DOCUMENT RESUME

ED 275 696 TM 860 571

TITLE	The Redesign of Testing for the 21st Century: Proceedings of the 1985 ETS Invitational Conference (46th, New York, New York, October 26, 1985).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	86
NOTE	107p.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Achievement Tests; Cognitive Psychology; Computer Assisted Testing; Educational Research; Educational Technology; *Educational Testing; *Educational Trends; Elementary Secondary Education; *Futures (of Society); Higher Education; *Measurement Objectives; Psychological Studies; *Research Utilization; Testing Problems; *Test Use
IDENTIFIERS	Pittsburgh School District PA

ABSTRACT

Future issues in educational and occupational testing were discussed at the 1985 Educational Testing Service (ETS) Invitational Conference. Gregory R. Anrig, ETS President, predicted that advances in cognitive psychology and technology would influence tests to serve individuals more than institutions, to help individuals learn and succeed, and to guide instruction on a continuing basis. The ETS Award for Distinguished Service to Measurement was presented to Paul Horst for his work in differential prediction, and for his contributions to psychological measurement, factor analysis, and the Psychometric Society. Nine papers were presented: (1) Changing Schools and Testing: An Uneasy Proposal by Theodore R. Sizer; (2) Cognitive Research and Future Test Design by Earl Hunt; (3) Measurement Research That Will Change Test Design for the Future by William C. Ward; (4) Technology Advances That May Change Test Design for the Future by Dorothy K. Deringer; (5) The Integration of Instruction and Testing by Robert Glaser; (6) Redirecting a School District Based on the Measurement of Learning Through Examinations by Richard C. Wallace, Jr.; (7) Barriers to New Test Designs by Robert L. Linn; (8) Technological Literacy as Means and Ends by Edward A. Friedman; and (9) The Perils and Promises of New Tests and New Technologies: Dick and Jane and the Great Analytical Engine? by George F. Madaus. (GDC)

The Redesign of Testing for the 21st Century

INVITATIONAL CONFERENCE PROCEEDINGS

Educational Testing Service

The Redesign of Testing for the 21st Century

Proceedings of the
1985 ETS *Invitational Conference*

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

The forty-sixth ETS Invitational Conference, sponsored by Educational Testing Service, was held at The Plaza, New York City, on October 26, 1985

Presiding: Gregory R. Anrig
President
Educational Testing Service

Conference Coordinator: Margaret B. Lamb
Editor, Conference Proceedings: Eileen E. Freeman
Production Coordinator: Joyce M. Hofstetter

Educational Testing Service is an Equ.'l Opportunity/Affirmative Action Employer.

Library of Congress Catalog Number: ISSN 0161-3758
Printed in the United States of America.

Contents

Introduction

The annual Invitational Conference sponsored by Educational Testing Service is designed to serve and to enhance the knowledge of those concerned with measurement in educational and occupational fields. In recent years, the focus of the conference has been on issues of particular priority at ETS. In 1983, the subject was the promotion of proper test use. Last year, we discussed "Equity, Access and Excellence" and focused on minority-related issues.

This year, however, our focus is a long-range one--"The Redesign of Testing for the 21st Century." In choosing this topic, we hope to raise issues that will see fruition in the years to come.

As a not-for-profit institution, ETS is founded on the principles of advancing educational measurement and serving education generally. For nearly 40 years, ETS has been in the forefront of psychometric research and development. Throughout that period, it has provided essential services to the educational community and has developed new testing instruments that have greatly improved the way we measure educational growth.

Now, however, it is time for ETS to look to the future, to make a new commitment to research and to service. No organization can succeed without a vision, a dream, a goal to be reached.

That is why I am pleased to announce today the initiation of Project Jessica, a long-term research, development, and application effort to create a new generation of testing and measurement services for the future. In order to support Project Jessica, the ETS Board of Trustees has established a special $30 million development fund.

Current forms of standardized testing serve important accountability and institutional needs. These needs will continue to exist in the future, as will the current array of achievement, admissions, and licensing tests.

Advances in cognitive psychology and technology, however, make possible new kinds of measurement instruments. This new generation of tests will have three functions: (1) it will serve individuals more than institutions; (2) it will aim primarily at helping individuals learn and succeed rather than simply yielding scores for institutional decisionmaking; (3) it will guide instruction and self-development on a continuing basis rather than compare performance among test takers.

This new generation of tests will be *helping* measures, enabling individuals to keep pace with rising standards in education and the workplace.

They will capitalize upon electronic technology for their development, design, and delivery.

We have named our new project after an actual four-year-old girl. Jessica reminds us that this initiative is to create new opportunities for children already born who will live most of their adulthood in the 21st century. We are committed to creating a new generation of testing measures to help all Jessicas with their personal, educational, and career development.

As you read the ideas presented by our invited experts, I urge you to keep Jessica in mind. Time is fleeting. Jessica is already going on five—and the 21st century is less than fifteen years away.

Gregory R. Anrig, *President*
Educational Testing Service

The 1985 ETS Award for Distinguished Service to Measurement

Presented to:

Paul Horst

The central theme dominating Paul Horst's long and productive career is the maximal utilization of human resources. Over the years, he has diligently pursued this concern through both theoretical formulations and practical applications. An important early contribution in this regard was his landmark 1941 monograph on *The Prediction of Personal Adjustment*, which remains an illuminating sourcebook to this day. He moved on to develop definitive quantitative techniques for differential predication, as well as multiple absolute predication, and for determining optimal test length for maximum battery validity, for maximum differential prediction, and for multiple prediction in general. On the applied side, Professor Horst's continuing commitment to optimal use of human resources reached fruition in his successful implementation of a multiple differential prediction program as a cooperative enterprise between the high schools and the postsecondary institutions in the state of Washington.

Professor Horst has also written four influential textbooks—*Matrix Algebra for Social Scientists, Factor Analysis of Data Matrices, Psychological Measurement and Prediction,* and *Personality: Measurement of Dimensions.* Their rigor and precision of expression have enhanced the quality of teaching and learning of their respective topics. The clarity of writing exemplified in these volumes is a consequence of Horst's lifelong reaction to the ambiguous and redundant verbal formulations characteristic of much psychological prose and provides the exception to his own Parkinsonian maxim, which he dubbed Horst's Last Law of Communication: "Most things that most people say most of the time don't mean much of anything unless proven otherwise."

For his seminal work in differential prediction, for his many theoretical and applied contributions to psychological measurement and factor analysis, and for his instrumental efforts in founding the Psychometric Society and its journal *Psychometrika*, Educational Testing Service is pleased to present its 1985 Award for Distinguished Service to Measurement to Paul Horst.

ETS Award for Distinguished Service to Measurement Recipients 1970-1985

1970 E. F. Lindquist

1971 Lee J. Cronbach

1972 Robert L. Thorndike

1973 Oscar K. Buros

1974 J. P. Guilford

1975 Harold Gulliksen

1976 Ralph W. Tyler

1977 Anne Anastasi

1978 John C. Flanagan

1979 Robert L. Ebel

1980 John B. Carroll

1981 Ledyard R Tucker

1982 Raymond B. Cattell

1983 Frederic M. Lord

1984 Louis Guttman
Henry Chauncey (special award)

1985 Paul Horst

Changing Schools and Testing: an Uneasy Proposal

THEODORE R. SIZER
Brown University

> *Proposed*: That Educational Testing Service, the College Board, and the American Council on Education form and finance a commission to create and administer a Secondary School Leaving Exhibition (SSLE), thereby providing a demanding, realistic alternative route toward a high school diploma to traditional school attendance. Successful completion of the SSLE alone would constitute the basis for award of a diploma by ETS, the CB, and the ACE.
>
> The SSLE would include conventional timed and untimed paper-and-pencil tests, essays and other formal presentations, portfolios of independent or group endeavors, extended problem-solving exercises, and an oral interview. Candidates would have some significant choice among themselves. SSLE administrators would emphasize the need to ascertain a candidate's strength—what he or she knew and could demonstrate—rather than seeking out weakness. Maximum feasible, but not slavish efforts at "objective" measurement would be made, and careful regard would be taken to the process of how "subjective" judgments would be rendered.
>
> After a period of trial, efforts would be taken with state authorities and institutions of higher education to accept this alternative criterion of substantive secondary school completion as equivalent to traditional high school diplomas.

In many ways this is a frightening idea. The proposal involves a national, comprehensive examination, one that would dangerously dominate both the standards and content of American secondary education. As the SSLEs that are recommended are more—much more—than mere machine-graded tests, they will be extremely difficult to design well and administer consistently. The financial costs will be substantial. Given the admittedly subjective judgments affecting parts of the SSLE, the likelihood of endless challenges to the system is high. The operation will spawn yet another bureaucracy in an educational system already smothered in administrative machinery.

All these objectives are valid. Indeed, the idea is riddled with problems. What gives it credence, however, is that it is less flawed than are alternatives to it, particularly those that are emerging from tidal shifts in American mores, economics, and politics that are newly affecting our schools. We should be uneasy with this proposal and even more so with current trends.

Seven of these tidal movements, or conditions that relate to them, bear mention. Some must be viewed as progressive, some regressive, others neutral. However categorized, they add up, paradoxically, both to a fresh fluidity in American education and to a new politicization of the schools' curriculum.

First is the growing public demand for demonstrated performance, both of students and of schools. This sort of challenge has periodically emerged in the history of American education, such as in the Taylorite efficiency movement in this century's earliest decades. The current spurt dates from the late 1960s; the formation of the National Assessment of Educational Progress, for example, is testimony to its growing vigor then. The seventies brought us management-by-objectives, PERT charts, etc., devices that require some sort of precise output. The hullabaloo over SAT score declines marked the late seventies, and the blizzard of state-mandated tests and the creation of new sorts of devices such as promotional gates mark this decade. The public (or, more accurately, that minority of the public that has political awareness and clout) wants to see evidence that its educational investment yields demonstrable returns. The key word is *accountability* (at least for other people's kids and schools in someone else's neighborhood). Mere attendance at school is not enough. An SSLE responds powerfully to this now well-established public demand.

Second, however, is growing skepticism, at least among some small, if potentially influential, groups about the quality of existing assessment or testing devices and regimens. David Owen's polemic, *None of the Above*, is one sort of evidence; the dismay of leaders in minority groups about the apparent discrimination in tests against their numbers is another. A former United States commissioner of education publicly calls for another "SAT score decline," simply to call attention to the apparent socioeconomic class bias of that program.

The absurd extravagances of the accountability movement—such as basing judgments of schools and school personnel largely on series of locally designed, forty-item, multiple-choice tests administered periodically to students (the practice in one large city)—is giving thoughtful leaders pause. Fairness, flexibility, recognition of the myth of total objectivity, acceptance of the fact that effective assessment cannot be done on the cheap: these issues are being heard often these days. A sophisticated, flexible, and responsive SSLE is congenial with this new critical, realistic mood.

A third trend is of a different order, unrelated to assessment issues—the issue of *choice*, of the ability (indeed the right) of students and their families to decide which schools to patronize. Like so many current educational movements, this initiative seems inconsistent with concurrent efforts to standardize concepts of excellence, of state action plans that assert through regulation what the "One Best Program" is. Curiously, many of the same people who are arguing for more sharply focused centralized standards are also calling for choice—which must logically assume some variety among schools. (These folks argue back that varied schools can have common standards, a notion that survives in theory but usually, alas, collapses in practice.)

While many in the public sector rail against *choice* (perceiving it ultimately as fresh competition from voucher or tax-credited-financed, privately-managed schools), they support it in practice within their own sector. Magnet schools are the vogue in most cities, and in a few urban communities (such as Manhattan's District #4) one finds virtually all the public schools to be of that sort, each with its own more-or-less distinctive program. A concurrent interest in school-site management (giving a principal and his or her staff significant authority over their school's program) reinforces this tendency. Variety, choice, magnet schools: all such notions, now quite acceptable, bespeak a public, and to some extent, a professional concern for different roads toward a high school diploma. A wisely designed SSLE could give independent structure and standards to a system increasingly interested in the variety inherent in a policy of parental student choice. It provides an acceptable common finish line to races run over differing routes.

There are powerful pressures against this variety, of course; state-directed standardized schooling is in a certain ascendancy. Extreme reaction to it may, however, already be coming visible. For example, the home-school idea has ceased being considered the wildly aberrant notion of the far Left; that group is now joined by the political Right, which has its own reasons for the ultimate in private education, ones different in most respects from the ideologies of the followers of the late John Holt. Now, as dismay at increasingly bureaucratized and depersonalized school routines grows among folks in the political middle, the full spectrum of ideological persuasion may soon be represented in the home-school movement. And if one adds to that the barely tapped appeal of the home computer-turned-instructor, backed by a neighborhood tutorial center run on a proprietary basis by the manufacturer of that computer (or its software), one can visualize an alternative to the traditional school that breaks with conventional wisdom far more sharply than even the most ambitious magnet school. The mix of a significant population that feels that existing schools ill serve their children and the smell of a potentially vast market for the education/technology industry makes this prospect a realistic one. The existence of an independent SSLE would be welcomed by many (if not all) families interested in this non-school/schooling approach. The SSLE would be absolutely essential for the education/technology industry's respectability; without it, its sales pitch lacks an autonomously established and monitored standard.

Many today slight, even ridicule, this "end run" of the status quo. They are wont to call attention to the Gallup poll, which shows that over half of all Americans give the public schools an A or a B grade. The flip side of that finding is more interesting: almost half of all Americans believe that the schools—those institutions which provide communities with honored rites of passage for youth—are mediocre or worse. In a word, lots of Americans have some doubts about the schools. In this climate, the potential for the notion of *choice* to evolve into a welter of schooling systems is certainly present. The existence of a respected, authoritative (albeit voluntary) SSLE would provide a common standard of school completion in what may, perhaps, be an increasingly fractionated school system.

A fourth, and paradoxical, trend: the movement largely within state government toward centralized control of schools. The legislative hiccup that followed the release of the report of the National Commission on Excellence in Education in the spring of 1983 has resulted in a rush of fresh legislation and regulation, much in the form of mandated practice on the schools. Assessment devices permeate these new systems; schools and teachers and communities are rated by performance on examinations. Centralized bureaucracies (in some but, mercifully, hardly all the states) write syllabi, select textbooks, instruct teachers on what will be taught to their pupils when and for how long, and oversee this entire process through elaborate reporting procedures and external tests.

In some states, what is emerging is a politicized curriculum, a set of academic mandates, shaped inevitably by pressure group politics at the state capitols, that are imposed on all public schools (and in at least one state on non-public schools, to some extent). Decision-making percolates up. We find these days the oddity of a state legislature (Texas) debating just how many days a student must have passed his or her tests in order to engage in interscholastic athletics. "No pass, no play" is the slogan; but just what is "passing," and how long must one have "passed" courses to pass state muster? This riddle is now to be solved by state solons, who apparently believe they have more wisdom on such matters than do teachers.

As is abundantly evident, this rush of new centralized control arose from the well-intentioned dismay of significant political leaders caused by the sponginess, and often appalling incompetence, of the existing schools. Some critics feel, however, that their remedy merely begets another disease, indeed a scary one—an overwhelming politicization of the school's curriculum. All curricula are political entities, of course; but when the inevitable tussle over their particulars takes place at a community level, citizens feel some reasonable control over their design. Removal of decisions to remote state capitols eviscerates locale initiative, produces citizens' detachment, and lessens the real leverage of the typical concerned citizen while increasing that of the sophisticated, well-financed special interest pressure group. The specter of a tightly controlled curriculum, tuned to a central government's politics-of-the-moment, is no longer merely a theoretical possibility—it is a real and present danger.

A Secondary School Leaving Exhibition addresses these trends in two quite separate ways. First, it provides an independent (i.e., nongovernmental), national (i.e., not hooked to one state's political situation) and authoritative (given its sponsorship) standard as a powerful countervailing force to governmentally mandated school practice. Second (as mentioned earlier), it provides a respectable alternative for families that find state-mandated programs ineffective or unacceptable.

Fifth, an especially important group greatly affected by newly regulated schools is the teachers. The best among them know that standardized programs and narrowly uniform testing devices ill serve children, for the commonsense reason that children differ one from the other. This happy variety among students may be inconvenient, but it is inescapable; and instructors who are forced to pretend that all (for example) thirteen-year-olds *must* be interested in the same thing at the same time and *must* traverse this subject matter at essentially the same rate and *must* be examined on its mastery in precisely the same way inevitably become frustrated and embarrassed by the compromises they are forced to make. They are being required to perform in ways they know are harmful to children. The existence of an alternative standard, such as the SSLE, gives them a sort of anchor to windward, some independent external standard that could properly reward flexibly operated schooling, if a way to distance their particular institution from strict state mandates could be negotiated.

As school systems face teacher shortages, especially of the ablest folk, the matter of teacher frustration and morale, of teacher-level authority, will become more important than it currently appears. Talented people only take jobs that entrust them with important things. A school system that presents teachers with a realistic, clear, but flexible "target" (such as the SSLE) but leaves to them the design of the paths to that destination will attract and hold good professionals. One that demeans them with unconstructive regulation will not.

A sixth trend is the growing heterogeneity of the school population. Within a decade, over a third of our schools' students will be from minority groups, especially Black and Hispanic. Most large cities will continue to be majority-minority. The demand of these groups for assessment procedures that are both responsible and fair—that is, which do not improperly, even casually, discriminate against any group—will increase. This pressure will be resisted: standards are standards, some will say, with some justification. Three plus two always equal five. But when one gets into more complex levels of scholarship, into domains where inventiveness and imagination are critical, the precision of standards gets more problematic. An assessment device such as that posted for the Secondary School Leaving Exhibition, which gives the student some choice in the exercises he or she will attempt, provides needed middle ground, however imperfect, in inter-class and inter-group controversy, providing some sensible, sensitive accommodations often difficult to make within a state education bureaucracy. Furthermore, by its very existence, the SSLE gives families in both minority and majority groups an alternative form to any one rigid state-mandated exercise.

A final reason for urging the adoption of an SSLE is the student population itself. Virtually every report on the state of this nation's education has remarked on the vagueness of direction in the American school, and thus of standards. Most find American students to be cheerful, remarkably compliant, but docile, directionless. The fuzziness of goals must be one important cause of this docility. Students with a concrete, achievable target, one (like the diploma itself) that they value and thus desire, will work to grasp it (as countless examples, from the Advanced Placement examinations run by the College Board to the hurdles the Army sets before award of a sergeant's stripes, bear absolute witness). A soundly conceived SSLE would provide such a target, one which a student could choose. Its very existence would provide a standard—*one* standard, an alternative-to-government's standard, not the only standard—for American secondary schools.

One reflects on the Herculean tasks of constructing and financing as complex a device as that proposed here and is dismayed, uneasy. A hundred critical questions spring to mind, some technical, some substantive. How specifically must subject matter be outlined to potential SSLE-takers? How can oral interviews be at all objective? What will the exercise cost each applicant; and, if not the applicant, who pays? Many more questions might be asked. It is easy to lose heart, buried under with doubts.

But one persists in view of the alternatives: A set of curricula prescribed by centralized state politics with no serious alternatives allowed. Or a school system without any generally accepted ultimate standards that might clarify the purposes of schooling for individual students as well as for the public at large. Or a new non-school educating system—a blend of home- and technology-driven learning—without some ultimate reasonable and rigorous external standard. These prospects are real. Major changes in how our schools work are guaranteed, whether or not our profession chooses to grapple with or shape them.

And so a Secondary School Leaving Exhibition seems a highly recommendable enterprise. Why should educators—those in the "system"—support it? Won't it put their jobs at risk, as many fear that vouchers will? An external graduation credential could undermine the status quo; indeed it *should* erode those aspects of current policy and practice that demonstrably ill serve students. However, a SSLE, even if—perhaps especially because—it is voluntary and controlled by the profession rather than political forces, concurrently could give a renewed and needed focus to school programs and, in its flexibility and absence of any mandates about how precisely one prepares for its exercises, could provide the authoritative autonomy to school staffs that the proudest and ablest of them know they need to adapt their programs to their particular students. In sum, it gives focus and freedom—qualities that wise educators hunger for these days.

Why turn for this service to ETS, the College Board, and the American Council on Education? Educational Testing Service has the scholarly and technical expertise to develop a program of this complexity. The College Board is one of higher education's most influential gatekeepers, and the American Council on Education has had forty years experience with high school equivalency examinations. These are the obvious partners, and their collaboration will give the risky experiment the leverage it requires.

And so, let us be uneasy; but let us also take note of trends now well under way. And let us have courage to try something new.

Cognitive Research and Future Test Design

Earl Hunt
The University of Washington

Tests for personnel selection are one of psychology's major technological contributions. The simplest view of technology is that it is the result of practical refinement of scientific knowledge. If this is correct, cognitive psychology, the scientific study of thought, should dictate our tests for evaluating people's cognition.

In practice, the situation is more complex. Demands for solutions to problems of perceived social importance pull science as much as science pushes application. We require screening tests because we perceive a need for objective classification measures. In addition, we need to fit measurement techniques into a rather rigid framework of cost-effectiveness. The measurement procedures we use then produce facts, to be explained by a theory of cognition. But a theory of test-score generation may not address central questions in cognition, or vice versa. This does not mean that the development of cognitive theories is unrelated to test development. It does mean that we must examine relationship carefully, by considering the logic of each effort singly, and then asking how they mesh.

The Testing Situation

Cognitive testing is done for the purpose of prediction. Psychometricians have developed a substantial body of mathematical methods to further this end. The methods are based on a straightforward model of mental competence. It is assumed that test scores are derived from a small number of measurable mental capacities called factors. Two of the best known are verbal comprehension, loosely, the ability to deal with language, and spatial-visual reasoning, the ability to manipulate visual images inside the head (Carroll, 1982). While there are other factors, I will use these two throughout to illustrate a variety of points. In psychometric prediction, a person is represented by his or her values on the various factors. A prediction is a mapping from every possible combination of factor scores onto the value of a score that represents success in some criterion situation.

Figure 1 presents a geometric version of the psychometric model. The possible mental competencies are represented by a space, whose dimensions are the factors. This is shown in the left-hand part of the display, using the verbal and spatial factors as an illustration. Individuals are represented by points in the space. In practice, though, unreliability in measurement does not allow us to locate the exact point for an individual. Instead, the best we can say is that (up to a specified level of probability) the person's mental abilities lie somewhere inside a (hyper) ellipsoid in the test space.

Predictions are made by mapping points in the test space onto vectors representing job performance. This is shown on the right of the figure. Each of the vectors represents a specific occupation; doctor, lawyer, physics major, etc. Applying straightforward statistical procedures, one

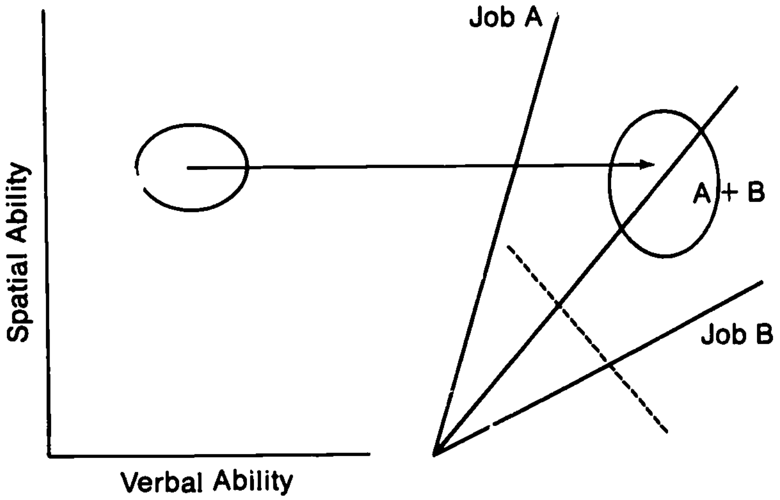


Figure 1. A graphic representation of the psychometric prediction model. People are represented as points in a space defined by the dimensions of test scores. (Verbal and spatial reasoning scores are used in the example.) Test inaccuracy will restrict locating each person to a region (ellipse) rather than a point on a vector representing a particular job. The orientation of the vector to the dimensions of the test space are determined by the nature of the job. If information on specific jobs is not available, test performance is mapped to a vector representing a hypothetical average job. This is shown here by job vectors A, B, and average job vector A + B.

maps every point in the test space to a point on each job vector. The distance between this point and the origin of the job vector represents the predicted degree of success for an examinee in a particular job. The orientation of the vectors with respect to the dimensions of the test space show how important each dimension is with respect to each job. This is also illustrated in the figure, which shows two abstract jobs, A and B. Job A is nearly parallel to the verbal dimension (a lawyer?).

Just as there is uncertainty about a person's position in the test space, there is uncertainty about the predicted point on each of the job vectors. The uncertainty can arise from three distinct sources. Often, we are unable to measure directly performance in jobs. Thus, we rely on averaging performance across jobs. This is particularly true in predicting performance in higher education. Undergraduates do not take exactly the same courses. Statistical analysis of aggregate measures such as grade-point average (GPA) map test performance onto an average of job vectors. The averaged vector represents an aggregate but non-existent job. I have shown this in the figure by applying the mapping to the vector A + B, rather than applying it directly to job A or job B. A second, and very large source of uncertainty comes from our inability to obtain reliable, valid measures of on-the-job performance. Anyone who is disappointed in the moderate correlations between screening test performance and GPA might look at the reliability of grades themselves. Or still worse, if one wants to become very discouraged, one need only examine the reliability of supervisor and/or interview ratings. Errors or inadequacy in the measurement of job performance (including educational performance) are probably much greater than errors of measurement in testing. Figure 1 reflects this, for the "ellipse of uncertainty" is considerably larger on the left side than on the right.

Any attempt to improve prediction must be reflected by a modification of the diagram in Figure 1. There are four ways in which the diagram could be improved. On the left-hand side, testing could be improved, either by developing more accurate measures of those psychological variables that we now measure, or by extending the range of variables that we measure. More accurate measures would shrink the ellipse, ideally to a point. If the range of tests was expanded to new psychological dimensions, the ellipse would become a hyperellipse, because the new measures would add dimensions extending out of the plane of Figure 1. This would allow the mapping to discriminate between points that are now treated as equivalent. For instance, one could distinguish between "high anxious" and "low anxious" individuals who had identical scores in spatial and verbal ability. This sort of distinction might be relevant for some occupations.

Testing could also be improved by concentrating on the right-hand side of Figure 1. The development of statistics for specific jobs, rather than for aggregates such as the GPA, would make it possible to develop locally accurate mappings. The development of more, or more reliable job performance measures would shrink the ellipse of uncertainty on the criterion side of the figure.

There is further problem. Any attempt to implement these abstract considerations must come to grips with reality. Testing operates under rigid economic constraints. By far the strictest of these is the amount of examinee time available. The inevitable passage of time between testing and the evaluation of testing introduces more uncertainties. Many high school graduates will commit themselves to a university, but not to a major within that university. One can only predict averages for such people. These constraints limit the improvements that can be implemented, whatever the benefits of those improvements might be.

The Enterprise of Cognitive Psychology

Theories in cognitive psychology are attempts to explain the process of thought. This purpose requires a quite different representation of a person than the mathematical representation so useful in testing. The point can be illustrated by a thought experiment. Suppose a person were asked to attempt the verbal/mathematical puzzles that Lewis Carroll scattered throughout in *Alice in Wonderland*. The number of problems that a person solved might be predictable from knowledge of his or her verbal, spatial, and numerical ability scores. Knowing the prediction equation would not tell us how any person solved any one of the problems. Explaining the process of problem-solving is just what a theory of cognition should do.

This point of view is hardly new, either in psychology or education. More than thirty years ago, Bloom and Broder (1950) presented a striking series of protocols showing that identical answers to questions could be produced by very different processes of reasoning. They went further, showing that examining the reasoning process told much more about a student than examining the answer sheet. Shortly thereafter (and apparently without knowledge of Bloom and Broder's work), Newell, Shaw, and Simon (1958) published the first of a series of papers that set the tone for modern theories of cognition. Newell et al. agreed that a cognitive theory should be stated as a design for a machine capable of doing some specified cognitive acts; playing chess, reading a newspaper, etc. In their first and subsequent papers (especially Newell, 1981, and Newell and

Simon, 1972) they carefully explained that they were concerned with abstract machines and that their position did not, in any way, amount to a claim that the modern digital computer is a model of the brain. The logical ramifications of Newell and Simon's approach has been e·'plored in considerable detail by Pylyshyn (1984). Much of the argument presented here is a specialization of his reasoning to the testing situation.

The cognitive psychology approach is based on the truism that thinking beings solve problems by manipulating mental models of the environment, instead of trying out responses until they find some that work. The problem-solver constructs these ınental models by combining his or her concept of the current problem with personal information about the world, as extracted from previous experiences. No one would argue with this, neither would anyone argue that, although mental models can be discussed abstractly, they eventually have to be realized by physical processes in the brain. The interesting thing is how these truisms limit possible theories of cognition.

A cognitive theory is inevitably a multilevel theory. Problems occur in the thinker's present environment. In order to solve them, the problem-solver must apply a variety of content-free information processing functions both to the stimuli at hand and to the problem solver's records of past situations. The application of the content-free processes, however, is controlled by content-sensitive problem-solving methods, based on the thinker's experience with previous problems that, in some way, the thinker perceives as similar to the current one.

The flow of information is shown in Figure 2. The top of the diagram represents "the environment," both past and present. The bottom of the environment represents physical processes. I show two of them, to stress the point that the argument is not limited to human thought. The physical equipment of the brain provides a problem-solver with certain functional, information-processing capacities. The person's experience determines how to use these capacities to build and manipulate internal representations of the external world. Perhaps the best example of this is the universal human ability to learn a language. All human beings possess the acoustic-pattern recognition, information storage, and information-retrieval capacities required to speak. Experience determines how they use these functions to learn to speak.

When a problem is presented it will be described, internally, in the light of information already in memory. Memory contains two separate types of knowledge; declarative knowledge of facts and procedural knowledge about how to do things. The latter can be thought of as "programs" that tell the brain what information processing functions to execute in different situations. For instance, when a problem first presents itself, the individual must apply a pattern recognition program in order to decide how to treat further input relevant to the problem. A good example is the process of reading comprehension. Most coliege-educated people have learned that the first one or two sentences of an essay state the general topic, so they use the information in these sentences to activate schema that guide understanding of the remainder of the essay (Kieras, 1978).

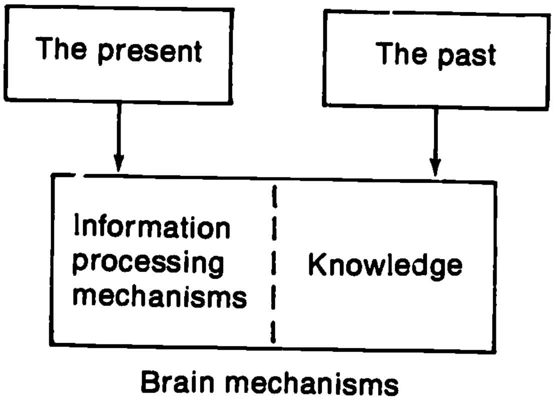


Figure 2. The cognitive psychology view of mental processes. Ultimately, the capacity of the mind is determined by brain mechanisms. The mind, however, is best thought of as a combination of content-free information-processing actions (e.g., the ability to hold an item of information in short-term memory) and a large set of learned rules that collectively determine our knowledge of the world. When a problem is presented in the present, it is solved by executing content-free, information-processing mechanisms, under the direction of previously acquired knowledge.

The Problem for Prediction

Emphasizing the process leads us to be more interested in how a person behaves in a problem-solving situation than in whether or not they meet some criterion of success. In order to construct an explanation of process (which is harder than making a prediction), one must have a considerable understanding of the criterion situation. How fast must information be processed in it? What declarative and procedural knowledge is required for success? Are there alternative ways in which problem-relevant information can be processed and, if so, how can a problem-solver determine what processes are best for a specific case? When these questions about the criterion situation are answered, then the cognitive theorist can consider the relationship between performance, the possession of knowledge of various sports, and the extent to which performance relies upon particular (knowledge-free) information-processing functions, such as generalized abilities, to recognize verbal codes or to manipulate mental images.

None of these questions makes a great deal of sense when the task is to predict global criteria, such as "success in college," because these criteria are abstractions from specific but often dissimilar situations. Obviously, the knowledge required to be a physicist is not the same as the knowledge needed to be a historian. Furthermore, content demands are not the only distinctions between situations. In particular, different criterion situations may require different mixes of what the cognitive psychologist calls "controlled" and "automatic" information processing. A related, but not identical, distinction can be made between situations that require the use of general and specific problem-solving methods. These considerations can be illustrated by conducting another thought experiment:

How would you answer the following questions?

1. What is your telephone number?
2. What is the telephone number of the Classified Advertisement department of the *New York Times*?
3. What is the telephone number of the editor of *Pravda*?

Most adults are expert at answering the first question. It is an example of highly automated information processing. Once you recognize the question, all you have to do is to execute precisely those information-processing functions required for the retrieval of information from long-term memory. Now suppose that a cognitive theorist were asked to predict how well different people could retrieve their own telephone numbers. It happens that there are reliable, individual differences in the speed of retrieval of information from long-term memory (Hunt, 1985); and these differences can be tested, so a predication could be made. However, in absolute terms, the differences are not large, so ecologically significant variations in individual performance on this simple tasks are likely to be small, at least in a population of healthy adults.

Now consider the second question. Few people could answer this question by executing a highly automated response. On the other hand, many adults know a series of serially ordered information processing steps. These include finding a telephone book, turning pages, and so forth. Problem-solving of this sort is characterized as "forward driven," because the stimuli present at each step virtually dictate the next step to be taken.

Constructing tests for forward-driven problem-solving is a straightforward process. The examinee must demonstrate knowledge of highly specific, content-bound, problem-solving behavior. One of the most striking findings from the recent spate of experiments on "experts," varying from studies of chess players to studies of cab drivers, is how much skilled performance relies on context-bound, forward-driven problem-solving. This finding lends further credence to an earlier point. In order to make really good predictions about a criterion situation, one has to understand the criterion behaviors. The problem this poses for predicting to vague criteria, such as GPA, is obvious.

Now, how would you find a telephone number in the Soviet Union? The Moscow telephone directory is not publicly available information. You would have to execute some general problem-solving procedures, such as "To find the answer to a question, identify a person who already knows the answer, and ask them to tell you."

From the viewpoint of cognitive theory, such strategies have several interesting characteristics. First, and perhaps foremost, the situations in which they should be used are loosely defined. Second, general problem-solving methods put more stress on abstractions, such as the generation of subproblems or the use of analogies. This means that it is harder to explain to a learner precisely what is to be done to execute a general strategy. Third, a general problem-solving procedure typically involves many more information-processing steps than do specific procedures. This means that the general procedures are slower and that they are more subject to disruption by any weakness in the information-processing functions that they use. It is a simple issue of reliability; the more use that is made of a primitive function, the more likely that function is to break when it is needed. This is particularly true of the manipulation of information in working memory. One of the striking characteristics of (specialized) expert problem-solving procedures is that they minimize short-term memory loads during problem solving (Ericsson, 1985).

Tests of general problem-solving ability can be constructed. This appears to be what tests of g and gf (fluid Intelligence) are, because they place people in relatively unusual situations and force them to figure out how to solve unfamiliar problems (Sternberg, 1981). However, for reasons that will be outlined below, this is done in only a rudimentary way.

As the telephone example shows, the more general a problem-solving procedure is, the less efficient it is likely to be in specific cases. The more specialized a procedure is, the closer the process comes to being limited by the efficiency of execution of a *small* number of basic information-handling processes. Going the other way, more general problem-solving procedures are likely to require execution of a large number of information-handling processes and, at the same item, to depend a great deal upon a person's knowledge of problem-solving in the abstract. Furthermore, there will be many cases where a person is "in between," relying in some cases upon the possession of knowledge and, in other cases, upon the possession of problem-solving skills.

The Implications of Cognitive Theory for Testing

Cognitive-process studies focus on information-processing functions, general problem-solving skills, or the way that people use knowledge in specific problem-solving situations. The tendency has been to avoid studying the interaction between these different levels of cognition, which is unfortunate for testing, because testing, perforce, has to deal with the whole person. Nevertheless, it is possible to fit cognitive research into testing, by combining the three levels of cognition—general problem solving, knowledge utilization, and information processing—with the three concerns of testing—amplifying old measures, developing new measures, and understanding criterion situations. Each research level has something to contribute to testing, although all levels do not contribute to all concerns.

At the information-processing level, we must make a distinction between processing linguistic and non-linguistic information. There are some interesting new computer-based technologies that can be tapped to provide tests of linguistic information-processing functions. For instance, the conventional vocabulary test attempts to determine how many words a person knows. The advent of computers in testing makes it possible to use laboratory-developed measures to tell us how fast a person can recognize a known word, either in isolation or in the context of other words. Clearly this is a basic function in reading.

Measures such as this are useful in developing a cognitive theory of why certain individuals are more adept at language processing than are others (Hunt, 1985). It is not clear, though, that measures of linguistic information-processing will add anything to our present ability to identify those individuals. The reason is that we already do a very good job of identifying verbal ability, because present-day tests can virtually present a work sample for reading. Work sampling is the best possible way to predict performance, when it is feasible, so there is only marginal room to improve our identification of verbal ability.

The situation is quite different if we look at spatial-visual information processing. Spatial-visual ability is typically tested by presenting geometric figures and asking people to move them around "in the mind's eye." Spatial-visual testing has been something of a disappointment. Although it can be shown to contribute to criterion performance in certain fields—notably mathematics, engineering, and architecture (McGee, 1979)—the predictions are not as accurate as one would hope them to be. Indeed, spatial-visual testing was dropped from the Armed Services Vocational Battery some years ago because it did not predict performance particularly well in any of the various military occupational specialties, even those that would ostensibly use the ability.

James Pellegrino and I have pointed out elsewhere (Hunt and Pellegrino, 1985) that the criterion situations offered as examples of spatial-visual ability, notably in aviation, almost all require that a person deal either with visual fields containing moving elements. These elements are missing from the conventional paper-and-pencil test format, simply because printed pictures are, of necessity, static pictures. On the other hand, cognitive psychologists interested in the processes of visual-spatial reasoning have developed a number of experimental techniques for studying how people react to moving elements. For instance, Poltrock and Brown (1984) developed a tes: of people's ability to add elements to a visual display that was being developed "inside the head." Their procedure tested a function that Kosslyn (1980) had identified as an essential element of visual imagery. In a yet unpublished work, Pellegrino and I have studied people's ability to project the paths of moving elements forward in both time and space. Some preliminary indications suggest that the ability to deal with these dynamic visual dısplays is not identical to the ability to deal with static displays.

Both Poltrock and Brown's measurement procedures and those that Pellegrino and I have developed rely upon computer-controlled display presentations and response recording. From the examinees' view, these testing procedures are somewhat like a video arcade game (2). Such procedures are not practical using the traditional paper-and-pencil format, but they will be practical when testing relies on microcomputer test stations. That day is surely not far off.

Similar techniques can be used to evaluate information-processing functions that are not now tapped by present tests. I will offer two examples. One is simply a measure of the speed with which people make rapid perceptual-motor choices, e.g., the time required to decide which of two lights has come on. Historically, Galton (1883) thought that this sort of measure should tap an important determinant of mental competence. As is well known, he was unable to find any interesting correlates. What is not so well known is that Galton's techniques for measuring choice reaction time would, today, be considered very questionable laboratory technique. Studies using modern procedures and repeated trials (e.g. Vernon, 1983) suggest that this ancient variable is worth further examination.

A second candidate for the expansion of testing is the ability to "pay attention." This has been suggested as a part of mental competence by many authors (e.g., Spearman, 1927). Unfortunately, attention is as slippery a concept as is intelligence. The most recent work in this field suggests that the important individual-difference component is the ability to direct one's attention to a particular part of the stimulus complex.

Researchers can evaluate this ability by requiring a person to shift rapidly from processing one stream of signals to processing another. A good deal of work in this area has relied on a dichotic listening paradigm in which people are asked to monitor signals presented in the other ear (Gopher, 1982). My colleagues and I have explored similar techniques that require shifts of attention to different places in the visual field. Recently, we have extended this work to include studying individual differences in the ability to shift from one form of semantic information processing to another, e.g., to shift rapidly from doing addition to doing multiplication. We have found that there is an ability to shift attention that generalizes across all these superficially quite different situations.

The examples that have been presented represent only a few of the sorts of information-processing functions that *can* be tested, providing that a computer-controlled testing format is used. But *should* testing of basic information processing be expanded? To answer this question, we must look at how individual differences in information processing seem to affect criterion performance. Two prototype cases can be considered. In one a person is executing a well learned, efficient problem-solving procedure. This person may be pushing the limits of his or her information-processing capacities. Thus, information processing, not task knowledge, is the limiting feature on *performance*. However, it may not be the limiting feature on success. I conjecture (and could hardly be disproved) that society does not provide many tasks where successful performance is dependent on superb information processing. Society has to be tolerant enough so that the average person can do the job.

More generally, in order to use information-processing tasks as predictors, one must have sufficient understanding of the criterion task so as to identify appropriate information-processing functions; and the people being tested should already know the task so well that their information-processing capacities, not their knowledge, place a limit on their performance. While there are military and industrial situations where these constraints are met, general education is quite another matter. The criterion will always be diffuse and by definition, students will be evaluated as much for their ability to learn as for their ability to perform after they have learned. What sort of measurement problem does this pose?

Information processing is important in learning and general problem solving, but in a different way. The information-processing functions must be able to support long, relatively complex problem-solving procedures. When we are dealing with healthy young adults, though, information-processing capacity may not be the limiting factor. Performance is probably more determined by knowledge of the general problem-solving procedures, and acknowledgement that solving a problem using these procedures is worth the effort it takes. In any case, for the purposes of prediction, it makes more sense to test the use of the problem-solving procedures directly, than to test them indirectly by looking at information-processing functions.

The tests that are pointed to as tests of "fluid intelligence" (Horn and Donaldson, 1980) or "dealing with novelty" (Sternberg, 1981) are attempts to evaluate general problem-solving ability. This is particularly true of inductive reasoning tests, which require the examinee to detect patterns in stimuli. Pattern detection is an especially important ability because, as has been noted, the first step in any problem-solving procedure is to decide how to categorize the problem itself.

Cognitive research on general problem solving, and especially on inductive reasoning, has boomed in the past few years. To what extent can that research be used to develop new testing methods? There is reason for both optimism and pessimism.

The optimism comes from the rather careful theoretical analysis that has been given of such things as "analogical problem-solving." The beautifully undefined gestaltist terms "restructuring" and "seeing the relations between parts of a problem" have been replaced by effective computing procedures for discovering analogies between current and past problems. A modern theory of problem solving specifies the data structure that is used to represent stored knowledge, how the data structure representing the current problem is created, and how the two are matched. A particularly interesting development, which is still in its infancy, is the use of this sort of theory to build intelligent computer-assisted instruction (ICAI) systems (Anderson, 1984; Clancey, 1984).

These systems contain both a model of the information to be taught and a model of the student's current knowledge structure. If the latter model has been built correctly, the program will be able to construct items that test the student's ability to expand his or her current model to encompass new cases.

Why not adapt the same philosophy in building assessment procedures?

There is a serious practical objection to applying ICAI techniques to testing. The data rate is too slow. This has nothing to do with computers—the problem is inside the student's head. Consider the problem of designing a simple computer program, an area that has actually been the topic of ICAI investigation (Anderson, 1984). A semi-realistic problem will require at least half an hour to solve, and a truly realistic problem should take several hours. Chance factors do enter in here. Some people may mistype a symbol, others may be lucky enough to have worked on a similar problem just before the test. Thus, as an evaluation device, problem-solving item reliability. The elementary mathematics of test theory show that tests consisting of only two or three unreliable items simply cannot be used in prediction situations.

The obvious answer is to make the test longer. A realistic "test" derived from cognitive theories of general problem solving and learning might require several *days* of the examinee's time. While we do not know whether or not this would be cost-effective, we do know that the capital costs of testing would increase dramatically. It is doubtful that society would agree to investigate the question.

The alternative is to combine instruction and evaluation. If the people to be examined are currently enrolled in an educational program that uses ICAI, evaluation procedures can be built into the teaching. There is nothing inherently wrong with this. Indeed, there is some attraction to the argument that if a teacher has spent a great deal of time trying to build up a student's problem-solving abilities, the teacher has a good idea of how adept the student is. Presumably, Socrates could have written good letters of recommendation. But all this assumes that "Socrates," in this case, an intelligent CAI program, exists. All that we have today are experimental models. Within five to ten years there may be working systems, but only for a few fields, simply because such programs take great ingenuity and time to write. This widespread availability of ICAI programs is probably fifteen to twenty years away, if only because of the effort required to develop them.

There is another issue that may impede the development of a combined ICAI-evaluation program. ICAI programs will change more than the form of evaluation. Evaluation and instruction will merge, lessening the need for an evaluation agency outside of the school itself. Truly intelligent CAI programs might substantially alter the role of the human teacher. In other words, our institutions will have to change. Any guess about the trauma induced by the change would be far beyond the scope of this article.

In closing, let us look at knowledge. As I have noted already, recent research in cognitive psychology has shown that content-specific, knowledge based problem solving is far commoner than the use of general problem solving methods. If test evaluation is to be for the purpose of prediction, a great deal of work needs to be done on criterion analysis—determining what knowledge is used by people who currently work at various criterion tasks. A particularly useful analysis would be of the path of learning from the point of the evaluation until the point of maturity. A specific example may help make this point. It has been suggested that erroneous naive models of physical phenomena impede the learning of formal physics (Caramazza, McCluskey, and Green, 1981). The matter is in some dispute. If the conjecture is correct, the test developer would like to know what the course of learning is as students move from holding naive to holding correct models. In particular, the test developer needs to know what naive beliefs are particularly hard to stamp out. Given this knowledge, the test developer can design a test to see not just what correct answers students can give, but also to see what incorrect beliefs they have.

Conclusions

How much can testing gain from modern cognitive psychology? The answer to this question may hinge more on the status of testing than it does on the state of research on cognition. So long as testing is viewed as something that takes place in a few hours, out of context of instruction, and for the purpose of predicting a vaguely stated criterion, then the gains to be made are minimal. The largest gains are likely in fields outside of the general "verbal performance" area, simply because the present verbal competency tests are so close to a work sample.

If testing can be expanded to prediction of success in specific fields, where a careful analysis of the cognitive demands of the criterion are possible, then substantial, though specialized, improvements can be made. Such situations are probably not very common in education, although they do occur in industrial and military settings. Substantial gains can also be expected if cognitive theories are applied to the diagnosis of individual pathologies of thought, e.g., to possible deterioration of performance associated with aging. Again, though, it is not clear that these are educational problems.

If the current enthusiasm for ICAI and expert systems can be transformed into reality, then a potential breakthrough in educational methods could be made. The breakthrough would involve assessment of people's problem-solving skills and knowledge bases as they were learning new material. This evaluation would take place over periods of days and perhaps months, and would not be suitable for inclusion in a traditional three-to-eight-hour testing session, held apart from normal instruction. While this is perhaps the most exciting "blue-sky" promise of a change in evaluation procedures, it can only take place if present tantalizing bits of scientific progress are transformed into solid technological works, and if there are major changes in the institutional procedures for testing and evaluation. Both these developments could easily take more than 25 years.

Footnotes

[1] The testing procedures described here were developed with the support of the Office of Naval Research, contracts N00014-84-K-5553 and the Naval Personnel Research and Development Center, Contract N66001-85-C-0017. The assistance of Professor James Pellegrino, Simon Farr, and Robert Frick is gratefully acknowledged. The opinions expressed are my own and do not represent opinions in the Office of Naval Research or the Naval Personnel and Research Development Center.

[2] At this point in the presentation a brief film showing the imaging techniques was shown.

[3] While this is true for testing, exactly the opposite is true for diagnosis of individual cases. If a person is known *not* to use general problem-solving procedures, it would be sensible to test to see if that person's information-processing capabilities could support the procedures.

References

Anderson, J.R., R. Farrell, and R. Savers. (1984) "Learning to Program in LISP," *Cognitive Science 8*(2) 87-129

Bloom, B., and L. Broder. (1950) *The Problem Solving Processes of College Students.* Chicago: U. of Chicago Press.

Caramazza, A., M. McCluskey, and B.F. Green. (1981) "Naive Beliefs in 'Sophisticated' Subjects: Misconceptions About Trajectories of Objects," *Cognition 9* (2) 117-124

Carroll, J.B. (1982) The Measurement of Intelligence in R.J. Sternberg (ed.). *Handbook of Human Intelligence* Cambridge: Cambridge U. Press.

Clancey, W.J. (1984) "Methodology for Building an Intelligent Tutoring System," in W. Kintseh, J.R. Miller, and P.G. Polson (eds), *Methods and Tactics in Cognitive Science* Hillsdal, N.J. Enlbaum Associates.

Ericsson, A. (1985) "Memory Skill," *Canadian Journal of Psychology 39* (2) 188-231

Galton, F. (1883) *Inquiries into Human Faculty and its Development*. London: Macmillan

Gopher, P.A. (1982) "A Selective Attention Test as a Prediction of Success in Flight Training," *Human Factors 24* 173-183

Horn, J.L., and C. Donaldson. (1980) Cognitive Development II: Adult Development of Human Abilities, in J. Kagan and O.G. Brian (eds.) *Consistency and Change in Human Development*, Cambridge, MA: Harvard U. Press.

Hunt, E., and J. Pellegrino. (1985) "Using Interactive Computing to Expand Intelligence Testing: a Critique and Prospectus," *Intelligence*. 9 (3) 207-236

Kieras, D.E. (1978) "Good and Bad Structure in Simple Paragraphs: effects of apparent theme, reading time, and record," *Journal of Verbal Learning and Verbal Behavior*. *17* 13-28

Kosslyn, S. (1980) *Image and Mind* Cambridge,. MA Harvard U. Press.

McGee, M.G. (1979) "Human Spatial Abilities: Psychometric Studies and Environmental, Genetic, Hormonal, and Neurological Influences," *Psychological Bulletin 86* 889-918

Newell, A. (1981) "Physical Symbol Systems," *Cognitive Science* 4 135-144.

Poltrock, S.E., and P. Brown. (1984) "Individual Differences in Visual Imagery and Spatial Ability," *Intelligence 8* (2) 93-138

Pylyshyn, Z. (1984) *Computation and Cognition*: Cambridge, Mass. Militi Press.

Spearman, C. (1927) *The Abilities of Man* New York: Macmillan

Sternberg, R.J. (1981) "Intelligence and Non-entrenchment," *Journal of Educational Psychology 73* 1-16

Vernon, P.A. (1983) "Speed of Information Processing and General Intelligence," *Intelligence 7* 53-70

Measurement Research That Will Change Test Design for the Future

WILLIAM C. WARD
Educational Testing Service

When I was invited to speak about the future of testing, my first thought was, "No problem. In the last several years that's been a constant topic around ETS, and I've had my share of opportunities to speculate. I'll just polish up the standard remarks and I'll be all set."

Then I looked at today's cast of characters and decided it wouldn't be quite that simple. One big chunk of my "spiel" has to do with the implications of cognitive science for measurement; and of course that piece was taken, appropriately enough, by Earl Hunt. Another chunk has to do with what's happening to bring technology within reach of the examiner. Again, that piece was spoken for. So, what's left?

Finally, I realized that the problem wasn't what to say, but what to leave out. Twenty minutes is just too short a time to deal with all the elements of our possible futures; I should be happy to lop off some major pieces. So I simply assert that some of the most important advances in measurement will grow out of the concepts and methods of the cognitive scientists; and, while I don't know what new technologies we'll have a decade or two from now, what is already on the shelf is more than enough to keep us busy. There is every reason to believe that more and more, we'll be using the computer in developing and in administering our tests.

With just those nods toward other's turf, I want to spend my time on four ways in which tomorrow's tests will be different from those we have lived with for the past many decades. Each represents an area in which some of the needed research and development is in the bank—but not all. I will allude here and there to what has been completed, but not very systematically. It's more interesting to think about the problems we have yet to resolve than the solutions that are already in the journals.

Adaptive Testing

Let me begin with an area in which the future is upon us—computerized adaptive testing. The concept of adaptive testing is more than familiar to many of those here, so I won't belabor it: An adaptive test is one in which each examinee answers different test questions chosen to ensure that each receives the best available test for his or her level of skills. In a full-fledged implementation of the process, the examinee's ability is estimated after each question, based on all the questions the individual has answered thus far. The computer then selects and administers the next question that is most appropriate in light of this estimate. This matching of questions with examinees yields very efficient measurement. Fewer than half as many questions are needed as in conventional testing; and it yields broad-range measurement—a test can measure accurately for individuals of widely different levels of skills.

The adaptive testing process rests on a foundation of more than 25 years of theoretical research in Item Response Theory. We also have a shorter but quite respectable period of experience in the practical use of IRT—in equating standardized tests and in scoring tests—that provides an indirect basis for confidence as we apply the theory to adaptive testing. And since the early '70s, we have had direct research on the adaptive testing process.

I don't intend to summarize 25 years of research. I'll simply refer you to Fred Lord's 1980 book (10), or Ron Hambleton's 1983 (volume 7). There are also several very useful reports, produced by a team led by Bert Green, that completed a comprehensive analysis of the issues to be resolved in preparing for adaptive delivery of the Armed Services Vocational Aptitude Battery (5, 6). The issues are legion, but the conclusion is that there are no critical measurement barriers to the delivery of this test adaptively.

Thanks to the hardware manufacturers, the economic barriers to practical adaptive testing are also rapidly falling. For about the last 18 months, ETS and the College Board have been piloting an adaptive basic skills test intended for use in college-placement decisions. When we started our development, the equipment needed to deliver such a test sold for more than $3,000. We can now do quite nicely with an off-the-shelf personal computer that retails for about $600.

All of this may sound as though adaptive testing ought to be considered today's technology, not an item for the future. However, we're far from finished with the research that is needed. Some of the issues and problems to be dealt with are as follows:

Violations of IRT Assumptions

First, we need a better understanding of the effects of violations of the assumptions of Item Response Theory. IRT makes several strong assumptions about a test and the domain that it measures. One of these is that of unidimensionality—all of the questions in a test module must measure a single dimension of aptitude or achievement. This assumption is unlikely to be strictly true in any complex domain, particularly in tests of achievement. Does this mean that the sphere of application of the adaptive process must be sharply limited? Happily, it appears not. Simulation studies, such as those done by David Weiss (16), have shown IRT to be robust in the face of reasonably large violations of unidimensionality. Empirical studies, such as those recently completed by Linda Cook and Dan Eignor (2), show IRT equating to be feasible for achievement tests in several content areas. This work implies that adaptive testing in these domains will also be feasible. But Cook and her collaborators also raise cautions (1)—some good tests are too heterogeneous to provide good IRT results, and the boundary conditions are not very well understood. We need research to better define those conditions and methods of testing for them.

A second critical IRT assumption is that of local independence—performance on one test item must be independent of that on other questions administered to an examinee. This assumption poses some difficulties as we contemplate translating traditional aptitude tests into adaptive form. In measuring reading comprehension, for example, we often have as many as six or eight questions associated with one reading passage. It wastes time to administer only one question per passage; it would be most desirable to select the two or three or four that are appropriate for a particular examinee. But can we? IRT equating studies are also relevant to this question, and have not been so positive here. The context in which an item is given can make a difference in the way it is understood.

A variety of activities is needed to understand context effects better and to find ways to get around them. Maybe we can identify the sources of these effects, quantify them, and make adjustments in item parameters as we test. Maybe we will need further developments in IRT models that parameterize sets of questions rather than individual questions. And maybe we will need to develop other item types that satisfy the local independence requirement but measure the same characteristics as those that don't. In reading comprehension, for example, the cloze technique may provide a promising alternative.

Construction of Adaptive Tests

Another set of issues needing attention has to do with ways to make the construction of an adaptive test easier. For example, with the three-parameter model, a rule of thumb is that 1,000 pretest cases should be used in calibrating new items. Having more than that can make a noticeable difference. Can models or data-collection strategies be devised that will decrease that number? Similarly, how can we best collect the data needed to calibrate new items within an administration of an adaptive test? That sounds straightforward, but it isn't. Hopefully, work under way by Fred Lord and Martha Stocking at ETS, in collaboration with Darrell Bock, Michael Levine, and Fumiko Samejima, will develop appropriate techniques.

Finally, can we find ways to calibrate items without ever pretesting them? A prototype is provided by work Isaac Bejar is doing in the measurement of spatial ability. The task is to determine whether two complex figures, presented in different orientations, are identical or are mirror images of one another. The difficulty of the task is related smoothly to the angular disparity of the two representations. This opens the possibility of pretesting each pair of figures at a small number of orientations, then interpolating to create the curve that describes the characteristics of that pair presented at any angular disparity. If the technique should prove workable, each pair of figures will provide a whole family of calibrated items. That means, in effect, a larger item pool for the same effort. It also means that the opportunity to create and administer within the test exactly the right variant of the item for the individual being tested.

A number of other issues could be raised but that's sufficient illustration. My projection for the future is that most of the issues will be resolved more or less to our satisfaction. Adaptive testing will become the norm in large standardized testing programs, particularly those that emphasize "academic" aptitude and achievement measurement. It will become so because it will provide accuracy, broad-range measurement, and efficiency in testing time. This savings in testing time will be important because we will want to measure more aspects of skills and abilities than we have been able to in the past, and we will need the time to do it.

Nonetheless, a caveat: When we consider the computer as the test-delivery vehicle, whether for adaptive testing or for any other kind of testing, a number of issues related to the comparability of scores across modes become salient. If the same test is given in paper and computer modes, is it really the "same" test? Some evidence suggests a need for caution. Several studies have found, for example, that there are differences between reading material presented on the computer screen and reading from paper copy—reading from the screen is about 25 percent slower (8,9). That might not be important at all. Perhaps if we just adjust norms or time limits for tests presented by computer, the difference will be inconsequential. Or perhaps when higher resolution screens are commonplace or typical examinees all have extensive experience in working with computers, the difference will go away. But it might be very important. Suppose there are not only main effects of mode of presentation, but also interactions with examinee characteristics. Some ponderous equity issues would have to be dealt with. Issues of construct validity could also arise if the cognitive processing of information is different between modes, either for everyone or from one group to another. Then, mode differences would mean differences in what abilities the tests measure, not just in the level of performance. This is an area in need of close attention, both to find what differences exist and to understand why.

Branching Tests

Let me turn to a second projection. Adaptive testing is just one way in which testing can employ branching to improve the quality of measurement. There are many other branching schemes made feasible by the information-management capabilities of the computer, and we can expect many tests of the future to take advantage of the possibilities.

One example is provided by problems of the "patient-management" type, which are popular in testing in the medical and allied health fields. These are complex simulations of real-world, problem-solving situations in which each decision made by the examinee creates a new situation to which he must react. Different examinees receive very different sequences of events depending on the appropriateness and timeliness of the decisions they have made. Such problems have many attractive features. One is verisimilitude—these problems look and feel more like the practice situation to which they are meant to predict than can a string of independent multiple-choice items. Another is richness—they can be scored in a variety of ways to reflect different aspects of performance. How accurate were the choices that were made? How efficient was the examinee in avoiding unnecessary steps? How much pain and suffering did the patient undergo in the course of treatment? And so on. And finally, they require the integration of what an examinee knows. Successful examinees must possess more than isolated items of information. They must be able to put together what they know into an effective course of action.

Such problems have been studied in paper-and-pencil form for some time. Christine McGuire and her colleagues at the University of Illinois have been developing and evaluating them for several decades (11, 12), and they are now stock-in-trade in many assessment programs. The computer is an ideal delivery vehicle, not only to manage the branching that is quite cumbersome in paper testing, but also to record the sequences of events that play an important role in scoring. We can anticipate that computerized versions of these problems will spread to a number of fields in which we need to know how effective an examinee is as a problem-solver, not just how large his repertoire of information is.

And yet, this is another area where major measurement research remains to be done. For example, a complex simulation may take an hour to complete—and yet it is only one behavior sample, limited to one problem situation—in a sense, only one test item. And just as with simpler tests, generalization from a test with only a handful of items is risky. Research is needed to find ways to improve the efficiency of information collection through such problems, so that a test can include a broad sample of problems and offer the best possible prospects of generalization to the domain of interest.

I can't leave this topic without alluding to another kind of branching test that, I believe, is a harbinger of new generation of tests. Garlie Forehand along with his colleagues at ETS and the College Board are conducting research and development on a diagnostic test of basic skills required for college work. Diagnostic testing is nothing new, at least in name; but the use of the computer as test administrator makes possible a much more powerful test. The test can be individualized, for example, to a degree not possible with paper testing. In each domain to be assessed, the student can be given a brief "challenge" test. If the student shows mastery, that domain is quickly abandoned for the next. If the student has problems, detailed "probes" are introduced to identify which component skills are the source of the difficulty. The result is that, very quickly, a profile of the individual's strengths and weaknesses is produced. Moreover, in some sequences, the identification of weaknesses can be made with such precision that the instruction needed to remedy them is self-evident. The computer can tell the student not only what went wrong, but how to do it right, and can print a page of exercises for practice.

This test presages the testing of the future in several respects. It is oriented toward guiding instruction for an individual, not toward providing a score that can be used to compare one student to another. And taking the test is, itself, a learning experience. As students are confronted with successive components of a complex problem, they are implicitly being given an analytic framework with which to tackle complex problems on their own.

Preparing for a generation of such tests will keep measurement researchers quite busy for some time to come. First, the effectiveness of the test will depend on how well we understand the structure of knowledge in a domain and how well we know what instructional intervention is appropriate, given a particular knowledge deficit. The cognitive and educational psychologists have given us beginnings in these areas, but only that, and we will need to look to them for increasingly deeper conceptualizations. Second, the efficiency and accuracy of measurement will depend, in part, on the development of new psychometrics models. One instance of this is the need for an efficient decision as to whether or not an examinee has mastered a content domain. My ETS colleagues are exploring the application of latent class theory to this problem. Another instance is the need for optimal branching from domain to domain. At the level of analysis that is desired, it is not practical to test each student on every domain—there's just not enough time available to do that. Models are needed to optimize the selection of domains for an individual, to get the most useful information within the constraints of the feasible.

Let me turn now to two ways in which tomorrow's tests will be different than today's. There isn't time for detail, so I'll just hit a few high spots.

Free-Response Testing

First, look for a decline in the hegemony of the multiple-choice item. Multiple choice is largely an artifact for the needs of large-scale standardized testing, and it's served us well in that context. But test users and test takers have never been fond of it, feeling that somehow it fails to get at what examinees really know and can do. And, to some degree, they have been right. Norman Frederiksen, Sybil Carlson, and I have conducted a series of studies with complex, ill-structured problems—problems like requiring an examinee to generate a set of alternative hypotheses to explain a sociological phenomenon (3, 15). We started with problems posed in free-response form: think of hypotheses and write them down. We then tried very hard to create multiple-choice and other machine-scorable versions of problems that would measure the same abilities; and we failed (14, 15). Generating ideas on your own, in complex situations in which you are not an expert, is just not the same thing as recognizing the best idea in a list someone else provides. And, I think, most would agree that generating ideas is one step closer than recognition to the kind of problem solving real people do in real-world situations.

What to do about free responses has been a problem, however. When we score our problems by hand, we invest about as much time in judging the quality of an examinee's answers as she spends in writing them; and that's not very practical for large-scale testing. We believe, though, that we are close to being able to administer and score these problems by computer. Short-term, the computer won't be terribly smart—we'll just feed it a list of key words and phrases to look for, and see if it can do as well as our human scorers in applying these. We'll probably have to keep human experts in the scoring process for some time. My guess is that the machine, again like most of our human scorers, will be able to deal with the large majority of protocols it encounters, but it will have to get help from someone with more expertise to cope with the remainder.

Long-term, we expect much smarter machines to be available. Sooner or later, expert systems with natural language-processing capability will be able to analyze freely written protocols and give us really intelligent scoring. In fact, we could almost have such systems today. An analysis recently completed by Roy Freedle (4) led to the conclusion that this kind of scoring is now feasible, if we restrict ourselves to problems somewhat less complex that the formulating hypotheses type. Today, the analysis is too expensive and too domain-specific to be practical; but wait for tomorrow.

Many Right Answers

My final suggestion for what tomorrow's test will look like is somewhat heretical. That is, look for a decline in the exclusive use of tests in which answers are scored simply "right" or "wrong." One basis for suggesting this comes from studies dealing with item types we have traditionally used in assessment—for example, item types used to determine vocabulary knowledge. When asked to produce an antonym for the word "frivolous," for example, many examinees show that they have partial knowledge. They may not be abl. to give you "conscientious" or "responsible" or any other of the ten or so fully acceptable antonyms for the word, but they can come up with "practical" or "studious" or some other word that is in the right neighborhood, even if the nuance isn't quite right. Should we say "sorry, that's wrong," or should we give partial credit for partially right answers? I have a little data that suggests that all-or-none scoring throws away useful information; a more reliable test score results from giving partial credit (13).

When we move beyond the old familiar item types to more complex ones—to ill-structured problems like that of generating alternative hypotheses—it's even clearer that right/wrong isn't sufficient. Complex problems often don't have one solution, but many. Some are more elegant or cogent or efficient than others, but those that are second-best are far from wrong. My friends who develop items for our testing programs shudder at the thought of having to justify and defend more than one acceptable answer; but as our questions become more interesting, that's a complication we'll have to live with in the interest of better measurement.

That's my list of some likely prospects for future testing. It's definitely incomplete, probably quite idiosyncratic, and almost surely wrong in major respects. But there's a saving grace in prognostication: When the future arrives, everyone is too busy dealing with it to look back to what you said and discover that you missed the boat. Meanwhile, I am sure of one thing: We're up to our eyeballs in possibilities for new, better, more useful ways of assessing than have been available in the past, and all of us interested in measurement have a very busy decade or two in front of us.

References

Cook, L.L., N.J. Dorans, D.R. Eignor, and N.S. Petersen. *An Assessment of the Relationship Between the Assumption of Undimensionality and the Quality of IRT True-Score Equating*. Princeton, NJ: Educational Testing Service, 1985.

Frederiksen, N., and W.C. Ward. "Measures for the Study of Creativity in Scientific Problem Solving," *Applied Psychological Measurement*, 2, (1978), 1-24.

Freedle, R.O. *A State of the Art Survey of Artificial Intelligence and Its Application to the Analysis and Production of Verbal Test Items*. Princeton, NJ: Educational Testing Service, 1984.

Green, B.F., R.D. Bock, L.G. Humphreys, R.L. Linn, and M.D. Reckase. *Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery*. Baltimore, MD: The Johns Hopkins University, 1982.

Green, B.F., R.D. Bock, R.L. Linn, F.M. Lord, and M.D. Reckase. *A Plan for Scaling the Computerized Adaptive ASVAB*. Baltimore, MD: The Johns Hopkins University, 1983.

Hambleton, R.K., ed. *Applications of Item Response Theory.* British Columbia: Educational Research Institute of British Columbia, 1983.

Heppner, F.H., J.G.T. Anderson, A.E. Farstrup, and N.H. Weiderman. "Reading Performance on a Standardized Test is Better from Print than from Computer Display," *Journal of Reading*, 28, (1985), 321-325.

Kruk, R.S., and P. Muter. "Reading of Continuous Text on Video Screens, *Human Factors*, 26, (1984), 339-345.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum, 1980.

McGuire, C.H. and D. Babbott. "Simulation Technique in the Measurement of Problem-Solving Skills," *Journal of Educational Measurement*, 4, (1967), 1-10.

Ward, W.C. "A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests," *Applied Psychological Measurement*, 6, (1982), 1-11.

Ward, W.C., S.B. Carlson, and E. Woisetschlager. *Ill-Structured Problems as Multiple-Choice Items.* Princeton, NJ: Educational Testing Service, 1983.

Ward, W.C., N. Frederiksen, and S.B. Carlson. "Construct Validity of Free-Response and Machine-Scorable Forms of a Test." *Journal of Educational Measurement*, 17, (1980), 11-29.

Weiss, D. *Robustness of Adaptive Testing to Error in Item Parameter Estimates and to Multidimensionality.*" Paper presented at the 1982 Item Response Theory and Computerized Adaptive Testing Conference, Minneapolis, 1982.

Technology Advances That May Change Test Design for the Future

DOROTHY K. DERINGER
Technical Assistant to the Manager of Educational Systems
Entry Systems Division
IBM *Corporation*

Megatrends author John Naisbitt, in his recent book, *The Year Ahead*, identifies ten important societal trends. Trend #4 is "Technology to Transform the Classroom." Naisbitt states, "The age of electronic education is upon us. In 1985 a mind-boggling array of educational software, interactive videodiscs, and classroom robots will be introduced. Personal computers will turn livingrooms into classrooms. Hundreds of thousands of adults will return to school via electronic university programs designed for home computers, cable television, or work-site educational extension programs."[1]

The majority of the papers at this conference have concentrated upon educational testing and research and development directions in testing. This paper focuses on the uses of technology in education, that major trend that Naisbitt identifies, and it discusses some of the educational changes that the use of technology stimulates. Finally, it suggests some areas in which the testing field can improve our understanding of the educational process in this technological era.

Technology has had and will continue to have an impact on education. However, if the technology is to improve education in the long term, two major challenges must be met. First, though many educators are knowledgeable about using computers in education, many more teachers and administrators need considerable education and training in using technology in education effectively. And parents need to become more knowledgeable, too. This need for education is an ongoing one. This paper concentrates on identifying technology trends, but the meaningful use of technology in education depends on a strong base of knowledgeable people. The second challenge is to integrate hardware and software into instructional and administrative systems that are both useful and easy to use. As in so many other application areas, our ability to design individual pieces of hardware and software is improving much more rapıdly than our ability to organize them into systems and to use them wisely. Successfully meeting these challenges will considerably enhance the ability of educators to use computers effectively.

Technology Trends Influencing Education

Two major technology trends that will influence applications in computing in the late 1980s and early 1990s are the increasing interconnection of computers through networks and the increasing sophistication of individual computers whether used on a network or standing alone. How might these technology trends influence education?

Some background information on typical school computer environments will help to assess the potential impact of these two major technology trends. The estimates of numbers of computers used in schools for instructional computing vary widely; however, the growth figures are impressive. Henry J. Becker of Johns Hopkins University reported in July of 1985 that there are one million computers in schools.[2] This is a considerable growth from the 52,000 computers and terminals counted by the U.S. Department of Education in 1980.[3] Though the individual numbers can vary widely, the trends show a rapid increase in the number of computers in school. The primary use of computers for instruction in 1983 reported by Dr. Becker was computer literacy with programming following as the second most frequent use. This may be changing. Dr. Becker is presently analyzing new major survey data, and his preliminary analysis indicates that the diversity of uses of computers is increasing in schools.[4] School computers are now most typically standalone computers. In secondary schools, they are frequently grouped in laboratories of 20 or more. Elementary schools tend to have computers distributed in the classrooms.

Approximately 40,000 computers are in use in school administration.[5] There is a long history of use of computers for activities such as payroll and school scheduling. But these applications are primarily single-task oriented. In larger districts, these may be done on a central computer with data carried to and from the computer via an automobile shuttle service. Tasks are performed in a batch mode. Smaller districts may have larger personal computers such as the IBM PC XT or PC AT and use these for applications such as attendance or payroll.

Home educational use of computing is an elusive concept. Nearly 20 percent of American families own computers.[6] Education is frequently listed as a leading reason for owning a home computer, but no comprehensive information is available on what education is actually taking place in the home using computers.

What are some of the more sophisticated schools doing? As schools become more sophisticated, some adopt instructional networks within classes. In Palm Beach County, kindergarten through 5th grade classes are using an instructional network under a federally supported project to teach English as a second language to Spanish-speaking children and Spanish to English-speaking children. The laboratory contains a PC XT as a student station and 15 PC Jrs. as student stations. According to their teacher, even the smallest children have no difficulty using the network. Other popular applications of in-class networks are for teaching programming. There is also some use of databases and outside information resources from standalone computers within the classroom.

In administration, there is a trend toward more sophisticated workstations, which have local data entry such as marksense card readers for attendance and grading functions. Some remote job entry systems access centrally located mainframe computers. And there is a high interest in networking and sharing of data bases, but the lack of appropriate software is inhibiting faster progress into advanced information-based systems by most school systems. There is considerable progress by individual districts such as that of Dr. Arvid Nelson, the Superintendent of Indian Springs School District in Justice, Illinois, whose school this year will have a computer in every teacher's classroom linked to a central computer for both instruction and administrative uses. However, there are as yet no comprehensive computer-based administrative systems operating in U.S. schools.

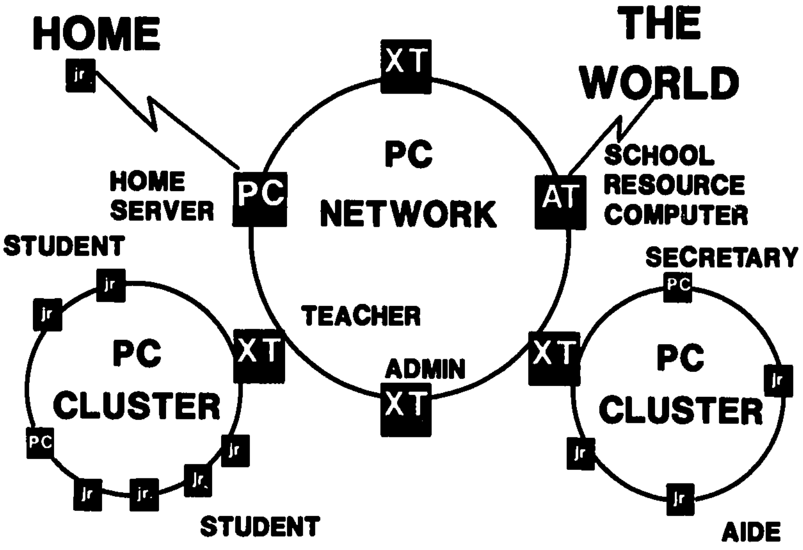
In the home, parents are using on-line encyclopedias and data bases for enrichment and homework for children. Again, the information is anecdotal and scarce about home education.

Therefore, schools in the mid-1980s have some experience and involvement in technology and a growing installed base of computers. How might schools be influenced by the trends of interconnectedness and sophisticated workstations?

Future Educational Environments

The trend toward increasing interconnectedness of computers is also remarked upon by Naisbitt. Trend #9 is "Computers and Telephones will Marry."[7] He states, "It takes no great leap of imagination to envision a time when teleports will be so thoroughly interconnected, much in the way cities are connected by complex and overlapping air travel route systems, as to create a worldwide network for the cost-effective and efficient transportation of information."[8]

One concept of how this trend might influence education is IBM's School of the Future demonstration that has been shown to over 400 decision makers from various schools, colleges, and universities. The School of the Future demonstrates what school might be like if every student, teacher, and administrator had his or her own computer in school and at home and all of these computers were networked together so that people could communicate with each other as well as share the use of programs and data. Figure 1 illustrates the network design of the School of the Future environment.



Access to such a powerful facility opens many new capabilities to schools. For the purposes of this paper, I would like to focus on just three. The first is that improved resources are available through communications links with the outside world. Several types of applications are possible: satellite delivery of courses where there is no local teacher in a speciality; delivery of coursewa-e via this link, with the school being billed only when the courseware is used; and correspondence of students with students in other locations, perhaps even in other countries.

The second major capability is that improved decision-making is possible through the collection and integration of information from a number of sources into comprehensive data bases. Administrative and instructional data bases could be linked to provide better analyses of the impact of investments of funds on measures of student progress. Reporting throughout the state's school administrative structures could take place faster and more efficiently with required reports compiled from data bases and passed over the network.

A third major capability is the exploration of the home as a place where students can be "in school." Students might be at home because they are temporarily ill or permanently unable to come to school, or they could be adults who, because of work requirements, could not come to the physical school building. Through the network, these home students would have access to the same information sources as school-based students, and they could correspond with their teachers and other students over the network. This network facility could put an end to the idea that school is a place surrounded by four walls.

Although many of these activities will take some time to be realized, one part of the concept of the School of the Future is taking place under the direction of Dr. Seymour Papert of the Massachusetts Institute of Technology. Dr. Papert has created a computer-intensive environment in a wing of the Hennigan School, a Boston public school. Two hundred students are using 100 networked computers in the fourth, fifth, and sixth grades. The students and the teachers are exploring music, language arts, and art, as well as technology, in this new environment.

A more subtle but profound result from this kind of environment is the student-to-student interaction and cooperation that take place when students have access to a powerful information resource. In Dr. Papert's project, 22 students as well as the Hennigan teachers learned about computers in the summer before school started. This summer session enabled the students to help their fellow students learn about computers and helped to create a more collegial environment in the school.

At the same time that more computers are being networked, the individual computer itself is becoming smarter. Sophisticated workstations have an important place in the business world, and researchers are experimenting with how they might be used in education. One of the most successful computer applications in the industrial world is computer-aided design (CAD) for engineering. These systems allow the construction of objects, their analysis, and rotation in real time. On the most sophisticated systems, the designed object, for example, an airplane wing, can be tested as well. These systems are being considered for educational uses.

Dr. Irwin Hoffman of the George Washington High School in Denver and winner of a Presidential Award for mathematics teaching has recently expressed concern that none of his students has experience in working in three dimensions. Solid geometry is rarely taught. Yet when students enter the world of work, particularly in areas related to manufacturing, work in three dimensions is necessary. Dr. Hoffman expects CAD systems to be useful in mathematics, chemistry, physics, art, and several other subjects. As CAD systems (which presently cost $100,000 and more) come down in price, their applications could become a key part of K-12 instruction. (Though some elementary CAD systems run on personal computers, they do not have real time rotational graphics and other more sophisticated features.)

Workstations are generally thought of as more powerful computers when compared to a personal computer. Dr. Andrea di Sessa at MIT is working on a workstation appropriate for students as young as third grade; it should also be useful for older students. The workstation will integrate text processing, database activities, and structured files as well as be a sophisticated computing environment. He is particularly concerned that it be useful to teachers to solve their own problems, not just something they learn in order to teach students.. The workstation has undergone limited testing with both teachers and young students. These sophisticated workstations promise to make it possible for even young students and teachers to do with their computers what only computer experts could do in the past.

Computers with a multiplicity of input/output devices will soon be commonplace. Videodisc, on the edge for so many years, is becoming widely accepted in industrial training. Indications are that it will also be used in K-12 education, as school districts try to make courseware that is more graphics-oriented.

In Benjamin Bloom's new book, *Developing Talent in Young People*, he states that of all the instructional strategies, the most effective is tutoring. He writes, "After forty years of intensive research on school learning in the United States as well as abroad, my major conclusion is: What any person in the world can learn, *almost* all persons can learn *if* provided with appropriate prior and current conditions of learning. . . . Under tutoring, the average student performs better than 98 percent of students taught by conventional group instruction, even though both groups of students performed at similar levels in terms of relevant aptitude and achievement before the instruction began."[9] If we can design expert systems to find oil and to diagnose illness, why not ones to teach arithmetic, problem solving, or debugging?

Research has been conducted on expert systems in just these areas of education. But up to this time, most of it has been experimental. An extended version of the InterLISP programming environment, the language in which many educational systems are written, is now available in a microcomputer environment. This will enable formative evaluation of some of these systems in the fi ild. Sleeman and Brown, in their book, *Intelligent Tutoring Systems*, predict the following developments in educational expert systems in the foreseeable future: a shift in emphasis to the cognitive and sociological aspects of these new kinds of learning environments; information now communicated by humans and not now written down will be embodied in expert systems. And they expect that an increasing amount of attention will be paid to the various aspects of student modelling and diagnosis of problems. These two trends of interconnectedness and sophisticated workstations could be thought about as an either/or situation. Either we will have increasingly sophisticated machines in education or we will increasingly emphasize the connectivity of machines. As in many other areas in life, perhaps the answer is that we will have both.

The New Technologies and Testing

There are several questions raised by these new educational environments that improvements in testing could help to answer. How can we test new educational objectives in this increasingly complex school environment? These new educational technologies offer rich experiences to students. If the student learns solid geometry in the dynamic, graphic world of a sophisticated workstation, how can we appropriately test his or her mastery of this knowledge with a paper and pencil test? If students continually work in teams or in small peer groups to accomplish class projects, why should we only test a student's solitary ability to solve problems? As students' daily experiences in school change, we need new testing tools to access their achievement.

How can we effectively provide feedback to teachers, administrators, and parents? As computers become more widely used in schools, we have the opportunity to provide more complete and immediate reports on student progress. Better information is needed at a variety of levels. Students need to understand how well they are mastering the subject at hand with a question or two, not a two-hour test. Administrators and teachers need better facilities to know how well each student is learning the curriculum, since these new tools will permit students to progress at their own rates through the material. As more learning takes place in the home, then parents can be informed of their child's accomplishment more frequently than once every report period. Because of the changes in schools noted above, fulfilling this need for better knowledge of progress is much more than creating data bases of existing test results. In fact, such an approach could be misleading if what we are testing is mastery of facts and details and what we are teaching is team problem-solving.

How can we respond to changing levels of expectations of the people who are using tests? As more schools design and administer sophisticated tests themselves, they become knowledgeable consumers. They understand more of the statistical underpinning of tests and their strengths and weaknesses. They also demand more rapid reporting of the results of nationally based tests, and they want more custom reports and analyses. These educated consumers are an opportunity for testing groups that can meet these demands; and involvement in testing at the local level provides an opportunity to improve feedback on student progress.

These technology trends of increased interconnection of computers and more sophisticated workstations offer opportunities to education, but they also create needs for better measurement tools to access educational progress. Dr. Andrew Molnar of the National Science Foundation has suggested that we may have a future with very big machines but very small ideas. In order for educators to establish the programs with big ideas that are powerful improvements to their systems, extensive work needs to be done within the school environment in thinking about what is needed and then planning for change. These educators need the support of testing professionals to evaluate and monitor their new educational directions.

Footnotes

1. Naisbitt, John, *The Year Ahead*, AMACOM, American Management Association, New York, NY 1984, p.23.

2. Becker, Henry J., "The Second National Survey of Instructional Uses of School Computers: a Preliminary Report," Center for the Social Organization of Schools, Johns Hopkins University, Baltimore, MD, July 1985, p. 1.

3. National Center for Education Statistics, *Student Use of Computers in Schools*, U.S. Department of Education, March 20, 1981.

4. Becker, *op. cit.*, p. 15.

5. Extrapolated from *K-12 Market for Microcomputers and Software,* Talmis, Inc., October, 1985, p. 70.

6. Software Access International in *The Wall Street Journal,* April 18, 1985, p. 1.

7. Naisbitt, *op. cit.* p. 52.

8. Naisbitt, *ibid*, p. 56.

9. Bloom, Benjamin S., *Developing Talent in Young People,* Ballantine Books, Inc., New York, NY, 1985, pp. 4-5.

References

Becker, Henry J., "The Second National Survey of Instructional Uses of School Computers: A Preliminary Report," Center for the Social Organization of Schools, Johns Hopkins University, Baltimore, MD. July 1985.

Bloom, Benjamin S., *Developing Talent in Young People*, Ballantine Books, Inc., New York, NY, 1985.

National Center for Education Statistics, *Student Use of Computers in Schools*, U.S. Department of Education, March 20, 1981.

K-12 Market for Microcomputers and Software, Talmis, Inc., October, 1985.

Naisbitt, John, *The Year Ahead*, AMACOM, American Management Association, New York, NY, 1984.

Sleeman, D., and J.S. Brown. *Intelligent Tutoring Systems*, Academic Press, Inc.: Orlando, 1982.

Software Access International in *The Wall Street Journal,* April 10, 1985.

The Integration of Instruction and Testing

Robert Glaser
Learning Research and Development Center
University of Pittsburgh

In this future-oriented conference, my prediction is that in the 21st century, testing, in relation to the educational process, will undergo significant redirection. The conditions necessitating this change have been accumulating over many years and now must be faced squarely. With each decade in the 20th century, we have increased the proportion of children attending schools; we have expanded both the range of social groups and the amount and kinds of education offered. Today's and the next century's challenge is to teach successfully all of the diverse children and youth who have become the active concern of our educational systems. New approaches to testing and instruction will be necessary to make it possible for everyone to meet standards of educational performance that—only three or four decades ago—were expected from a smaller segment of the population.

The then-acceptable route to educational attainment, in which high standards were achieved by selective testing, is no longer adequate. Dropping the reluctant or difficult learners or testing primarily to segregate them in programs that make few demands and offer few opportunities will not be a viable alternative. Simultaneously, we must assure that our most talented and most difficult students optimize their learning.

At present, tests (with the exception of the important informal assessments of the good classroom teacher) typically are not designed to guide the specifics of instruction. We use them primarily as indicators to signal general rises or declines in school performance. They serve as an index to the standards of schools, but they are not designed to shape progress effectively toward these standards and can do so only indirectly, if at all. In tl.e 21st century, tests and other forms of assessment will be valued for their ability to facilitate constructive adaptations of educational programs.

To accomplish this, students and teachers will need information that can inform instructional decision rather than just predict academic success or offer a percentile or grade-level index of relative standing and global attainment. The information required will be analogous to that used by an opera teacher or a swimming coach to guide the development of further competence and proficiency. Testing and teaching will be integral events. A test that monitors access to education only and does not monitor the progress of education will not be tolerated for either the slow or the quick learner. Relationships between test score information and the nature of competence in school subject matters will be empirically studied and conceptually better understood.

In comparison with our cu.:ent well-developed technology for aptitude measurement and our techniques for achievement test standardization, techniques for measuring the growth and development of human competence are not well developed. In the 20th century, a strong theory of achievement testing has not emerged. Lee Cronbach (1970) recognized this state of affairs 15 years ago when he wrote: "The design and construction of achievement test items has been given almost no scholarly attention. . . . Demands for content validity have suddenly become insistent, thanks to demands for genuine diagnosis and mastery testing, for national assessment and local accountability, for data that describe learners rather than rank them, (however,) the art of test construction has so far not coped very well with these demands" (pp. 509-511). Cronbach went on to say that some important ideas have been generated, like criterion-referenced testing, items as samples of operationally defined content universes, and analysis of information-processing requirements of tasks, but that much work lay ahead to clarify these ideas and turn them into useful procedures.

In recent years, the general outline of theoretical grounds for forms of assessment that can assist educators in monitoring the characteristics of new learning and attained levels of ability has emerged. There is a wide recognition of the need to ascertain the critical differences between successful and unsuccessful student performance by appraising the structures of knowledge and cognitive processes that reveal degrees of competence in a field of study. The design of measurement techniques that can guide instruction will be based on the now accumulating studies of learning that identify the performance components that facilitate or interfere with the eventual attainment of higher levels of achievement. In essence, this is the theme of my remarks: that the measurement of achievement should rely on our knowledge of learning and of the course of acquisition of competence in the subject matters that we teach. In the near future we should be able to develop assessments of learning that are more indicative of competence than tests with which we are now familiar.

The usual forms of achievement test scores do not provide the level of detail nece:sary for making appropriate instructional decisions. An array of subject-matter subtests differing in difficulty is not enough (Linn, 1983). Sources of difficulty need to be identified for specific problems in learning and performance. Tests also should permit learners to demonstrate the limits of their knowledge and the degree of their expertise. The construction of tests that are diagnostic of different levels of competence is a difficult task, but recent advances in the psychology of subject-matter competence and research on the functional differences between experts and novices in various fields are good starting points for framing the theories that should underlie achievement measurement.

From this perspective, consider our customary practices. It has always been startling to me that most of the technology of testing has been designed to occur after test items are constructed. The analysis of item difficulty, discrimination indices, scaling and norming procedures, and the analysis of test dimensions and factorial composition take place once the item is written. In contrast, in the next century, sustained attention to theory will be required before and during item design. We will rely on what we know about the cognitive properties of acquired proficiency, and the structures and processes that develop as individuals move from beginning to advanced learners. The assessment of achievement will be integrally tied to the study of the nature of learning. Modern learning theory is taking on the characteristics of a developmental psychology of performance changes—the study of changes that occur as knowledge and complex cognitive strategies are acquired. In the future, achievement measurement will be designed to assess these performance changes. It will be cast in developmental terms to identify attainment at various levels of acquisition, emphasizing not only content considerations but structural and process considerations involved in sources of difficulty and in facilitators of the growth of competence (Messick, 1984).

I am encouraged to make this prediction about the future of achievement testing because a marked change is taking place in our knowledge and theories of human learning and intelligence. In the course of this century, theories of psychological measurement have focused on the testing of general processes—general forms of intelligence (verbal, numerical, and spatial) and on general aptitudes of various kinds that showed correlational relationships to overall success in school and in other forms of learning. Similarly, the study of learning also has sought for evidence of general processes and general conditions of learning—pervasive laws that influence all kinds of learning, such as forms of conditioning, the nature of reinforcement and feedback as a consequence of learning, and the conditions of practice such as massed and spaced learning. Such broad-based analyses, though they helped in explicating important principles of learning, could only assist learning in a general way, on the basis of rather weak heuristics, such as categorizing classes of learning deficits that impede ability to learn.

In contrast, in recent years the study of human performance has become more oriented toward stuuying the specific types of knowledge and skill that people acquire and face in their lives. This change has led to considerable emphasis on learning in the knowledge-rich domains that correspond to the academic disciplines and the subject matters of schooling. This new emphasis will make it feasible to identify strengths and weaknesses involved in performing academic tasks. Rather than attempting to identify a general underlying deficit, we will concentrate more precisely on helping the learner recognize incomplete or partial knowledge that can become a focus for more direct instructional attention (Brown & Campione, 1984, in press).

Two advances in the study of human cognition are particularly noteworthy here. One is the information-processing analyses of the performances that contribute to proficiency in academic tasks. The other is the increased understanding of the nature of competent performance that has resulted from study of the characteristics of experts and novices in various domains of human endeavor. In the analysis of school tasks, elementary arithmetic and mathematics provide a good example. Progress has been made in mapping the development of children's grasp of the principles that underlie counting skill and their understanding of the concept of numbers and numerical reasoning (Geldman & Gallistel, 1978; Greeno, Riley, & Gelman, 1984), of the acquisition of arithmetic facts (Ashcraft, 1982; Siegler & Shrager, 1984), of knowledge and tactics for solving arithmetic word problems (Kintsch & Greeno, 1985; Riley, Greeno, & Heller, 1983), and of principles underlying place-value notation that is basic to computational skill (Resnick, 1982, 1984). These efforts and work on the diagnosis and categorization of error patterns in arithmetic performance (Brown et al.) will provide a basis for informed diagnosis of a child's understanding or misunderstanding in early mathematics learning. It will become easier to identify the incomplete knowledge and procedure and incomplete conceptual understanding (Resnick, 1984) that contribute to weak performance and that can be remedied in the course of instruction. We will be able to appraise the knowledge that reveals degrees of competence and that determines functional differences between superficial and more lasting achievement.

Let me turn now to several ideas for "learning assessment"—a term that might be better used than "tests." These ideas, which I will consider in several areas, are: the analysis of rules of performance, assessment of prior knowledge, the coordination of basic and advanced performance, and the nature of competence and expertise.

Analysis of Rules of Performance

One technique of learning assessment will be the analysis of task performances in a way that mimics an important skill of teaching, that is, the ability to synthesize from a student's performance an accurate picture of misconceptions that lead to error or of attainment that can lead to new learning. This task goes deeper than identifying incorrect or correct answers and pointing them out to the student and the teacher. Rather, it attempts to identify the nature of the concept of the rule that the student is employing in some systematic way. The assumption is that in most cases the student's behavior is not random or carelcss, but is driven by some underlying misconception or by incomplete knowledge.

Such diagnostic procedures are based on the decomposition of a complex skill into component procedures that contain elements of the underlying ability. Misconceptions that result from incorrect implementation of the various component skills are identified through a student's patterns of error on a set of tasks. From an apparently confusing array of student responses, patterned scoring procedures have been able to identify systematic sources of error. For example, studies of errors in subtraction (Brown & Burton, 1978) illustrate the point well. In some cases the student subtracts the smaller digit in each column from the larger digit, regardless of which is on top. Or when the student needs to borrow, he or she adds ten to the top digit of the current column without subtracting one from the next column on the left; or when borrowing from a column whose top digit is zero, the student writes nine but does not continue borrowing from the column to the left of the zero. Students' problems in working with fractions (Tatsuoka, 1981) show similar systematicity. Often the student converts mixed numbers to the wrong improper fractions but uses the correct combination rule or omits the whole number after using the correct procedure on fraction parts.

Similarly, in writing, a student puts in a comma every time an *and* occurs, rather than when the *and* introduces an independent clause; or a student may connect any relative clause that comes at the end of a sentence to the independent clause before it with the phrase "in which"; or the student dctermines the boundaries of sentences by the erroneous rule, "Put a period at long pauses" (Hull, in press; Shaughnessy, 1977a, 1977b).

Scoring systems that identity systematic bugs of this kind have important implications for testing, because students are evaluated not on the basis of the number of errors on their tests, but rather on the basis of the misconceptions or incomplete rules that influence their performance. Diagnosing performance in this way links testing to instruction. It encourages the teacher to see that the apparently random, careless, or lazy behavior of a student is frequently rooted in a complex and logical process of thought toward which teaching can be directed. A diagnostic testing emphasis of this kind is useful and impressive to teachers; they view it as an important aspect of their own skills and as a way of respecting the systematic intelligence of their students.

Assessment of Prior Knowledge

Consider another aspect of student performance that might be assessed to assist instruction. It is well known that comprehension and learning are based on current beliefs and that students attempt to understand and think about new information in terms of what they already know. This being the case, then it seems best to base teaching on the forms of knowledge that they currently hold. High levels of learning and understanding can be fostered by insuring contact between new information and the student's prior knowledge, which then can be restructured through instruction. The possible benefits of assessment of this kind have been indicated by studies in various subject areas, particularly by research in science education.

In science, the information with which students enter classrooms is based upon intuitive theories derived from prior experiences, from the perspective of common-sense interpretations of scientific phenomena. Common misconceptions are prevalent in students' beliefs about velocity and acceleration, free fall, electric circuits, photosynthesis, etc. These informal theories are not readily abandoned, and they frequently come up against scientific principles that are counter-intuitive and not easily assimilated to students' current notions. As science education researchers point out, "When a student's naive beliefs are not addressed, instruction may only serve to provide the student...with new terminology for expressing his erroneous beliefs" (McCloskey, Caramazza, & Green, 1980, p. 1141). If learning entails restructuring or replacing of these ideas, then it is not enough to assess whether or not the student knows the science information that was taught—one must also assess what the beginning student believes as a basis for instruction (Messick, 1984). Thus, we point to another important aspect of performance diagnosis that is relevant to the integration of learning and instruction.

Coordination of Basic and Advanced Skills

Consider now the coordination of basic skills and advanced performance. Studies of competent performance have made it clear that human ability to perform many attention-demanding tasks is rather limited. If the simultaneous processing of the many tasks that make up a complex activity require conscious attention, then difficulties arise because attention must be switched from one task to the other. However, if performance of some of the tasks becomes sufficiently automated through practice and requires little conscious attention, then effort can be devoted to other, frequently higher level ones.

This orchestration of task components has been of special interest in the study of reading, particularly in investigations of the relationships between word-level reading skills and advanced processes of comprehension. A reader's attention may vacillate between the decoding skills of recognizing words and the skills of comprehension that integrate text ideas into memory. Shifts in attention are apparent in the beginning reader, who alternately concentrates on sounding out a word and then on considering what the word means in the context of what is being read. Although these component processes may work well when tested separately, they may not be efficient enough to work together. Because attention to each process takes time, slowness of a component process in interaction with other processes can lead to a breakdown in overall proficiency (Perfetti & Lesgold, 1979). Low levels of reading performance often reflect the interfering effects of slow, inefficient word decoding on the execution of higher level comprehension tasks.

Such interference effects between the component processes of a complex performance have important implications for learning assessment. Certain processes need to attain a certain level of efficiency so that other processes can be carried out simultaneously and in a coordinated manner. Hence, to optimize the success of learning where such coordination is important, it should be useful to assess the level of basic skill efficiency that is required to minimize interfering effects with higher level processes. The important index of performance is not whether the two processes can be carried out independently, but whether proficiency has reached a point where one process facilitates another. This suggests devising methods for diagnosing competence in basic skills in ways that indicate their success in freeing attention for advanced levels of achievement.

The Nature of Competence

Let's turn to yet another aspect of human performance that could influence learning assessment. Over the past 15 years, developments in cognitive psychology and artificial intelligence have spurred increasing investigations of the nature of proficiency and high levels of competence. The central questions are how knowledge becomes organized and how the processes that use this knowledge develop over long periods of learning and experience. Just what are the factors that enable expertise and the amazing efficiency, judgment, and problem-solving abilities shown by individuals who are very good at what they do?

A great deal of effort is now being devoted to understanding the cognitive structures and abilities of the skilled performer and analyzing the processes involved in the transformation of novice learners into increasingly expert individuals. As we gain understanding of the nature of competence, we should begin to see possibilities for advances in techniques for assessing attainment at various levels of proficiency.

One of the most salient and consistent findings of this research is that proficient individuals develop organizations of knowledge that enable them to perceive rapidly meaningful patterns in their memory. This allows them to form representations of problems that lead to appropriate, meaningful action. Novices, on the other hand, represent problems in qualitatively different and superficial ways that make problem situations more difficult to solve. Adept pattern recognition and problem representation are indices of competence which might be included in assessment of developing expertise.

There are many evidences of this phenomenon. The classic work was carried out in studies of skill in chess (Chase & Simon, 1983; de Groot, 1965, 1966; Simon & Chase, 1973). The striking difference between chess experts and weaker players is not the experts' superior general intelligence or their superior ability to keep all the moves of a game in memory, but rather their ability to recognize patterns quickly on the chessboard for their meaningful strategic implications. The estimated size of a chess expert's pattern vocabulary is roughly 50,000 configurations, in contrast to the thousand patterns of an average player and the very few patterns of a novice. The chess expert is a superior recognizer, rather than a deeper thinker. This explains how they are able to play many individuals at one time; for the most part they rely on pattern recognition abilities (so-called chess intuition) to generate potentially good moves (Chase & Chi, 1981).

Analogous abilities are found in those who perform well in the subject-matter domains of schooling. Investigations of students solving problems of elementary physics have studied the phenomenon of physical intuition, which is much like the chess expert's intuition (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Rees, 1982; Larkin, Mc Dermott, Simon, & Simon, 1980; Simon & Simon, 1978). Good solutions are associated with the perception of significant patterns. In contrasting novices and graduate students, it seems clear that the proficient performer rapidly perceives the deep central principles that underlie the problem. His or her knowledge is organized around central principles of physics that inform solution procedures, whereas, the knowledge and perceptions of the novice are organized around the surface features and physical description of the entities in a problem. Upon looking at a problem, the proficient individual says, "That's a Newton's Second Law Problem." The less proficient individual says, "It is a pulley problem, or an inclined plane problem." Both students may solve the problem, but the way in which the problem is initially perceived and represented determines the selection of problem-solving procedures, which results in differences in efficiency and the ability to handle difficult situations.

Similar results have been obtained in other subject-matter areas. For example, proficient students in high school and college algebra develop rapid perceptions of the semantic structure of algebra problems (Hinsley, Hayes, & Simon, 1978). After reading the first sentence or two of a problem and before carrying out steps toward a solution, they quickly categorize the problem as belonging to a class of problems—a triangle problem or a ratio problem or a river current problem. They say, "Oh, that's a triangle problem and it's solved by using the Pythagorean theorem." For these students, problem categories rapidly trigger appropriate solutions in memory. This ability of proficient individuals suggests a possibility for learning assessment. We should be able to develop procedures to test problem perception, and to observe the forms in which it occurs in the course of developing competence.

Toward Principles for the Measurement of Achievement

Let me now attempt to summarize the ideas I have described in a form that could suggest a framework for the design of learning assessment instruments—instruments for determining levels of knowledge and skill that are attained in the course of instruction. These ideas should be considered as a basis for test-item construction coordinate with or prior to psychometric considerations. As I have tried to show, achievement measurement can now begin to be grounded by modern cognitive theory that conceives of learning as the acquisition of knowledge and competence. At various stages of learning, there exist different integrations of knowledge, different degrees of procedural skill, differences in rapid access to memory and in representation of the tasks one is to perform. These different indices signal advancing expertise or possible blockages in the course of learning (Glaser, Lesgold & Lajoie).

As I envision it, achievement measurement theory based on this kind of knowledge is at an early stage. Many of the essential ideas are yet to be worked out, but enough work has been done to indicate the shape of a guiding framework. A tentative set of "dimensions" can be proposed in an effort to characterize components of developing proficiency that might underlie the assessment of achievement. These dimensions are certainly covered to some extent in traditional forms of achievement assessment, but also may require new methods of measurement. In any case, whether or not items take on new characteristics, they will be informed by a theoretical base that will underlie more systematic rationales for interpretations of the meaning of test scores. Consider as a representative sample the following four dimensions: Principled performance and active knowledge, theory change, problem representation, and automaticity to reduce attentional demands.

1. Principled performance and active knowledge. As competence is attained, elements of knowledge and components of skill become increasingly interconnected and rule-based, so that individuals access rules for their performance rather than fragmentary pieces of information. This is apparent in various subject-matter domains; a beginner's knowledge consists of incomplete definitions, erroneous rules, and superficial understandings; but from the pattern of a student's test responses, systemati ties of performance can be determined to explain behavior. The diagnosis of these principles of performance becomes a candidate dimension for the assessment of achievement that can inform instruction.

Related to this point is the suggestion of learning theory that the course of acquisition of knowledge proceeds from an initial accumulation of information in declarative form to a form that is more active and useful. In essence, we can know a principle or a rule or an item of specialized vocabulary without knowing initially the conditions under which it is to be used effectively. Studies of the difference between experts and novices indicate that beginners may have requisite knowledge, but this knowledge is not bou..d to the conditions of applicability. When knowledge is accessed by experts, it is associated with indications of how and when it is to be appropriately used. Assessments of the development of achievement in an area of knowledge through this progression from declarative to active information can be a useful measure of competence. Test items can be composed of two elements—information that needs to be known and information about the conditions under which use of this knowledge is appropriate.

2. Theory change. Learning takes place on the basis of existing mental models and theories held by students which either enhance or retard learning. With appropriate instruction, students test, evaluate, and modify their current theories on the basis of new information, and, as a result, develop new schema that facilitate more advanced thinking. However, as I have indicated, students can hold naive theories at the beginning of a course that make learning difficult. Even after instruction, these naive theories may persist. Although students have learned, in some mechanical fashion, to solve problems, they may have little understanding. Thus, theories of knowledge become a target for assessment. The characteristics of a theory held by a student might indicate whether it is a tractable theory, amenable to change under certain instructional conditions, or whether the theory held is more intractable, resulting in learning difficulties that require more thorough instruction.

The nature of students' theories adds an important dimension to achievement assessment. They can be measured to determine not only the levels of task complexity that a student is capable of handling, but also the level of thinking demanded by the requirements of school curricula. The demands of school problem-solving tasks may require understanding less sophisticated than the teacher envisions. This discrepancy poses a dilemma, because when proficiency is assessed, the student will have acquired and retained the model required by actual performance, not the one prescribed by stated teaching objectives.

3. Problem representation. It is now known that novices recognize the surface features of a problem or task situation and more proficient individuals go beyond surface features and identify inferences or principles that subsume the surface structure. This growing ability for fast recognition of underlying principles indicates developing achievement and could be assessed by appropriate pattern recognition tasks in verbal and graphic situations. Since certain forms of representation appear to be highly correlated with the ability to carry out the steps of a problem solution, test items might concentrate on assessing the initial understanding that is displayed by problem representation, rather than emphasizing the details of arriving at the correct answer.

4. Automaticity to reduce attentional demands. As I have indicated, investigations of competence make it evident that human ability to perform competing, attention-demanding tasks is limited. When subtasks of a complex activity simultaneously require attention, efficiency of the overall task is affected. This fact has particular implications in diagnostic assessment of the interaction between components of performance. Although component processes may work well when tested separately, they may not be efficient enough to work together. If a task demands an orchestration of skills, then measurement procedures should be able to diagnose inefficiencies. A criterion for assessment becomes the level of automaticity required for subprocesses to have minimal interference effects and to have progressed to a point where they can facilitate total performance and new learning.

Conclusion

To conclude, achievement testing, as I have defined it, is a method of indexing stages of competence through indicators of the development of knowledge, skill, and cognitive process. These indicators reveal stages of performance that have been attained and that provide a basis for further learning. They also show forms of error and misconceptions that result in inefficient and incomplete performances which need instructional attention. Achievement measurement defined in this way needs to be informed by theories of the acquisition of subject-matter knowledge, and by a focus on various dimensions of proficiency, such as rules of performance, automaticity, forms of representation, and procedural efficiencies that can index the growth and development of competence.

I have speculated on some possible dimensions, and further research is required, but we have grounds for anticipating important advances. It is likely that new theoretical sophistication will be brought to achievement measurement. In the 21st century, learning assessments will not provide merely a score, a label, a grade level, or a percentile. Rather, we will have also "instructional scoring" that indicates to the student and assists the teacher's judgment in making apparent the requirements for increasing competence.

References

Ashcraft, M.H. (1982). "The development of mental arithmetic: A chronome'.ric approach," *Developmental Review*, 2, 213-236.

Brown, A. L., & J.C. Campione (1984). "Three faces of transfer: Implications for early competence, individual differences, and instruction." In M. Lamb, A. Brown, & B. Rogoff (Eds.), *Advances in Developmental Psychology*, (Vol. 3, pp. 143-192) Hillsdale, NJ: Erlbaum.

Brown, A.L., & J.C. Campione (in press). "Psychological theory and the study of learning disabilities," *American Psychologist.*

Brown, J.S., & R.R. Burton (1978). "Diagnostic models for procedural bugs in basic mathematics," *Cognitive Science*, 2, 155-192.

Brown, J.S., & K. VanLehn (1980). "Repair theory: A generative theory of bugs in procedural skills," *Cognitive Science*, 4, 379-426.

Chase, W.G., & M.T.H. Chi (1981). "Cognitive skill: Implications for spatial skill in large-scale environments." In J. Harvey (Ed.), *Cognition, Social Behavior, and the Environment.* Hillsdale, NJ: Erlbaum.

Chase, W.G., & H.A. Simon (1973). "Perception in chess," *Cognitive Psychology*, 1, 55-81.

Chi, M.T.J., P.F. Feltovich, & R. Glaser (1981). "Categorization and representation of physics problems by experts and novices," *Cognitive Science*, 5, 121-152.

Chit, M.T.H., R. Glaser, & E. Rees (1982). "Expertise in problem solving." In R. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence.* Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1970). [Review of *On the Theory of Achievement Test Items*]. *Psychometrika*, 35, 509-511.

de Groot, A. (1965). *Thought and Choice in Chess*. The Hague: Mouton.

de Groot, A. (1966). "Perception and memory versus thought: Some old ideas and recent findings." In B. Kleinmuntz (Ed.), *Problem Solving*. New York: Wiley.

Gelman, R., & C.R. Gallistel, (1978). *The Child's Understanding of Numbers*. Cambridge: Harvard University Press.

Glaser, R., A.M. Lesgold, & S. Lajoie (in press). "Toward a cognitive theory for the measurement of achievement." In R.R. Ronning, J. Glover, J.C. Conoley, & J.C. Witt (Eds.), *The Influence of Cognitive Psychology on Testing and Measurement.* Hillsdale, NJ: Erlbaum.

Greeno, J.G., M.S. Riley, & R. Gelman (1984). "Conceptual competence and children's counting," *Cognitive Psychology*, *16*, 94-143.

Hinsley, D.A., J.R. Hayes, & H.A. Simon (1978). "From words to equations: Meaning and representation in algebra word problems." In P.A. Carpenter & M.A. Just (Eds.), *Cognitive Processes in Comprehension.* Hillsdale, NJ: Erlbaum.

Kintsch, W., & J.G. Greeno (1985). "Understanding and solving word arithmetic problems," *Psychological Review, 92*, 109-129.

Linn, R.L. (1983). "Testing and instruction: Links and distinctions," *Journal of Educational Measurement* 20(2), 179-189.

Larkin, J., J. McDermott, D.P. Simon, & H.A. Simon (1980). "Expert and novice performance in solving physics problems," *Science, 208*, 1335-1342.

McCloskey, M., A. Caramazza, & B. Green (1980). "Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects," *Science, 210*, 1139-1141.

Messick, S. (1984). "The psychology of educational measurement," *Journal of Educational Measurement, 21* (3), 215-237.

Perfetti, C.A., & A.M. Lesgold (1979). "Coding and comprehension in skilled reading." In L.B. Resnick & P. Weaver (Eds.), *Theory and Practice of Early Reading*. Hillsdale, NJ: Erlbaum.

Resnick, L.B. (1982). "Syntax and semantics in learning to subtract." In T. Carpenter, J. Moser, & T. Romberg (Eds.), *Addition and Subtraction: A cognitive perspective* (pp. 136-155). Hillsdale, NJ: Erlbaum.

Resnick, L.B. (1984). "Beyond error analysis: The role of understanding in elementary school arithmetic." In H. Cheek (Ed.), *Diagnostic and Prescriptive Mathematics: Issues, ideas, and insight* (pp. 181-205). Kent, OH: Research Council for Diagnostic and Prescriptive Mathematics.

Riley, M.S., J.G. Greeno, & J.I. Heller (1983). "Development of children's problem-solving abilities in arithmetic." In H.P. Ginsburg (Ed.), *The Development of Mathematical Thinking* (pp. 153-196). New York: Academic Press.

Shaughnessy, M. (1977a). *Errors and Expectations*. New York: Oxford University Press.

Shaughnessy, M. (1977b). "Some needed research on writing," *College Composition and Communication, 28*, 317-321.

Simon, H.A., & W.G. Chase (1973). "Skill in chess," *American Scientist, 61*, 394-403.

Siegler, R.S., & J. Shrager (1984). "Strategy choices in addition and subtraction: How do children know what to do." In C. Sophian (Ed.), *Origins of Cognitive Skills*. Hillsdale, NJ: Erlbaum.

Simon, D.P., & H.A. Simon (1978). "Individual differences in solving physics problems." In R. Siegler (Ed.), *Children's Thinking: What develops?* Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K. (1981, January). *Diagnosing Cognitive Errors: Statistical pattern classification and recognition approach* (Research Report 85-1-ONR). Urbana-Champaign: University of Illinois at Urbana-Champaign.

Redirecting a School District Based on the Measurement of Learning Through Examinations

RICHARD C. WALLACE, JR.
Superintendent, Pittsburgh Public Schools

America:1 school children are the most tested in the world and the least examined. (Resnick & Resnick, 1985, p. 17)

Testing school children in America is a highly valued part of schooling. Parents expect that tests will be given to determine the annual progress of their children and to compare their children with national standards. Administrators and board members want tests administered so they can use the results to compare their schools to national norms and make internal comparisons among schools and plan school improvement programs. Resnick (1981) contends that the public's belief in the value of testing stems from: a) the long-standing cultural need to make the most efficient use of human resources; b) the desire to reward talent regardless of social class origin; and c) the need to establish standards that somewhat limit the autonomy of local school districts.

A distinction should be made between tests and examinations. Standardized achievement tests are viewed as loosely linked to curriculum goals established by a school or school district. On the other hand, examinations are perceived to be closely matched to a school district's learning outcomes (Resnick & Resnick, 1985). Tests are usually machine-scored and contain predominantly multiple-choice type items. Examinations contain both short- and long-answer essay questions in addition to items in which responses are provided. Because of the essay requirement, the author contends that examinations provide students with a greater opportunity to demonstrate their learning accomplishments.

The subject of this paper will be Pittsburgh's Syllabus Driven Examination Project (SDEP). SDEP is a direct response to the challenge to improve the quality of education in American schools and the critical-thinking ability of students. It is also a direct outgrowth of the Monitoring Achievement in Pittsburgh (MAP) program to be described later. SDEP will provide all secondary students with a syllabus for each major academic course for each year of high school. Examinations, to be administered on a quarterly basis, will include multiple-choice test items and short- and long-answer essay questions. Students will be provided with sample examination questions and will be given the opportunity to practice the behaviors required by the exams. Quarterly examinations will provide students and parents with knowledge of student progress toward the goals of the school district.

The author believes that the beginning of the 21st century will find a wide-scale use of examinations in American schools. These examinations will build upon the best features of European examination systems and the traditions of the New York State Regents Examinations. However, the author believes that, to be maximally useful, examinations should be developed by local education agencies (LEAS).

In this paper I will present the rationale for SDEP and the results of the feasibility study. I will explore the European antecedents of SDEP and will describe the MAP programs in reading, writing, and critical thinking as prerequisites to SDEP. Finally, I will propose a new participatory role for the American testing industry, a role that will involve close collaboration with LEAS to develop, implement, and monitor the use of examinations in the 21st century.

European Examination Systems—Implications for American Schools

Since the early 19th century, European countries have used examinations administered by external agencies (e.g., ministry of education) as a means of setting standards and controlling the entry or exit of students to and from the educational system. Madaus and Greaney (1981) point out that examinations serve political and social as well as educational purposes. They trace the history of the Irish Primary Certificate and indicate that the Irish Parliament imposed examinations to establish standards for primary schools in Ireland. Madaus and McDonagh (1979) trace the growth of the current English examination system. They point out that socioeconomic conditions provided the impetus to open the universities in England to all academically able students without regard to their social class origin. Examinations now control entry to higher education in England.

In most European countries, past exams are made public. The tradition of past exams has a powerful influence on students, teachers, and curricula (Madaus and McDonagh, 1979). The exams serve several purposes. First, they provide valuable information about the content areas that are judged to be most important. Second, they provide students with models of examination questions that are likely to be encountered in future exams. Third, they provide the prudent teacher with very clear signals about what should be taught, how it should be taught, and how it should be measured. Past exams provide a *de facto* curriculum for schools; they also provide teachers with the tools to prepare students for exams.

The negative aspects of European examination systems relate to the stress induced and the tendency to use exam results as the single criterion to make important decisions about the educational and life future of the young. Both teachers and students endure great stress in preparing for exams. Teachers obviously teach to the exam, and this condition may have the undesirable side effect of narrowing the curriculum. Students cram for the exams, since important decisions about their educational or occupational future hinge on performance on the examination. Such cramming does not always contribute to learning that lasts. The use of the single criterion of exam performance to make critical judgments about the student's future is, in the author's judgment, placing far too much importance on a single indicator.

There are, however, positive aspects to European examination systems. They include the presence of a syllabus for each examination, sample questions provided to students for practice, and copies of prior examinations as models for expected learning outcomes. Perhaps the most powerful and useful aspect of exams are the standards that they contribute.

Advanced Placement Exams

The Advanced Placement (AP) program of the College Board provides school districts in the United States with an approximation of the syllabus-driven examination that the author will describe. The AP program provides teachers and school districts with the opportunity to offer beginning college-level courses and exams to high-ability students. If students pass the exams, they may be granted advanced standing in cooperating colleges and universities. Students are provided with a course description and examples of multiple-choice and essay questions typical of those found on the AP exams. Local districts, in turn, develop and implement courses of study that reflect the AP goals.

Since 1982, the essay portion or "free-response" section of the exams for social science and English has included a Document-Based Question

(DBQ). The DBQ offers reading material that presents several points of view related to a theme or issue. Students read the documents and then respond to an essay question that requires them to analyze the documents and synthesize a response. Students are expected to cite evidence from the text provided and other sources as they respond to the essay question. The AP exams assess both the general knowledge of students and their ability to think critically. The essay portion of the AP exam is graded by college and secondary teachers trained for that purpose.

Both the European examination systems and the Advanced Placement exams have important components that will be used in developing the SDEP in Pittsburgh. The use of a syllabus that provides students with a brief statement of the content of a course and a set of expected outcomes will be an important part of the SDEP system. Also, the provision of sample test and essay questions will provide students with examples of how knowledge acquisition will be measured. The Document-Based Question of the AP program provides students with the expectation that they will analyze information from multiple sources and synthesize a coherent response to an essay question. These features will be incorporated into the SDEP.

Pittsburgh's Syllabus Driven Examination Project

How will the SDEP differ from the European examination systems? How will it avoid the negative consequences? How will it differ from the AP exams? First, the SDEP is designed to be administered to all high school students, not just the academically able. Second, SDEP will be administered on a quarterly basis in all major subjects in every year of high school, rather than one time only at the end of secondary school. Third, the exams will be used to gauge progress of students toward the learning outcomes that have been established in the syllabus rather than to make only pass/fail judgments. Fourth, the exams will not be the sole criterion in determining whether a student has passed a course or performed adequately for promotion or graduation. The exams will directly influence the grade that a student earns in a course. However, performance on homework assignments, class participation in discussions, the quality of term papers, performance on teacher-made tests, and other relevant indices will be used as well to determine the final grade in the course. Therefore, the consequences of failing an examination will not be so immutable as is the case with European systems.

The primary and most positive reason for the introduction of the SDEP is to raise academic standards for all students. The exams and syllabi are expected to influence the quality and type of classroom instruction. They will place a premium on the ability of students to analyze and synthesize the knowledge that they gain, as well as to express that knowledge in response to a challenging essay question.

SDEP—the Process of Development

In the spring of 1985, the Pittsburgh Public Schools embarked on a feasibility test of the SDEP. The district contracted with Carnegie-Mellon University faculty, led by Professors Daniel Resnick and Pet^r Stearns; Stearns directed the project. The development and pilot testing took place at the Schenley High School Teacher Center, the district's site for the revitalization ɔf its secondary school faculty (Wallace, et al., 1984).

The SDEP development proceeded as follows: The project staff and the teachers at Schenley engaged in a review of the most important learning outcomes in a tenth-grade course, World Cultures. Throughout that review, the participants were required to probe deeply into the subject matter, examine the existing curriculum, and achieve consensus about what learning outcomes were most valuable to students. Concepts, generalizations, and themes were sought as outcomes as opposed to names, dates, and facts. The next task for the staff and teachers was to develop actual items and questions for students. They were constructed to reflect the high-level discussions to take place in the classroom.

Next, the project staff examined existing textbooks and curriculum guides to judge their utility in achieving the outcomes and the exams that were developed. Typically, textbooks were found to be too inadequate for the purposes of this program. The staff considered them to be too broad and superficial in content presentation to address the critical thinking required of students in SDEP. Often, textbooks tend to present predigested knowledge that does not permit multiple interpretation. While some predigested knowledge is important as background information, primary source materials are needed to promote critical thinking skills in pupils.

Two dilemmas faced the project's staff and teachers. The first was to locate sufficient and appropriate instructional materials that would provide enough information to stimulate the depth of classroom discussion desired. The second dilemma stemmed from the teachers' belief that there was not enough time to cover the existing textbook. Teachers believed that if a substantial amount of time were taken for discussion with students, they would not be able to "cover the ground" as they had in the past.

Lessons were designed to use the instructional materials and deliver the expected students' learning outcomes. Again, this was not as easy as one would think. Decisions had to be made about how much time to devote to discussion. Additional issues included the way to incorporate discussion texts into lesson design and how training will be provided for teachers and students to encourage them to participate in discussions.

Finally, a syllabus was prepared to provide students and parents with a statement of expected outcomes, sample test items, and examination questions. Sample responses to essay questions were provided to give students an example of criterion behavior expected.

The feasibility study for the SDEP took place during the spring of 1985. It involved the development of a syllabus, instructional sequences, and an examination for a four-week unit on Russia and the Cold War, part of the World Cultures course taught to all tenth-grade students in Pittsburgh. The unit was taught to students at three high schools. The results of the feasibility study indicated that a syllabus-driven examination system can be implemented for students of general ability in American secondary schools. The participating teachers reported that the students were able to attain the established learning outcomes. However, the feasibility study also indicated that teachers found the new mode of instruction difficult, and some of them were initially resistant to the idea of a significant reorientation of curriculum, instruction, and testing. A pilot test will continue in the fall of 1985 to complete the course in World Cultures and begin U.S. History and English.

Monitoring Achievement in Pittsburgh-MAP Writing and Critical Thinking

Is it realistic to assume that all students can perform well on examinations that typically have been restricted to high-achieving, college-bound students? In Pittsburgh, the answer is yes! This assumption can be made because of the district's recent emphasis on the enabling skills of reading, writing, and critical thinking, which are part of its MAP program.

The MAP writing program emphasizes both grammar and composition. Standards are established for student writing at each grade level. Elementary students, for example, are expected to write good topic sentences in their compositions, use logical structure, and bring their writing to an appropriate conclusion. The writing demands become more complex as a pupil progresses through the grades. Senior high school students are expected to use varied sentence structure and rich vocabulary as well as to meet all of the standards set for previous grades. Both direct instructional programs for students and training programs for teachers have been developed and implemented to promote the development of students' writing skills. Four writing samples are required of each student each year, and they are analyzed according to established criteria. As with all MAP programs, individual student's results are sent home to parents. Teachers receive a classroom profile that details the relative strengths and weaknesses of pupils based on the teacher's analysis of students' writing.

The MAP critical thinking program is similar with respect to the analysis of student essays. The critical thinking program stresses the ability of pupils to state a position (both orally and in writing) with respect to a particular topic of study in the social studies program. In preparing oral and written responses, students are expected to support their positions with evidence from appropriate texts or other relevant sources; students are expected to elaborate upon the evidence presented and draw their argument to a conclusion. The MAP critical thinking tasks are similar to the Document-Based essay Questions presented to college-bound students in certain AP exams. Thus, the MAP writing and critical thinking programs provide students with the skills that will enable them to address the tasks required by examinations.

The development of the ability to think, speak, and write critically presupposes that a student can engage in a productive discussion with respect to an issue or in response to a text. Therefore, an emphasis on teacher questioning and discussion has assumed a central role in the implementation of MAP critical thinking and SDEP programs.

Pittsburgh's Questioning Categories

The pilot testing of the MAP critical thinking program in 1982 revealed that teachers were skilled in conducting discussions. Recitations are used by teachers as a quick check of student understanding of a topic being presented. Discussion, on the other hand, requires complex thought processes and the development of an attitude of tentativeness toward knowledge (Dillon, 1984). In discussions, an emphasis is placed on multiple interpretations of texts. Students are encouraged to accommodate more than one point of view on a topic or issue.

Discussions can be difficult to conduct. The form of questioning that fosters a discussion is a highly complex skill, contrary to what may first appear to be the case. Consequently, the developers of the MAP programs reviewed the research related to teacher questioning and then developed a format for questions and a model for discussion in the classroom. The product of their work is called the Pittsburgh Questioning Categorier (PQC).

The PQC identify three types of questions: literal, inferential, and evaluative. Literal questions usually require recall of information. Inferential questions require interpretation, generalization, or prediction based on textual material. Evaluative questions require judgments regarding the merit, importance, and value of ideas. The model for conducting discussions requires teachers to frame a "main question" to lead off a discussion. Main questions must pose an air of uncertainty and must have at least two plausible answers. Based on student responses to the main question, teachers ask probing questions that may ask a student to clarify, elaborate, or justify a response. Teachers may also ask management questions designed to elicit broad participation from the class or redirect the discussion to keep it on task.

For students to perform well on the SDEP, they must take active part in discussions that will encourage them to analyze and articulate their ideas. Effective discussions among students themselves and with teachers will bring about greater understanding of the text and their responses to it. Discussions, combined with the monitoring of student achievement in composition and critical-thinking writing tasks, should prepare Pittsburgh's students to respond effectively to a formal examination system. The SDEP is a logical extension of the MAP programs in the Pittsburgh district that have been implemented in all schools since 1982.

A New Role for the Testing Industry

If the widespread use of examinations in American schools is to continue into the 21st century, we must recognize an important new role for the American testing industry. My fundamental assumption is that examinations such as SDEP would not be national or state exams, but local or perhaps regional exams. The exams must be responsive to the specifications of the local or regional educational agency.

To develop such examinations and accompanying instructional systems requires that local or regional education agencies: a) achieve consensus on important learning outcomes; b) identify instructional materials and methodology to be used; c) prepare a student syllabus for each course; d) design and conduct in-service training for teachers; e) develop examinations to address those outcomes; f) implement and evaluate the program.

It is the view of the author that the testing industry is in a position to assume a new and more interactive role, and, in collaboration with local school districts, to design and implement examination systems. The author believes that it is appropriate for the testing industry to:

1. Identify, train, and establish networks of college and university personnel to work with local educators to identify the most important learning outcomes in major academic disciplines.

2. Develop examinations in collaboration with local agency educators and university personnel or provide technical assistance to LEAs to develop, pilot test, and refine their own examination system.

3. Develop, validate, and maintain a secure bank of multiple-choice test items to be used as part of the examination system.

4. Develop training and/or technical assistance programs to prepare teachers of LEAs to grade written examinations.

5. Provide technical assistance to LEAs to validate a sample of graded examinations to provide quality control.

6. Conduct research on the positive and negative impact of such exams on students, teachers, and school districts.

7. Conduct research on the ways in which classroom techniques or instructional sequences are influenced by the imposition of examinations.

The proposed role for the testing industry would bring it into a much more active collaboration with LEAs than has been the case. The shift from developing, validating, and scoring *standardized tests* to an active role in assisting LEAs to develop, administer, and score *examinations* would be a significant change. It requires a shift from a product orientation to a service orientation. It is my view, however, that the testing industry is in a unique position to contribute in a constructive way to the improvement of the quality of American education. Raising the level of performance of students in this country requires that we raise the level of expectation for all students. This is particularly true of the urban youth of the nation. The author asserts than an examination system can serve to raise academic standards and improve student achievement in the schools.

The results of much educational research indicate that members of minority groups and the poor will respond to higher expectations for achievement if the learning environment is consistently supportive of their efforts and if their progress is monitored carefully. If new standards of excellence are presented by means of examination systems and if students' acquisition of basic skills and critical thinking is carefully monitored, students will respond to examinations successfully.

If this goal of excellence can be realized through the use of examination systems, the quality of education in America in the 21st century will be significantly better than that recorded in the 20th century. The testing industry, in collaboration with local school districts, can and must play an important role in the transformation of the schools.

References

Dillon, J.T. (1984). "Research on Questioning and Discussion," *Educational Leadership*, 42(3), pp. 50-56.

Madaus, G.F. & V. Greaney (1985). "The Irish Experience in Competency Testing: Implications for American education," *American Journal of Education*, 93(2), pp. 268-294.

Madaus, G.F. & J.T. McDonagh (1979). *Minimal competency testing: Unexplored negative outcomes*. Paper presented at the 9th annual conference on large scale assessment sponsored by the National Assessment of Educational Progress, June 11-14.

Resnick, D. P. & L.B. Resnick (1985). "Standards, Curriculum and Performance: A historical and comparative perspective," *Educational Researcher*, 14(4), pp. 5-21.

Resnick, D.P. (1981). "Testing in America: a supportive environment," *Phi Delta Kappan*, 62(9). pp. 625-628.

Wallace, R.C., J.R. Young, J. Johnston, W.E. Bickel, & P.G. LeMahieu (1983). "Secondary Educational Renewal in Pittsburgh," *Educational Leadership*, 41(6), pp. 73-77.

Barriers to New Test Designs

ROBERT L. LINN
University of Illinois at Urbana-Champaign

Implicit in the theme of this conference, "The Redesign of Testing for the 21st Century," is the idea that scientific and technological advances can provide a foundation for substantial improvements in testing. There are several reasons to think that this idea is timely. First, advances in cognitive science are providing new understandings of cognitive processes that have major implications for the design of instruction and testing. Second, advances in technology, particularly low-cost microcomputer technology, promise many new possibilities for testing. Third, there seems to be a reawakening of interest in instructional uses of tests in the measurement community. Finally, these changes are occurring in a larger context of heightened concern about education in which testing is often viewed as a powerful tool for achieving reform.

Together, these four forces have the potential to reshape testing. They could lead to improved measures of developed abilities for the traditional purposes of selection, classification, certification, and guidance. More importantly, they could lead to measures with greater instructional utility. But the reshaping will not be easily accomplished. A number of barriers will need to be overcome if the envisioned improvements in testing are to be even partially realized. As the title of this paper indicates, some of these barriers are technical in nature. However, I believe that the more serious barriers are economic and ideological; and I will, in fact, give at least as much emphasis to these as to technical barriers. In any event, it is important to understand the obstacles to change, whatever their nature, in order to overcome them.

Efficiency

To begin this process, it is useful to consider the current system of large-scale testing. Standardized testing is such a familiar part of American education that it hardly needs to be described in any detail. Almost all school districts administer a variety of standardized tests each year. They provide information to parents and school boards and serve general demands for accountability. Grade-to-grade promotion and high school graduation in many schools depends, in part, on passing a test. Standardized tests are also used to evaluate compensatory education programs, to identify students for remedial and special education programs, and to identify students for gifted education programs. As students move up the educational ladder, the use of standardized tests continues. Most colleges and universities, even those that do not have selective admissions policies, require applicants to submit test scores. Upon completion of school, students must undergo additional testing in order to be certified or licensed to practice in a growing number of occupations.

Standardized tests serve a relatively wide variety of educational functions, ranging from the symbolic to ones that significantly affect individuals and educational institutions. The range of functions served by standardized tests and the amount of testing has grown tremendously during the last half century. Despite this tremendous growth in testing and a variety of technical refinements that have been made, the fundamental nature of standardized tests has remained remarkably unchanged during this period. This is not to say that the refinements are unimportant. High-speed optical scanners and computers have made testing an extremely efficient enterprise. Psychometric advances, especially the development of Item Response Theory, have led to improvements in item analysis, test design, test equating, and in procedures for detecting item bias. Item Response Theory also provides the basis for the design and implementation of computerized adaptive testing. Efficiency has been enhanced by new item types, such as the quantitative comparison items used on the Scholastic Aptitude Test. These and other advances in the field of testing are significant, but they have not led to major alterations in the fundamental nature of what is tested or in the valid use of scores for improvement of learning.

The relative lack of change is not a consequence of an absence of efforts to develop alternative procedures. Numerous attempts have been made at expanding the range and nature of tests. Ingenious tasks and item formats have been devised. Various combinations of hundreds of tests have been administered to thousands of people in a continuing search for better measures and a better understanding of the facets of human ability. However, testing is a highly pragmatic undertaking, one that is largely ruled by two masters—the predictive-validity coefficient and the economic viability of the product. From the perspective of both of these masters, the combination of multiple-choice tests and machine-readable answer sheets is the clear winner. This combination is not only highly efficient and cost-effective. it yields predictive validities that have proven hard to exceed in a host of studies that have attempted to demonstrate better predictive power for a wide range of experimental measures.

The extraordinary efficiency and relatively good predictive validities provided by the existing technology represent major barriers for the redesign of testing. Despite the remarkable reductions in the cost of computer technology, there remains a substantial gap in the likely total cost of running a fully operational computerized testing system for, say, a million candidates a year, and the cost of testing those same million people with a current testing system such as the College Board's or American College Testing Program's. It may be that future reductions in costs and increases in availability of microcomputers will eventually lead to a crossover in the relative costs of paper-and-pencil and computerized test administration, but even if it does become a more economical means of administering tests, computer administration by itself will not necessarily lead to fundamental changes in what is measured or in the validity of the measures.

The use of computers for test administration that is currently receiving the most attention, or at least the greatest financial resources, is computerized adaptive testing (e.g., Green, 1983), i.e., testing where item selection is based on the test taker's previous responses. Adaptive tests have both intuitive and psychometric appeal. The less able test taker is not needlessly frustrated by the presentation of a large number of items that are clearly too difficult and the more able test taker does not have to waste time or risk boredom answering numerous items that are too easy. The primary psychometric advantage is one of increased efficiency. According to Ward (1984, p. 17), for example, with adaptive testing "the length of a test battery can be cut by 50 to 60 percent and still maintain a measurement accuracy equivalent to that of the best standardized conventional test."

Increased efficiency is certainly not something to belittle. If adaptive testing can, in fact, cut the average testing time in half without losing precision of measurement, that would save several hours of testing time for millions of test takers each year—time that could be used for instructional purposes or for expanding the range of characteristics that are measured. Computerized testing could also enhance flexibility. For example, it could eliminate the need for administering secure tests such as the SAT (Scholastic Aptitude Test), ACT (published by the American College Testing Program), or the GRE (Graduate Record Examinations) on only a few selected days a year. Test takers could instead schedule a time to take the test at a computer terminal at their convenience. All that could be accomplished, however, without changing the fundamental nature of what is being measured and without any noticeable increase in the predictive power of the test information.

As long as standardized testing is driven primarily by the traditional goals that are well served by a global ranking of students on one or two dimensions such as verbal and quantitative ability, it seems unlikely that we will see a revolutionary redesign of tests. Better measures of cognitive processes or measures that provide better information for guiding and enhancing learning cannot be expected to compete in terms of the standards of efficiency and predictive validity. We need to focus on different goals and use different standards for evaluating the effectiveness of the measures if we are to have a significant redesign of testing. At its most general level, the goal I have in mind is the effective use of tests to enhance learning and cognitive development.

Instructional Testing

So stated, this goal does not sound unusual. Publishers of standardized achievement tests give lip service to the instructional use of test results. They provide an impressive array of scoring services that promise to provide teachers and students with diagnostic information for guiding student learning. However, there is little evidence that teachers find the results particularly useful for this purpose. Indeed, many would agree with Bejar's (1984, p. 175) conclusion that "standardized tests frequently have little or no impact on instruction because the test results offer little help in designing instruction that is optimal for the individual student." Assuming, as I do, that Bejar is correct in this assessment, it seems important to understand the reasons that existing standardized tests do not have more instructional value, to consider the types of new test designs that would improve this situation, and to identify the barriers that will need to be overcome if we are to redesign tests in ways that enhance their instructional utility.

Current achievement tests do a good job of assessing a student's general level of knowledge in a particular content domain. They provide a reasonable basis for comparing the current performance of students and are relatively good predictors of future performance. A low score relative to a student's grade placement on, say, a reading comprehension test is apt to be a valid indicator that a student will have difficulty reading and understanding assignments in the typical textbooks used at the grade level. Such global information, however, is more likely to confirm what the teachers already know about the student than to provide them with new insights or clear indications of how best to help the student. The global score simply does not reveal anything about the causes of the problem or provide any direct indications of what instructional strategies would be most effective.

One recent response to the limitations of global achievement test scores has been proliferation of tests designed to give highly specific information. These tests, which are referred to by a variety of labels such as criterion-referenced, objectives-referenced, curriculum-embedded, or mastery tests, splinter the content domain into tiny skills and specific bits of knowledge. For example, short tests for specific objectives such as "recognize the phoneme-grapheme correspondences for diphthongs" or "divide for syllabication a two-syllable word with medial consonant letters" (Smith and Arnold, 1983) can be found in assessment systems accompanying basal readers. Although information about the accumulation of discrete facts is potentially relevant, it is insufficient, for as Snow (1980, p. 43) noted at the 1979 ETS Invitational Conference, "achievement is no longer to be understood as simply the accretion of facts and content specific skills."

A clear definition of the subject-matter content is essential, but insufficient by itself. An understanding of the learner's cognitive processes — the ways in which knowledge is represented, reorganized, and used to process new information — is also needed. The importance of the latter is strongly suggested by recent research in cognitive psychology and artificial intelligence. A number of authors (e.g., Bejar, 1984; Curtis and Glaser, 1983; Glaser, 1981; Messick, 1984; Pellegrino, 1985; Snow and Peterson, in press; Snow, 1980; and Sternberg, 1984) have summarized this work and its implications for testing. Other papers at this conference also address this topic. I will not attempt to provide another review of that work, but merely to point to three strands of that work to show its potential relevance for improving the instructional utility of testing and to consider the barriers that stand in the way of realizing that potential. For convenience, I'll refer to these three lines of work as *cognitive components*, *error analysis*, and *cognitive structures*.

Cognitive Components

The cognitive components approach is typified by the work of Sternberg (1977, 1980) and Pellegrino (1985) who have attempted to identify the basic mental steps or cognitive components involved in inductive reasoning. This work and its implications are well illustrated in a recent paper by Pellegrino (1985). For example, Pellegrino describes four processes involved in the solution of verbal-analogy problems of the type commonly found on a variety of I.Q. and verbal-aptitude tests. These are: (1) encoding "in memory the important attributes of each term in the analogy"; (2) comparing the specific attributes of each term in the analogy and inferring the relationship between the first two terms; (3) application of the inferred relationship to the third term of the analogy; and (4) "evaluating the potential answers and responding" (Pellegrino, 1985, p. 51).

Component scores that provide information about the speed and accuracy of performing each of these processes have been devised in the laboratory. With computer-based test administration, it would be feasible to obtain separate component scores of the type described by Pellegrino on an operational basis. The natural question, however, is what advantages would these component scores have over the global scores that can be obtained so efficiently now with a conventional verbal analogies test? It is unlikely that the added expense and complexity of the component scores could be justified in terms of improving conventional predictions. Rather, the justification of such scores for practical application and use would need to be based on quite a different standard, namely, their utility for facilitating development of the inferential reasoning ability for the people taking the tests.

Pellegrino (1985, p. 54) suggests that component scores "could pinpoint a person's weak areas of cognitive functioning and provide some basis for designing individualized instruction and training to improve cognitive skills." There is laboratory research to suggest that this lofty goal may, at least to some extent, be achievable. This is a worthwhile goal, but one that will require a substantial amount of research and development effort.

If test publishers are to play a significant role in such an effort, they will need to expand their markets, put much greater effort into linkage between testing and associated instructional materials, and add to their traditional approaches to test validation. Evidence that the separate measurement of components, when linked to individually targeted instruction and training, can improve cognitive skills will need to supplant the traditional reliance on correlation coefficients. Of course, such information is quite consistent with the notion of construct validation, but as Cronbach (in press) has recently pointed out, serious efforts at construct validation are the exception rather than the rule. What too often passes as construct validation in test manuals is an undigested array of correlation coefficients. Such evidence is simply inadequate for validating a set of scores to identify cognitive components that aid in the development of human intellectual abilities.

Error Analysis

A second line of work that seems to have immediate implications for instructional testing is the analysis of errors. As Brown and Burton (1978) and Tatsuoka and Tatsuoka (1983) have shown, student errors are often systematic, and detailed analysis of errors can lead to the identification of the nature of student misconceptions. Although there is only limited evidence to show that this identification leads to more effective instruction, it is, at least, highly plausible that it should.

The analysis of student errors requires a different type of testing analysis than is typically used to support the development and use of global test scores. It can be highly labor-intensive activity. Logical analysis of the content as well as painstaking analysis of student responses is required. There are, however, qualitative differences between the results of such tests and the traditional global score on a standardized achievement test. The latter may tell a teacher that a student performed better than only ten percent of the normative sample on an arithmetic test, but provides no real indication of the nature of the student's difficulty. A test designed to diagnose errors, on the other hand, may indicate that when asked to add fractions with different denominators, a student consistently gets the wrong answer by separately adding the numerators and denominators. The latter information suggests specific corrective action whereas the number-right score does not.

Of course, the information provided by error analysis is also more complex. The single numerical score of a traditional test, with all its supporting psychometrics, is replaced by an array of information about the categories of errors made by a student. This may call for a new type of psychometric analysis with different scoring procedures and ways of characterizing, validating, and reporting what Tatsuoka (1983) refers to as a "rule space." The work on the psychometrics of error-analysis procedures is still in its infancy. It will take considerable time and effort to bring it to a level of maturity needed to support large-scale operational testing programs that help teachers identify student errors in instructionally useful ways.

Although the primary illustrations of error analysis come from the area of arithmetic, considerable progress has also been made in the areas of people working in educational technology are more comfortable with the passive, behavioristic view of the learner. The warning of Adams and Jones (1983, pp.27-28) bears repeating:

> The Learning Theory model for computer use is simplistic and flawed. It takes for granted that education is the acquisition of facts or concepts treated as facts. It is Gradgrind Redivivus and, like the offerings of that utilitarian it can still be found lurking in some of the current demands for standards or a return to "basics".... It is a barrier to questioning and ignores personal experience as the foundation for growth. It lacks any sense of how the whole person is involved in real learning.

In the same vein, and with remarkable prescience, Philip Jackson, speaking in 1967 of the teacher and the machine, observed that: "... many of the technological tools . . . designed for use in the schools are being promoted by men who talk and think like engineers (p.15)." He is dated only in limiting his observation to men. Today, for too many people, the implied model of the brain is the digital computer. In the 1984 Reith lectures (p.44) on the British Broadcasting Corporation (BBC), John Searle corroborated the wide acceptance of this metaphor:

> Because we do not understand the brain very well we are constantly tempted to use the latest technology as a model for trying to understand it. In my childhood we were always assured that the brain was a telephone switchboard. ("What else could it be?") I was amused to see that Sherrington, the great British neuroscientist, thought that the brain worked like a telegraph system. Freud often compared the brain to hydraulic and electromagnetic systems. Leibniz compared it to a mill, and I am told that some of the ancient Greeks thought the brain functions like a catapult. At pre:ent, obviously, the metaphor is the digital computer. (p.44)

The metaphor is powerful—revealing something dangerous about our culture's values and anxieties. "It denies free will—just like astrology and bio-rhythms—and thus it is comforting because it removes responsibility" (Hall, 1985, p.8). The mind-as-computer metaphor is captured in Laurence Lerner's archetypal digital computer **A.R.T.H.U.R.**, whose "credo," in part, runs as follows:

> I believe in the binary structure of reality: One substance but two possibilities: One process but two alternatives (No path can be left if it is taken: No switch can be on if it is off).... I believe in logic: If I can do what a man does, I am a man. Socrates was a man. I am as good as Socrates. (Lerner, 1971, p.32)

One ancillary issue that must be raised under "technology and the view of the learner" is equity, not the obvious economic equity issues raised by differences in the ability of districts to afford the new technology, nor those raised by TV ads aimed at yuppie parents, depicting the home computer as providing a competitive edge for their children. Rather, I am concerned that the less able child might be trapped in electronic workbooks full of repetitive, practice-and-quiz routines, while the brighter student enjoys exploratory software.

Educational Technology and One's View of Teaching

It has been too readily assumed in some quarters that teaching can be operationally defined in much the same way that Cy Bernation defined golf, and for much the same reason: Teachers play the game badly. This mentality first surfaced early in this century in what Callahan (1962), called *Education and the Cult of Efficiency*. It resurfaced in the '60s when the term "teaching machine" came into vogue. Pedagogy was to be transformed into a science, and the teaching function was to be programmed as far as possible (Jackson, 1968). Some people will remember talk of developing a "teacher-proof" curriculum.

It never happened. But, today, with the microcomputer—the great promise of "artificial intelligence" —and "expert systems" and a new generation of adaptive machines, new and bolder promises are being made. Some claim that these developments will eventually permit us to do what we could not do in the '60's—to build a machine that will largely replace the teacher. In the view of some, such as Evans (1979), we may even replace schools themselves. Enthusiasts of artificial intelligence point, with understandable pride, to those chess programs that can challenge a master. The master teacher, however, must employ a much broader constellation of abilities than the chess master, and do so in the extremely complex social system of the classroom.

In speaking of teaching machines and artificial intelligence, the choice of adjective and noun, in both instances, is unfortunate and downright misleading. Machines can't teach; at best, they can instruct or tutor, and with careful selection of software and human monitoring, they perform this more limited function well. The danger of attaching the adjective "artificial" to intelligence can be seen by comparing this usage to the ways in which we use "artificial" to modify heart, kidney, or limb. The comparison conveys the spectra of damaged or defective intelligence, restored by artifice. Could the term "artificial intelligence" have emerged from an "artificial imagination"? Further, the concept of intelligence in artificial intelligence is as limiting as the same concept embodied in intelligence tests; and look at the confusion we have had to live through on that score.

Searle (1984, p.31) describes well the fundamental difference that will always exist between human and artificial intelligence:

> The reason that no computer program can ever be a mind is simply that a computer program is only syntactical, and minds are more than syntactical. Minds are semantical, in the sense that they have more than a formal structure—they have a content.

Consequently, I don't see teachers under threat, whatever grandiose claims are made for artificial intelligence. Engineers using artificial intelligence will never design hardware or software that will be the equivalent of human teachers.

Teaching is a creative, improvisational performance involving a dialectic between the fixed conventions of organization and curriculum on the one hand, and a teacher's personal interpretation and style on the other. Teaching has its framework of stereotyped obligatory procedures and behaviors, but otherwise it allows great freedom of treatment. The former can, of course, be operationalized, the latter cannot be. Teaching can be compared to the blues, which has:

> no single "authentic" form but somewhat altering from singer to singer and even from verse to verse,...with its balance of constraint with freedom, fixed model with fluid treatment, communal taste with individual fantasy, traditional constancy with novel creative moments, sameness with difference. (Lloyd, 1967, p.63)

No "singing machine" can replace Bessie Smith; no "teaching machine" can replace Mr. Chips.

At the heart of teaching is the inescapable element of relationship. Isaac Asimov (1963) has an interesting science-fiction story in which a child of the future, whose entire instruction and evaluation is through an anonymous machine, discovers that, in some bygone age, students actually went to a place called a school and spoke and met with a human teacher.[2] This elicits the nostalgic sigh that provides the story's title, "The Fun They Had." No machine, however benign or "smart," can provide the essential educational element of human relationship.

Precisely because it involves relationships, teaching is more often than not spontaneous, unplanned, unpredictable; a creative performance that relies heavily on the subtle interaction between and among students, the class as a whole, and the teacher. No machine can recognize—let alone decipher—the fleeting cues in posture—a yawn, giggle, whisper, or furtive or bored look—and know how to alter its tactics in response. Moreover, you cannot totally define all educational objectives with the overt precision that a computer program presupposes. But most of all, teaching entails a moral and ethical relationship between human beings (Jackson, 1968). No program, no matter how sophisticated, will ever be able to care about, or feel responsible for children.

The Impact of Changes in Testing on Education

Two technologies are under consideration today: the electronic technology and the technology of testing. For me, at any rate, the former is arcane, the latter relatively simple. Basically, testing technology comes down to a simple alternative—the examinee can provide a product or answer (the supply mode) or choose an answer (the selection mode). The selection mode, in turn, is buttressed by a complex psychometric system.

As the manner of testing has evolved over time, so too have the ways in which acceptable answers have been judged. Hoskin (1979) has an interesting analysis of the historical evolution of the supply mode of examining. Initially, the medieval guilds required that an apprentice supply a product as a final proof of competence.[3] This, in turn, was evaluated according to well-defined criteria by the master. In the medieval universities, the apprentice scholar demonstrated mastery of the subject in a "performance of display" before masters once a year. This oral disputation, which consisted of responding to previously known questions, was qualitatively evaluated by the master according to the examinee's ability to demonstrate a traditionally approved form of rhetoric. On the basis of their total performance, candidates were sorted into classes. Around the year 1750, people were ranked within classes rather than left undifferentiated.

With the advent of a modest technological breakthrough, the written examination, there was a further development. While the questions were still known in advance and the oral mode persisted, both the oral and written products of several days of examining had to be qualitatively assessed according to traditionally acceptable answers. At the heart of the system was the presumption that examiners singly and in concert could rank a *total performance*. Inevitably, partiality crept in and, in 1972, William Farish introduced the then momentous innovation of assigning quantitative marks to individual questions. "Quality" was now mathematized; individual marks could be summed, the individual's performance and consequent overall ranking compared to others. "The blunt weapon of banding yielded to the precision tool of the mark" (Hoskin, 1979, p.144).

The combination of writing and ranking led inevitably to a homogenized examination system: a common set of questions arising from a common curriculum. The quantitative symbol of the mark had assumed supreme significance, and a generation later, the question posed assumed a factual "right" answer that left little room for individuality or rhetorical flourish. "Narrow specialization, and examination based on the principle of testable knowledge became the new parameters of undergraduate education...and with it a new intellectualist ideal, what we now call proficiency or the acquisition of skills" (Hoskin, 1979, p.45). These changes, in turn, drastically modified what was taught, how it was taught, what was learned, and how it was learned.

In order to hold headmasters in the Boston Public Schools accountable, Horace Mann imported the written essay supply exam to America. Here it evolved into the short-answer supply form. Presently, teachers, for the most part, limit their tests to recall of information and favor the use of the short-answer form (Goodlad, 1983).

The development of the selection-type item early in this century and the recognition and exploitation of its efficiency and commercial possibilities after the first World War introduced national norms, the answer sheet, and eventually, the scoring machine. Since these developments, formal district or statewide testing programs have been limited, by and large, to the use of the selection mode; thereby limiting, according to many critics, the usefulness of such tests in the instructional process. The selection mode did not remov human judgment from testing; human judgment goes into deciding which domain to measure, evaluating items for content validity, and in the setting of cut scores. But what was lost by use of the selection mode almost exclusively in formal testing programs was reflection, discernment, and evaluation, all of which are crucial in the supply mode. Many people, at least from the time of Starch and Elliott (1913), have been worried by this subjective element inherent in the supply mode. Their resolution was the multiple-choice format, which has come to dominate formal testing. With hindsight, I believe that the avoidance of the supply mode in formal testing programs was an overreaction to the problems associated with it—the product of a limited epistemology, attractively packaged in administrative convenience. We simply swapped one set of problems for another with our overreliance on the multiple-choice format.

Recently, policymakers have discovered the accountability potential, and the power to influence teacher and student behavior inherent in attaching rewards and sanctions to multiple-choice test performance. The selection mode is assuming an overriding significance. The quantitative score has become synonymous with qualitative evaluation, not only of individual students and teachers but of the system itself. Thus, we have moved inexorably from a qualitative, reflective evaluation of answers supplied by students to a quantitative, mechanical assessment based on the optical scanning of marks on an answer sheet.

The Promises of New Tests and New Technologies

As I said in my introduction, the peril of coupling new tests with new technologies is that it can accelerate and further legitimize a mechanistic, solely quantitative evaluation of the person. On the other hand, the promise of the new technology is that it can facilitate the reintroduction into testing of free-form answers, human judgments, and evaluations of them.[4]

It seems a pity that, to date, the new technology has been used primarily to make the selection mode more efficient. Computer-adaptive testing utilizes Item Response Theory and the computer's computational power to permit an examinee to take a much shorter version of a selection test. This work is interesting, and undoubtedly will be used successfully in many of the smaller certification programs, and eventually for traditional school district testing programs which use tests like the *Iowa, Metropolitan, California*, etc. All of this awaits more information on the differential validity of this approach; it is not a simple one-to-one transformation from one mode to another. We already know that children react to the technology in very different ways (Turkle, 1984). And we need to keep in mind the trait/method literature.

The lack of adequate numbers of terminals rules out this more efficient use of the selection format for large-scale system or statewide certification testing programs, at least for the foreseeable future. Further, this approach simply canonizes the status quo—the predominance of the selection mode.

I see the promise of new tests and new technologies as being primarily within the world of the classroom as instructional aids, rather than as grading or certifying devices. Formative evaluation—continuing feedback to improve student performance rather than to simply grade it—can be greatly facilitated. And such is the versatility of the new technology that it permits a return to the supply mode. Oral, written, and visual input and output are not only possible, they are fast becoming feasible as well.

The microcomputer and commercially available software promise much more emphasis on writing and language skills. Recently, reformers have put heavy emphasis on mastering the English language, with writing seen as the key (e.g., Boyer, 1983). Word processors, spelling checkers, grammar checkers, electronic thesauruses, and programs that count types of words and varieties of images and monitor syntactical usage all help teachers in their work to improve student vocabularies and writing skills (Foley, 1985). But no machine can read and critically evaluate whether a student's writing "works." This requires qualitative human evaluation and feedback, however subjective and fallible.

Another promise centers around higher-order analytical skills. One of the reasons such skills are so underdeveloped in our school children is that they are inadequately taught and inadequately tested. Now, the current repertoire of digitized photographs, music and speech, video documents, computer graphics, and document and voice recognition opens up the possibility of presenting a wide range of stimuli to students, also permitting a wide range of free-form supply answers. These features can be tapped to develop higher-order skills and test for them in ways not currently possible.

In both the physical and social sciences, simulation programs can allow teachers to test for abilities that previously were cumbersome or administratively awkward to assess. But here again, I would hope that qualitative assessments oi student and teacher responses, not machine feedback, would be a built-in feature of such programs. These types of testing programs must be very subtle and must be mediated by the astute judgment of teachers.

Another promising possibility of new tests and new technologies lies in their potential to provide truly diagnostic information for teachers. Most state testing programs promise diagnostic information, but none, in my opinion, delivers. Teachers are weary of commercial or state tests telling them what they already know—that Dick can't read, or Jane can't compute. What they would really appreciate is more detailed information as to why this is so, and what strategies they could adopt to deal with it. Brown and Burton's (1978) diagnostic program, *BUGGY*, which analyzes the answers given by a child to arithmetic questions to determine whether there is a specific "bug" or defect in the child's procedure, is a step toward fulfilling this kind of need. However, *BUGGY* is not yet available for micros; and notice that perennial easy arithmetic, not language or reading, is the subject of analysis.

A final promise—already *partially* fulfilled—is the provision to teachers of computerized item pools containing both supply and selection exercises. Such item pools can incorporate a wide range of stimuli not heretofore available. Teachers can use these computerized pools of items to build quizzes and tests, or to give students practice with certain kinds of skills.

But promises will remain mere promises if policymakers continue to prefer traditional multiple-choice tests in accountability schemes to make important decisions about students, teachers, or school systems. If the present trend continues of using student performance on multiple-choice tests as a necessary condition in decisions about graduation, promotion, merit pay, or district certification, then it will surely dictate what is taught, how it is taught, what technology is developed and marketed, and how it is employed. Already, ads are appearing in national magazines for hardware/software systems that purport to help raise a student's performance on a state mastery test.[5]

In the pursuit of the very real promises mentioned above we should not be blind to the alterations that they may bring about in the life of the classroom. The more time the student spends interacting with a machine, the less time teachers have to make powerful informal evaluations of individuals and the class as a whole.

One final observation: We are inclined to think that children will be fascinated and motivated by new tests linked to new technologies. It is presumed that the new technology will remove the drudgery, boredom, and anxiety associated with traditional forms of testing. Even a drill test with the imaginative appeal of Space Invaders, I suspect, has a severely limited life expectancy. The fate of all toys, no matter what their educational possibilities, is to remain toys and eventually to be discarded in favor of reality. Jackson (1968, p.49) sums up this limitation best when he points out:

> The same tendency that leads to the ultimate rejection of make-believe will likely have some effect on the students' willingness to "converse" and "reason" with a computer console. Engineers can add sound, color, canned applause, and even low-heeled oxfords but their product will forever remain a toy teacher not a real one.

Conclusion

In conclusion, what of Dick and Jane and these great analytical engines? We need to temper our enthusiasm for new tests and new technologies with a measure of Ludditism. We need to listen carefully to the opposition. It may be expressing an instinctive feeling that some very real values are at risk.

Time has proved the Luddite at least partially right. It is only now that we fully realize the social, ecological, and human costs of the industrial revolution. The nineteenth-century opponents of the "dark satanic mills" may not have expressed themselves with the media hype of twentieth-century activists, but now, with hindsight, we see that many of their intuitive fears were justified.

We always want our children to have the advantage of technological developments we ourselves have begun to enjoy later in life. Our concern that they have access to technologies unavailable to us in our school days leads us to assume that any skill can be grafted onto children without displacing another or absorbing the limited energies that should be engaged in the more important areas of human development. Postman (1981), who calls himself a media ecologist, has some sobering reflections on the impact of television, which has brought about what he calls "the day our children disappear." No one fully anticipated the tremendous impact TV has had on our children. I would hope that we do not make the same mistake with new tests and new technologies.

Provided the human element remains dominant, and the child is not further robbed of his or her childhood, then we can endorse the remarks of the president of Harvard when he says:

> In the end, therefore, with all the exaggerated claims and the media hype, we can still look upon the new technology with cautious enthusiasm. At the very least, [schools] should manage to use technology to engage students in a more active process of thir.king and problem solving that will help them leam more effectively. At best, the new machines may also be a catalyst to hasten the development of new insights into human cognition and new ways of helping students leam. (Bok, 1985, p.8)

Footnotes

1. When Disraeli was asked to evaluate Babbage's project, he characterized it as "indefinitely expensive, the ultimate success problematical and the expenditure incapable of being calculated." To which Babbage replied that this was "excusable in the Chancellor of the Exchequer who was himself too practically acquainted with the fallibility of his own figures, over which the severe duties of this office had stultified his brilliant imagination." (Moseley, 1964. p. 226)

2. In Asimov's story, the "regular" teacher is the machine. The little girl hated most "the slot where she had to insert homework and test papers. She always had to write them out in a punch code they made her leam when she was six years old, and the mechanical teacher calculated the mark in no time." (p. 26)

3. Waterford glass has a bowl called an apprentice bowl that incorporates the repertoire of standard "cuts" used by a qualified artisan. This is strongly reminiscent of the medieval craft tradition.

4. A factor that will continue to retard the realization of the full power and promise of new tests and new technologies is the current incompatibility of different hardware and software, even within the same brand name (Macrae, 1984). Right now, the situation is analogous to having a Bruce Springsteen tape that can be played only on one model of Sony tape decks. Undoubtedly, this problem will eventually be solved, but until then, schools are at the mercy of the particular brand of computer they adopt; and some software developers may be hesitant to enter the educational market. Incompatibility in technology is nothing new. The development of sign language for the deaf in Europe proceeded on parallel sectarian lines so that deaf Protestants and Catholics had difficulty communicating with one another for many years.

5. See, for example, the WICAT Systems ad in the September 24, 1984, issue of *Newsweek* (p. 11). The ad points to a comparison between the performance of a class that used WICAT and another one that did not. The ad states that 90 percent of the WICAT class passed the California State Objective Mastery Test, compared to only 64 percent of the other class.

References

Adams, A. and E. Jones. *Teaching Humanities in the Microelectronic Age*. Milton Keynes: The Open University Press, 1983.

Asimov, I. "The Fun They Had," *Fifty Short Science Fiction Tales*. Edited by I. Asimov and G. Conklin. New York: Collier Books, 1963.

Boyer, E.L. *High School: A Report on Secondary Education in America*. New York: Harper & Row Publishers, 1983.

Brown, J. S. and R.R. Burton. "Diagnostic Models for Procedural Bugs in Basic Mathematical Skills," *Cognitive Science*, 2: 155-192, 1978.

Callahan, R. E. *Education and the Cult of Efficiency*. Chicago: University of Chicago Press, 1962.

Evans, C. *The Micro Millennium*. New York: Washington Square Press, 1979.

Foley, J. "Computerized Assessment of Writing for Instructional Improvement," unpublished paper delivered at the NCME Annual Convention, Chicago, 1985.

Goodlad, J. *A Place Called School: Prospects for the Future*. New York: McGraw-Hill, 1983.

Hall, D. "On Language," *New York Times Magazine*, July 14, 1985, 6-8.

Hoskin, K. "The Examination, Disciplinary Power and Rational Schooling," *History of Education*, 8 (2): 135-146, 1979.

Jackson, P. W. *The Teacher and the Machine*. Pittsburgh: University of Pittsburgh Press, 1967.

Lerner, L. *A.R.T.H.U.R., The Life and Opinions of a Digital Computer*. Amherst, Massachusetts: University of Massachusetts Press, 1975.

Lloyd, A. L. *Folk Song in England*. London: Lawrence & Wishart, 1967.

Macrae, N. *The 2024 Report: A Concise History of the Future, 1974-2024*. London: Sidgwick & Jackson, 1984.

McIrvine, E. "The Admiration of Technique." In R. Theobald (Ed.), *Dialogue on Technology*, 33-44. New York: Bobbs-Merrill, 1967.

Moseley, M. *Irascible Genius, a Life of Charles Babbage, Inventor*. London: Hutchinson, 1964.

Postman, N. "The Day Our Children Disappear: Predictions of a Media Ecologist," *Phi Delta Kappan*, 62 (5): 382-386, 1981.

Pynchon, T. "Is it O.K. to Be a Luddite?" *The New York Times Book Review*, (October 28,1984) 1:40-41.

Searle, J. *Minds, Brains and Science*. London: British Broadcasting Corporation, 1984.

Starch, D. and E.C. Elliott. "Reliability of Grading Work in Mathematics," *School Review*, 21: 254-259, 1913.

Stephen, L. and S. Lee (Eds.). *The Dictionary of National Biography*. London: Oxford University Press, 1973.

Turkle, S. *The Second Self: Computers and the Human Spirit*. New York: Simon and Schuster, 1984.

Vandenberg, D. *Human Rights in Education*. New York: Philosophical Library, 1983.

Weizenbaum, J. "Two Minutes with Mr. Chips," *Boston Magazine*, (May 1985) 27.