessment – has surfaced. For example, Alan Purves, director of the international writing assessment, recently expressed disenchantment over the inability to establish comparable ratings among judges (Rothman 1990). This problem plagues not only writing assessments, but performance assessments of other subjects as well.

Studies by psychologist Richard Shavelson (1991) also cast doubt on the viability of a performance assessment as a sole tool for assigning grades to students. Based on his project, which develops science and mathematics performance assessments, he reported that large differences in a student's scores can occur depending on which performance assessment task is administered. In other words, different performance assessments that attempt to measure the same content do not rank students in the same order.

Computer-based technology promises to make assessment efficient and has demonstrated some impressive results. Optical mark-reading equipment, mark-sense answer sheets, microcomputers, hypermedia, artificial intelligence, and other applications have advanced assessment practices. Technologists claim that computers will be able to score complex, constructed responses; maintain cumulative mastery maps of student progress; present multimedia simulations; videotape student performances for further analysis; and train teachers on the administration, scoring, and interpretation of assessment results. Many of these components have been demonstrated separately. What is lacking, as yet, are large-scale systems that integrate a substantial part of the K-12 curriculum and instructional programs.

In the past, several computer-based innovations have been heralded as cure alls. Although the feasibility of such technologies was demonstrated in university laboratories and military and business environments, their effectiveness in school settings was less well-documented. Presumably, with national, or at least widely shared, goals that value technology for education, new technologies may become feasible for the nation's schools. Nevertheless, it is unlikely that computer-based technologies will be a panacea to our assessment ills in the immediate future. As in the case of all forms of assessment, open-mindedness and healthy skepticism are in order.

ERIC
Full Text Provided by ERIC

References

Blando, J.A., and Ryan, P. "Student Assessment: The Role of New Technologies." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992.

Bunderson, C.V.; Inouye, D.K.; and Olsen, J.B. "The Four Generations of Computerized Educational Measurement." In *Educational Measurement*, 3rd ed., edited by R.L. Linn New York: Macmillan, 1989.

Carlson, D. "Changing the Face of Testing in California." *California Curriculum News Report* 16 (January-February 1991): 1, 10.

Committee to Develop Standards for Educational and Psychological Testing. *Standards for Educational and Psychological Tests*. Washington, D.C.: American Psychological Association, 1985.

Darling-Hammond, L. "The Implications of Testing Policy for Quality and Equality." *Phi Delta Kappan* 73 (November 1991): 220-25.

Dorre-Bremme, D.W., and Herman, J.L. *Assessing Student Achievement: A Profile of Classroom Practices*. CSE Monograph Series in Evaluation No. 11. Los Angeles: University of California, Center for the Study of Evaluation, 1986.

Educational Testing Service. *ETS Standards for Quality and Fairness*. Princeton, N.J., 1987.

Linn, R.; Baker, E.; and Dunbar, S. "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20, no. 8 (1991): 15-21.

Madaus, G. "The Effects of Important Tests on Students: Implications for a National Examination System." *Phi Delta Kappan* 73 (November 1991): 226-31.

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

Neill, D.M., and Medina, N.J. "Standardized Testing: Harmful to Educational Health." *Phi Delta Kappan* 70 (May 1989): 688-97.

Niemiec, R.P., and Walberg, H.J. "Comparative Effects of Computer-Assisted Instruction: A Synthesis of Reviews." *Journal of Educational Computing Research* 3, no. 1 (1987): 19-37.

Resnick, L.B., and Resnick, D.F. "Tests as Standards of Achievement in School." In *Proceedings of the 1989 ETS Invitational Conference: The Uses of Standardized Tests in American Education.* Princeton, N.J.: Educational Testing Service, 1989.

Rothman, R. "New Tests Based on Performance Raise Questions." *Education Week*, 12 September 1990, pp. 1, 10, 12.

Shavelson, R.J. "Authentic Assessment: The Rhetoric and Reality." Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1991.

Shavelson, R.J., and Baxter, G.P. "What We've Learned About Assessing Hands-on Science." *Educational Leadership* 49, no. 8 (1992): 20-25.

Snow, R.E., and Lohman, D.F. "Implications of Cognitive Psychology for Educational Measurement." In *Educational Measurement*, 3rd ed., edited by R.L. Linn. New York: Macmillan, 1989.

Tyler, R.W. "The Functions of Measurement in Improving Instruction." In *Educational Measurement*, edited by E.F. Lindquist. Washington, D.C.: American Council on Education, 1951.

U.S. Congress, Office of Technology Assessment. *Testing in American Schools: Asking the Right Questions.* Report No. OTA-SET-519. Washington, D.C.: U.S. Government Printing Office, 1992.

U.S. Department of Education. *America 2000: An Education Strategy: Sourcebook.* Washington, D.C.: U.S. Office of Education, 1991.

Viadero, D. "RAND Urges Overhaul in Vt.'s Pioneering Writing Test." *Education Week*, 10 November 1993, pp. 1, 18.

Ward, W.C., Rock, D.A., and La Hart, C.L. *Toward a Framework for Constructed Response Items.* Report No. RR 90-7. Princeton, N.J.: Educational Testing Service, 1990.

Weinstein, C.E., and Meyer, D.K. *Implications of Cognitive Psychology for Testing: Contribution from Work in Learning Strategies.* Englewood Cliffs, N.J.: Prentice Hall, 1991.

Weinstein, C.E.; Schulte, A.C.; and Palmer, D.R. *The Learning and Study Strategies Inventory (LASSI)*. Clearwater, Fla.: H&H Publishing, 1987.

Wittrock, M.C. "Cognition and Testing." In *Testing and Cognition*, edited by M.C. Wittrock and E.L. Baker. Englewood Cliffs, N.J.: Prentice-Hall, 1991.

Phi Delta Kappa Fastbacks

Two annual series, published each spring and fall, offer fastbacks on a wide range of educational topics. Each fastback is intended to be a focused, authoritative treatment of a topic of current interest to educators and other readers. Several hundred fastbacks have been published since the program began in 1972, many of which are still in print. Among the topics are:

Administration
Adult Education
The Arts
At-Risk Students
Careers
Censorship
Community Involvement
Computers
Curriculum
Decision Making
Dropout Prevention
Foreign Study
Gifted and Talented
Legal Issues
Mainstreaming
Multiculturalism
Nutrition
Parent Involvement
School Choice
School Safety
Special Education
Staff Development
Teacher Training
Teaching Methods
Urban Education
Values
Vocational Education
Writing

For a current listing of available fastbacks and other publications of the Educational Foundation, please contact Phi Delta Kappa, 408 N. Union, P.O. Box 789, Bloomington, IN 47402-0789, or (812) 339-1156.

ERIC
Full Text Provided by ERIC

Phi Delta Kappa Educational Foundation

The Phi Delta Kappa Educational Foundation was established on 13 October 1966 with the signing, by Dr. George H. Reavis, of the irrevocable trust agreement creating the Phi Delta Kappa Educational Foundation Trust.

George H. Reavis (1883-1970) entered the education profession after graduating from Warrensburg Missouri State Teachers College in 1906 and the University of Missouri in 1911. He went on to earn an M.A. and a Ph.D. at Columbia University. Dr. Reavis served as assistant superintendent of schools in Maryland and dean of the College of Arts and Sciences and the School of Education at the University of Pittsburgh. In 1929 he was appointed director of instruction for the Ohio State Department of Education. But it was as assistant superintendent for curriculum and instruction in the Cincinnati public schools (1939-48) that he rose to national prominence.

Dr. Reavis' dream for the Educational Foundation was to make it possible for seasoned educators to write and publish the wisdom they had acquired over a lifetime of professional activity. He wanted educators and the general public to "better understand (1) the nature of the educative process and (2) the relation of education to human welfare."

The Phi Delta Kappa fastbacks were begun in 1972. These publications, along with monographs and books on a wide range of topics related to education, are the realization of that dream.