DOCUMENT RESUME

ED 080 539	TM 003 044
TITLE	Invitational Conference on Testing Problems (New York, November 1, 1969).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	1 Nov 69
NOTE	123p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Academic Achievement; *Achievement Tests; *Conference Reports; Educational Accountability; Evaluation Methods; Measurement; *Program Evaluation; Statistical Analysis; Systems Analysis; *Testing Problems

ABSTRACT

The 1969 Invitational Conference on Testing Problems had as its theme "Toward a Theory of Achievement Measurement." Papers presented in Session I, The Nature of Educational Achievement, were: (1) "Concepts of Achievement and Proficiency" by William E. Coffman; (2) "The Functions and Uses of Educational Measurement" by Winton H. Manning; (3) "Social Consequences of Educational Measurement" by Edgar Z. Friedenberg; and (4) discussion by Chester H. Harris. Papers presented in Session II, The Measurement of Educational Achievement, were: (1) "Validation of Educational Measures" by Lee J. Cronbach; (2) "Integration of Test Design and Analysis" by Louis Guttman; (3) "Knowledge vs. Ability in Achievement Testing" by Robert L. Ebel; and (4) discussion by Goldine C. Gleser. Papers presented in SessionIII, Measuring the Performance of Systems and Programs, were: (1) "Systems Analysis of Education" by Thomas K. Glennan, Jr.; (2) "The Role of Evaluative Research" read by Louis Guttman for the late Edward A. Suchman; (3) "Controlled Experimentation: Why Seldom Used in Evaluation?" by Julian C. Stanley; (4) "Accountability in Public Education" by Leon M. Lessinger; and (5) discussion by Michael Scriven and Robert H. Ennis. (KM)

PROCEEDINGS

of the

1969

Invitational Conference

on

Testing Problems

Toward a Theory of Achievement Measurement

Library of Congress Catalog Number: 47-11220
Printed in the United States of America

Invitational Conference on Testing Problems

November 1, 1969
Hotel Roosevelt
New York City

PHILIP H. DUBOIS
Chairman

EDUCATIONAL TESTING SERVICE
Princeton, New Jersey
Berkeley, California
Evanston, Illinois

ETS
BOARD OF TRUSTEES
1969-70

ETS OFFICERS

Foreword

The final Invitational Conference of the 1960s was built upon a theme that seems appropriate for the decade ahead: Toward a Theory of Achievement Measurement. Speakers in the morning session were concerned with the problems of defining and measuring educational achievement. They raised provocative questions about the meaning and essence of achievement, the purpose, effects, and function of achievement testing in our society, and the problems of content validation and test design.

The papers 'n the afternoon turned from achievement and the individual to the question of measuring the performance of systems and programs. Speakers explored topics such as the ways in which systems analysis can aid decision making in education, the application of systematic evaluation techniques to educational problems, and the ways in which the concept of accountability can be implemented in our public schools.

It was obvious from the overflow crowd at the first session that many felt these topics would make an exciting program. The speakers and the discussants who presented critiques of the papers confirmed that promise. A great deal of the credit for this exciting program must go to the chairman, Professor Philip DuBois, who worked so hard to organize it. We owe him a special debt of thanks. I should like to express thanks as well to the speakers and the discussants whose papers we are proud to publish in these *Proceedings.*

Henry Chauncey
PRESIDENT

Preface

As a major Ameri:an industry, education has an enormous capital investment, a huge payroll, and a tremendous expense budget, yet all too little in the way of product accounting and quality control. While the effectiveness of American education is ultimately to be gauged in changes not only in individuals but also in the nature of society, the attentioı: of employees (teachers and professors) is characteristically focused on processes rather than outcomes.

Educators, in fact, are divided as to the basis of evaluation. Some maintain that the grade at the end of a course should reflect merely the attainment of the student at a fixed point in time, irrespective of how that attainment was reached. A student who ends a language clas: with considerable prior knowledge might be awarded a good grade even though gain was negligible. Other teachers would attempt to evaluate change—that is, the amount learned during the course.

Theoretically there is a reconciliation o: this dilemma. If students at the beginning of a period of instruction were more or less equal both in ability and in attainment, relative standing at the end of the instructional period would reflect both gain and relative levels of achievement.

In practice, this procedure would require good achievement tests prior to instruction. This goal would not seem to be unrealistic since there are now in existence enormous pools of achievement items to which thousands of new items are added each year. Computer methods could certainly be devised to assist in the implementation of pretesting in language skills, mathematical skills, and the common subjects of instruction in grade school, high school, and the university.

A much greater limitation is in the division of students to be trained into comparable groups. It is doubtful whether this can ever be accomplished physically except to a limited extent in a few of our larger educational institutions. Nevertheless, the modern computer with its enormous potential for storing and comparing data might eventually come out with an evaluation of an individual's training in which he would be compared systematically with his peers undergoing similar training in other institutions throughout the country.

The concept of evaluation by means of achievement tests is not without its critics. The format of the objective item has sometimes been criticized as leading to superficiality of knowledge and inability of students to express themselves in connected discourse. Reliance on modern examinations has sometimes been the basis of charges that education is conducted too mechanically and without regard for social values. Nevertheless, it is a historical fact that the introduction of formal examining procedures in universities during the late Middle Ages and early years of the modern period coincided with university reforms that made possible higher education as it exists today. Similarly, it is only through the development of systematic testing procedures that our public schools have been able to deal effectively with student enrollment, which has multiplied so rapidly over the past five decades.

Educational Testing Service, which has sponsored provocative discussions of new trends and practices in its annual Invitational Conference on Testing Problems, has played an important role in American developments in evaluation. In fact, the prestige of this conference as a forum is such that it was relatively easy to assemble a group of speakers whose prestige as leaders in educational measurement is outstanding.

In developing this program, it seemed appropriate to consider some of the issues of achievement measurement that are still unresolved. The general topic "Toward a Theory of Achievement Measurement" breaks down into three main areas: definitions, functions, and consequences of testing; techniques; and problems related to evaluation of more or less complete educational systems. These discussions, reflected in the organization of the three sessions, involve numerous questions that will require continued discussion over the years. As partial solutions are reached, we can anticipate implementation of new concepts, instruments, and procedures.

Techniques of educational measurement, which in their current form are largely the product of the past 50 years, have had enormous influence in making mass education possible. Recent years have witnessed the development of measures that are applicable to large numbers of individuals at relatively low cost and, at the same time, yield more information usable in teaching th:n has been available hitherto. The fact that we can now compare educational achievement in widespread geographical areas and in diverse types of instruction has made possible for the first time an applied science of education.

Consider for a moment the innovations in achievement testing in the past five decades: the objective item, test standardization, automatic scoring, electronic reporting and statistical analysis, and—even more important—increasing sophistication on the part of the educators as to how measurement techniques may be applied in reaching educational goals.

Nor is the end in sight. Recent years have witnessed much development and experimentation in the integration of measurement into the planning of instruction, in the use of feedback in motivating and guiding the learner, and in determining what sorts of educational situations and hardware are most effective. We can expect that ways will contir,ue to be found in which to utilize modern technical developments, including computers, communication and recording devices, document reproduction, and transportation, in helping educators to attain the goals they set for themselves. As changes are introduced, both in the individual learner and in the social setting in which he learns, evaluation will be required. The future of the measurement of educational achievement seems assured.

The papers presented herein speak for themselves, and the reader wi'l find that they constitute real contributions to the literature of educational evaluation. I wish to acknowledge my indebtedness for the cooperation shown by all speakers and discussants as well as the continuing and indispensable help of Miss Anna Dragositz, who coordinated the project at all stages.

Philip H. DuBois
CHAIRMAN

Contents

Session I

Theme:
The Nature of
Educational Achievement

Concepts of Achievement and Proficiency

WILLIAM E. COFFMAN
University of Iowa

One of the episodes in the motion picture *Isadora* depicts the birth of Isadora's first child. There are two scenes. In the first, Isadora, obviously in pain and taking her usual positive approach to events, is demanding that the attending physician *do something*. The doctor, wearing the frock coat of the turn-of-the-century practitioner of the healing art, just stands there waiting for nature to take its course. The second scene opens as the newborn baby is placed into the arms of a tired but obviously triumphant mother who shouts, "I did it all myself! I did it all myself!'

The viewer of the film does not doubt that he has witnessed an achievement in the sense of definition 2a in *Webster's Seventh New Collegiate Dictionary:* "a result brought about by resolve, persistence, or endeavor." If he is a medical educator, or if, like me, he has agreed to speak at an Invitational Conference on Testing Problems, he might wonder whether the achievement represented the ultimate in proficiency. What, for example, would have been the effect if Isadora had been able to consult, prior to the labor, one of today's specialists in natural childbirth? Of if the attending physician had been more inclined toward intervention? I doubt, however, that Isadora at the moment of achievement was concerned with such questions. The labor was accomplished, the product was good, and the satisfaction was complete.

Now, one might draw a more or less extended analogy between the achievement of Isadora and achievement in today's schools depending on his educational philosophy and his conception of the nature of human learning. At this point, let us simply note that the achievement was a tangible product, that the achiever had had some choice in the matter of whether or not to risk pregnancy. that thc [illegible]er nad immediate knowledge of results. that the evaluation was made by the achiever herself, and that the judgment was absolute, not relative. The same cannot often be said in the case of achievement in school.

Those who undertake to propose goals of instruction in school are usually careful to be comprehensive; after all, the school is concerned with the well-being of developing human beings. and developing human beings have many legitimate needs. On the other hand, time is limited and so choices have to be made. Implicit in the choices of things to emphasize are the differing viewpoints about what achievements and proficiencies should be the outcomes of schooling.

One well-established viewpoint is that the schools exist primarily for the purpose of transmitting accumulated knowledge. Through successive generations, mankind has accumulated a vast store of knowledge about himself and his environment and has organized this accumulation in systematic ways that facilitate its transmission and use. It is the responsibility of the educator to abstract from this accumulation those elements that are of greatest significance and organize them into teachable units. The viewpoint is well expressed by Ausubel (2):

> Actually . . . the transmission of subject matter can be considered the primary function of the school. Most of the thinking that goes on in school is and should be supplementary to the process of reception learning, that is, concerned with having students assimilate subject-matter content in a more active, integrative, and critical fashion. Development of thinking or problem-solving ability can also be considered an objective of schooling in its own right, although it is a lesser objective than the learning of subject-matter and is only partly teachable; but under no circumstances is it a proper substitute for reception learning or a feasible primary means of imparting subject-matter knowledge.

Ebel (7), after recognizing the complexity of the problem of deciding what pupils should achieve in school, comes to much the same conclusion as Ausubel:

> If we look at what actually goes on in our school and college classrooms and laboratories, libraries and lecture halls, it seems reasonable to conclude that the major goal of education is to develop in the scholars a *command of substantive knowledge.* Achievement of this kind of cognitive mastery is clearly not the only concern of teachers and scholars engaged in the process of education. But the command of substantive knowledge is, and ought to be, the central concern of education.

For Ausubel, school achievement is marked by progress in the construction in the minds of students of an ever increasingly complex set of cognitive structures that will enable the student to interpret experience. By capitalizing on the work of generations of scholars, the school can short-cut the agonizingly slow process of building cognitive structures through direct experience. To a considerable extent, the abstractions that constitute the structure of the networks can be taught directly so that the bits of information can be incorporated without the necessity of wasting time trying to "discover" how things go together in meaningful ways. For Ebel, the task of the test maker is to construct questions that differentiate between automatic, rote verbalizations and responses reflecting meaningful relating of the questions to the structure that has been developed in the mind of the student.

This view of the purpose of schooling, often simplified or distorted, is probably held by a majority of teachers today. A systematic survey would probably show that pupils, too, think of school as a place where they are to learn subject matter, organized and communicated through textbooks, lectures, and discussions and examined by questions of one type or another for which there exist right answers that the informed can recall or figure out in one way or another. The frame of reference is generally that of some standard, more or less flexible, set by the teacher—or perhaps by some more impersonal "they" representing the authority of the school or the society. To achieve in this setting is to accumulate points—by answering questions in class, handing in homework, supplying answers to periodic quizzes, and writing a final examination. If one accumulates enough points, usually a certain percentage of all the possible points one might accumulate, one has achieved. If one accumulates fewer points, one fails and must have another try at it or withdraw from the competition. My recollection is that I thought of school achievement in this manner when I was passing through the system some 40-odd years ago. And my daughters did not appear to have a very different conception 30 years later in spite of subsequent developments in the field of testing. I recall vividly their discussions of the importance of accumulating 85 percent rather than 84 percent since that seemed to be the difference between a B and a C on the periodic report card. True, the school did administer standardized tests at periodic intervals on which were reported grade-equivalent or percentile scores, but these were strictly peripheral. The thing that counted was doing the "regular" class work.

Anybody who has taken the trouble to look closely at wha' goes on in schools knows that there is great variability in the extent to which pupils do learn what teachers attempt to teach them. There was a time when the problem could be solved by a process of attrition; those who didn't learn simply dropped out of school. Today, however, individuals stay in school even if they don't learn the subject matter or develop the skills taught at a particular level. So long as they do not become too troublesome, they are simply ignored. In extreme cases, however, such as in certain inner-city schools, teachers have become primarily disciplinarians leading the class through a caricature of meaningful learning. And even in more affluent settings, all is not well. A teacher like John Holt, who takes seriously his commitment to guiding meaningful learning, becomes an angry man as he struggles to reconcile the objectives of instruction with the facts of individual differences.

One solution to the problem is to throw out the concept of absolute standards and replace it with one of relative standards; individual differences are inevitable. And even the good student soon forgets much of the detailed content of instruction Identify the broad skills and understandings that remain after the specifics are forgotten and measure them with general examinations appropriate for a wide range of ability. Report scores in units that are related to the actual performance of reference groups of school children. Over time, it will then be possible to chart the progress of individuals in this frame of reference. This is the system characterized by ability grouping, some choice of courses at the secondary school level, and the school testing program as the monitor.

Some years ago I made a study of the achievement of two classroom groups in a small city school system. The groups were measured annually beginning in Grade 4 and continuing through Grade 8 with tests in the Stanford Achievement Battery. There were a few striking individual growth patterns, but in general, when fluctuations attributable to errors of measurement were discounted, the picture that emerged was one of constantly increasing scores on a gradient the slope of which was determined by the initial status. In other words, in this particular system, children seemed to be increasing in their ability to answer the kinds of questions in the test at a rate proportional to their initial level of ability. At the time I noted that even those in the lower third of the distribution seemed to be improving over time and commented: "It is also interesting to speculate on the attitudes which have been built up in this group of pupils, who, over a four-year period, have increased their scores as much as an average group would be expected to increase in about three and one-half years. They have been in the bottom third of their class. They have received the D's and F's. To my knowledge they have not had an opportunity to see plots of their growth lines. I suspect that periodic reports of scores on standardized tests might be more rewarding to them than periodic grades reflecting their position in the group" (6).

There's some question, however, whether it would be possible to substitute a record of systematic improvement in broad areas of cognitive skills for relative position in the group as a measure of success in school. I recall with deep concern the remark of one mother on hearing that I was the Director of the Iowa Testing Programs. "Oh yes," she commented, "some of my friends report proudly each year how their children have scored at the 99th percentile. I keep quiet, because mine never seems to be able to get above the 75th percentile." For her, and possibly for her children, success in school consisted of scoring at or near the top of the examinations, whether they were teacher-made tests graded on some absolute scale or standardized tests reflecting relative position in the group on broad intellectual skills. In his book *Schools Without Failure*, William Glasser (11) reports his findings with respect to this matter:

> From talking with many children over the past several years about grades, I find that they believe that the line between passing and failing in our grading system lies just below B; that is, a child who gets mostly C's is essentially a failure in school because the only real passing grades are B and A.

For Glasser and for Holt, the solution to the problem lies in fomenting a revolution that would transform the schools into markedly different institutions. Stop trying to plan in detail the content of the curriculum and pass it out in little doses to children. Stop grading, and stop the testing that is the basis of grading. Get rid of this whole rigid system. Toward the end of his book *How Children Fail*, Holt (12) writes:

> The alternative—I can see no other—is to have schools and classrooms in which each child in his own way can satisfy his curiosity, develop his abilities and talents, pursue his interests, and from the adults and older children around him get a glimpse of the great variety and richness of life. In short, the school should be a great smorgasbord of intellectual, artistic, creative, and athletic activities, from which each child could take whatever he wanted, and as much as he wanted, or as little.

Presumably, in such a system each child, like Isadora, would make his own evaluation after examining the product. Teachers, like those described by Phil Jackson in his book *Life in Classrooms* (14), would make their evaluations by examining the process rather than the product. If children were engaged in meaningful activities that experience had indicated might lead to useful learning, then the educational program in that classroom would be judged effective. Presumably, the administrators would obtain evidence for their evaluation by looking at the process of interaction of teachers and pupils. And given the drive of the human organism to make sense out of his world, it just might be that such a system could provide the society with enough scientists to keep the machinery going and expanding, enough artists to interpret and beautify the culture, enough politicians to manage the human interactions, and a host of happy people able to do their thing. There may be a place for freedom of choice in the market place of education as well as in the market place of the economic system.

It is likely, however, that the cost in a free market of educational activities would be as high as that in a completely unregulated free market in goods and services. Not all teachers are likely to have the sensitivity of a Holt or the skill in questioning of a Glasser or the ability to recognize good learning situations of the teachers observed by Jackson. Furthermore, how shall the content of the smorgasbord be developed? And how shall the output of the system be assessed if not in comparison with the performance of reference groups?

Over 10 years ago, Dorothy Adkins (1), after analyzing and evaluating the implications of Skinner's learning theory, proposed that the solution was in teaching each learner to the mastery level those materials that he is capable of mastering. Such a procedure would require a variety of teaching methods and more or less continuous progress appraisals based on tests of defined educational objectives. The work of Gagné (8, 9) offers one possible solution to the problem of developing suitable teaching methods and testing procedures. Identify, by detailed analysis, the hierarchical structure involved and aim test questions at key points in the structure, insuring that the prerequisites are mastered before proceeding to the next level. Glaser (10) has pointed up the implications of such an educational program for testing. If all students master a unit, the concept of item difficulty as the percentage of the group marking the correct answer and the concept of test score as relative position in the group becomes meaningless. What are needed are criterion-referenced measures: The pupil demonstrates achievement by answering the questions correctly.

In 1963, Carroll (5) suggested that perhaps the major factor accounting for individual differences in learning was not, as Adkins assumed, difference in the capacity of the pupils to reach higher and higher levels of cognitive organization but rather differences in the time required to incorporate and organize experience. If so, by individualizing instruction and by using carefully prepared unit materials, it should be possible to insure that all students achieve mastery of the units they have the time to complete. There should be no failure, only differences in the number of units completed. Bloom (4) states the expectations thus:

> Most students (perhaps over 90 percent) can master what we have to teach them, and it is the task of instruction to find the means which will enable our students to master the subject under consideration.

Does this mean the end of educational achievement testing as we have known it for many years? I doubt it. In the first place, it is not yet clear that children will necessarily accept the adult value system implicit in the subject-matter specialists' view of what ought to be taught. What if Snygg (15) is right in his view that phenomenal field theory is applicable to all school learning? Bloom has demonstrated that the concept of mastery learning is dramatically successful for the highly selected students at the University of Chicago who, like Isadora, have chosen to try for the prize. It will be interesting to see how he fares with a representative sample of elementary or secondary school pupils. Some may point to the reported success of Bereiter and Engelmann with disadvantaged preschool children (3). It's my impression, however, that they are simply providing the children with the minimum tools required to play the educational game so painfully described by Holt and Glasser.

Secondly, there's some doubt in my mind just how long the demonstrated mastery can be guaranteed to persist unless there are solid anchors in the world of ongoing experience. Once we have learned to drive a car, we don't forget, but then, every time we take hold of the wheel and press the accelerator, we receive immediate feedback. I imagine that the society will want some program of periodic testing to monitor even a system based on the concept of mastery learning.

There is a mountain of unfinished business in this area if we are to provide something more than a distorted view of a small aspect of the output as we do today through tests of cognitive skills and subject-matter knowledge. Lindquist (13) posed the challenge almost 20 years ago, and we have not yet found a way of meeting it. In his chapter in *Educational Measurement* he wrote:

> If the descriptions of educational development of individual students provided by tests are to be truly comprehensive, tests and measuring devices must be developed for many more educational objectives than are now being measured at all. In general, satisfactory tests have thus far been developed only for objectives concerned with the student's *intellectual* development or with his purely *rational* behavior. Objectives concerned with his non-rational behavior, or with his emotional behavior, or objectives concerned with such things as artistic abilities, artistic and aesthetic values and tastes, moral values, attitudes toward social institutions and practices, habits relating to personal hygiene and physical fitness, managerial or executive ability, etc., have been seriously neglected in educational measurement.

I would only add that we need to be concerned not only with what the schools are trying to accomplish—that is, the educational objectives—but also with what the unintentional concomitants are. We do not intend that an educational program produce fearful or deeply anxious children or teachers. We ought to know when it does if we are to reach valid conclusions about its effectiveness.

REFERENCES

1. Adkins, Dorothy C. Measurement in relation to the educational process. *Educational and Psychological Measurement,* 1958, 18, 221-240.

2. Ausubel, David P. *The psychology of meaningful verbal learning.* New York: Grune & Stratton, 1963. P. 13.

3. Bereiter, C. and Engelmann, S. *Teaching disadvantaged children in the preschool.* Englewood Cliffs, N. J.: Prentice Hall, 1966.

4. Bloom, B. S. Learning for mastery. UCLA, CSEIP *Evaluation comment,* May 1968, 1 (2).

5. Carroll, John. A model for school learning. *Teachers College Record,* 1963, 64, 723-733.

6. Coffman, W. E. Patterns of growth in basic skills in two elementary school classrooms over a four-year period. *The Seventeenth Yearbook,* NCMUE, 1960. Pp. 141-151.

7. Ebel, R. L. *Measuring educational achievement.* Englewood Cliffs, N. J.: Prentice-Hall, 1965, Pp. 38-39.

8. Gagné, R. *The conditions of learning.* New York: Holt, Rinehart and Winston, 1965.

9. Gagné, R. Learning hierarchies. *Educational Psychologist,* Nov. 1968, 6(1).

10. Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist.* 1963, 18, 519-521.

11. Glasser, William. *Schools without failure.* New York: Harper & Row, 1969. P. 63.

12. Holt, John. *How children fail.* New York: Pitman, 1968. P. 180.

13. Lindquist, E. F. (Ed.) *Educational measurement.* Washington, D. C.: American Council on Education, 1951. P. 137.

14. Jackson, P. W. *Life in classrooms.* New York: Holt, Rinehart and Winston, 1968.

15. Snygg, Donald. Another look at learning theory. *Educational Psychologist,* October 1963, 1 (1).

The Functions and Uses of Educational Measurement

WINTON H. MANNING
Educational Testing Service

The main argument presented in this paper is that the functions and uses of educational measurement that have developed in the past are insufficient for the future because they have been too much shaped by the practical problems of educational institutions to the neglect of other functions. The institutional problems that have demanded priority in the past are primarily those that arise from our having viewed the educational system as a training resource designed to supply the manpower needs of industry. Although this aspect cannot be wholly overlooked if we are to maintain a highly technological society, it is nevertheless unsatisfying, particularly if we believe that educational measurement offers the best means we have for effecting improvement in the quality of the educational process. Aside from the rhetoric that is employed, differences in the uses of tests in education as compared with industry or the military are probably more apparent than real, and this fact alone should signal that something is wrong.

If there is merit in this argument, then it follows that we should be particularly concerned with identifying and fostering the development of measurement functions in education that are uniquely appropriate to the needs of young people and to the traditions of rational, objective, scientific inquiry into the process by which young people are educated. Put another way, two issues that arise when one examines the functions of educational measurement are:

1. Because educational measurement is oriented mainly toward the solution of practical problems of educational institutions, its functions have been those of providing a means for accomplishing certain tasks of social and educational engineering—that is, successively sorting people into hierarchies of talent and accomplishment for the world

of work—rather than as an instrument in the construction of educational theories that are amenable to scientific investigation.

2. Educational measurement has been mainly employed in the solution of problems confronting educational institutions as they seek to shape human resources for economic development, rather than in the solution of problems of individuals as they seek to use the resources of educational institutions for self-development. As a consequence, testing has not progressed as far as it should as a means for assisting students to encounter successfully those problems of self-understanding, choice, and decision making that they confront as maturing individuals in a modern technological society.

This way of looking at testing suggests there exists an imbalance in the development of educational measurement—a practical, or a-theoretical, bias on the one hand, and an institutional, or a-personal, bias on the other. In taking this view, my intention is not to disparage the traditional uses of educational measurement, or even to view these as becoming less important in an absolute sense in the future. However, forces do exist that seem likely to extend the functions of tests into new directions in the future. Two of these are:

First, opportunities afforded by educational technology—particularly computer-assisted instruction, testing, and guidance—seem likely to make the development of scientific educational theory more feasible and, hence, more attractive. Correspondingly, if we are wise enough to use it that way, the new cybernetic technologies also offer us the opportunity to devise measurement procedures that more effectively serve the individual student for purposes of his own development. In brief, we have potentially within our grasp the means for employing measurement directly in the interests of students, rather than indirectly through the presumably beneficent ministrations of mediating institutions or agencies.

Second, the revolution in values, attitudes, and beliefs of young people throughout the world augurs for profound changes in educational institutions at all levels, and as a consequence, the way in which educational measurement is employed may shift from the traditional institutional concerns to new problems and, hence, to new functions. I believe these revolutionary social forces may well serve to undermine the practical and institutional biases that presently too much characterize the functions and uses of educational measurement; or to be more exact, I hope that this will be one of the consequences.

With this general introduction, let me turn to a brief discussion of some of the older uses of measurement in education and to a consideration of some functions that seem likely to grow in importance.

Traditional Uses of Educational Measurement

As Henry Dyer (4) pointed out several years ago, we can readily identify at least three important functions that educational tests have been des'gned to serve in the past. These are:

1. *Selection and distribution* in which tests are used as a basis for the selection of students for programs, or the distribution of students among programs, and where the distribution system is aimed at providing an optimal match between student abilities and limited educational resources
2. *Diagnosis or prescription* where tests are used as a basis for identifying the nature and extent of educational deficiencies, and for prescribing educational treatments designed to remedy these deficiencies, thus aiming at maximizing the number of children reaching a given level of achievement
3. *Evaluation* where tests are used to assess the effectiveness of educational programs so that there is a systematic basis for comparison of educational outcomes and, hence, for the improvement of educational practices

It would be tempting to consider how the uses of tests for selection, diagnosis, and evaluation have evolved over the past 50 years and to speculate on how these functions are likely to change in the decades ahead. At the level of higher education, for example, demands for open admissions to college are causing not only an intensive study of the effectiveness and equity of tests that are used in selection but also a profound re-examination of the moral and ethical bases for the process of selection itself (12). Similarly, the concern for equalizing educational opportunity has led to enhanced demands for tests that diagnose educational deficits, particularly of disadvantaged, minority children, in ways that will permit prescriptive rather than random efforts at remediation (5, 10). Finally, the concern with evaluation of educational programs has grown dramatically in the past decade, partly as a consequence of the promise that this approach holds for improving the educational process and partly as a result of the increased demand for accountability in education (11).

There is little question that these functions of tests will continue to be among the major pivots around which measurement research and testing programs will continue to revolve in the years ahead. However, there are other uses of educational measurement that are important and which I should like to call to your attention: 1) the uses of tests in the development of educational theory, and 2) the educative or guidance function of educational measurement.

Educational Tests as Instruments of Educational Theory

Measures of educational outcomes may be conceptualized as real traits that are acquired as a consequence of instruction. Aside from determining whether the sample of behavior displays reasonable consistency, the validity of such measures is established by assessing the extent to which the proposed interpretation of the test corresponds to some real trait or, in other words, the investigation of the construct validity of the measure. To do this requires study of the content of the items, the interrelationships or structure that exists among the items, and the relationships of the test responses to behavior that is ınanifested external to the test itself. In Loevinger's analysis (9), for example, the three components of construct validation are described in terms of: 1) the substantive validity of the test (which for achievement tests may be seen as equivalent to the problem of content validity); 2) the investigation of the structural validity of the tests or the cxtent to which test items parallel the structural relations of other manifestations of the trait being measured; and 3) the external validity of educational tests, or the degree to which the test is related to behavior displayed outside the testing situation.

Seen from this perspective, the development of a theory of educational achievement is not materially different from the task of developing personality theory or psychological theory. The problem of educational testing as a theoretical instrument rather than as an applied technique is, it seems to me, central to many of our other concerns, and is indeed the basis for the motivation of most behavioral scientists who · ıoose to study education.

The concept of a "trait" is, of course, essential to the whole operation, and it is here that educational theory confronts some particularly thorny problems. The objectives of education, when stated in terms of specific behaviors, often focus more upon products than upon processes of behavior. Although the taxonomies of Bloom (2) and Krathwohl (7) appear to be much concerned with assessing the various ways of acquiring and using knowledge, there is reason to assert that educational measurement practices tend to subordinate the question of describing *how* one goes about seeking a solution to a problem, and to enhance the goals of teaching students to display acceptable solutions. Whether this is so or not, there is little question that educational measurement has been mainly developed as a technique for evaluating the quality of outcomes, rather than for describing the characteristic strategies that individuals use in reaching those outcomes.

The reasons for this are many, including particularly the origins of educational testing practices within a meritocratic selection framework, and the noticeable tendency in educational research to avoid studying behavior that depends as much upon affective as upon cognitive processes. Furthermore, if progress through the educational system is seen as a competitive race for the rewards that society can bestow on the successful, there is reason enough to understand why we have to emphasize measurement of traits of ability or accomplishment to the exclusion of other interesting characteristics of the learner.

At the level of early childhood education, the contrast is quite evident when one compares the behavior that is sampled by conventional intelligence tests and the approach to measurement that, stemming from Piaget's work, has characterized the project known as *Let's Look at Children* (8). Here the emphasis is not on whether the child has learned to perform such acts as stringing beads or defining words, but rather upon understanding the kinds of cognitive processes that the child uses when he is confronted with an interesting and challenging problem that requires an explanation of the way in which he sees his world.

Obviously, a major deterrent to the development of tests that permit us to observe problem-solving strategies has been the awkwardness and expense of such procedures. For example, tab tests and similar approaches have not seemed, in most cases, to be dramatically better than conventional tests for conventional purposes, and, hence, it is difficult to justify their far greater expense.

The great potentiality of the computer as a medium for testing seems to be in the capacity it affords for real-time interaction with the subject and the consequent ability we acquire to record objectively the strategy that the individual uses in seeking a solution as well as its outcome. However. as some have pointed out, serious questions arise when one considers computer-based, problem-solving tests. What aspects of behavior we should look at, and how these observations should be combined to yield quantitative or qualitative descriptions that are useful, are not always immediately evident. For most persons. the characteristic way in which we go about searching for solutions to a problem may have more significance for their later performance of social and occupational roles than the degree to which we have mastered the content of a discipline or subject in school. From the standpoint of education, the development of tests that are oriented toward assessment of problem-solving styles would have, therefore, in my judgment, a salutory effect on educational practices. Furthermore, the possibility is growing that educational objectives will again come to be specified in terms of narrow product-oriented behaviors, leading ultimately to a kind of neo-positivism of the classroom. An important deterrent to such excesses of naive behaviorism may be that, through development of measures of both process and outcome, we will be able to construct theories of human behavior that will lead to more faithful statements of educational objectives and, hence, to more fruitful hypotheses about educational treatments. Richard Atkinson (1) has suggested that the lack of a real theoretical foundation for the teaching of reading was not clearly evident until the problem of devising a program for teaching reading to children by means of computer-assisted instruction was confronted. I hope that something analogous to this may occur to theories of intellectual growth when we confront the problem of assessing problem solving within the framework of a computer-based test.

The search for consistencies in problem-solving strategies and the attempt to devise and test educational theory through construct validation of such tests will require as much attention to Loevinger's second and third components—structural validity and external validity—as to her first component, substantive or content validity. Among other benefits, not the least would be that of placing educational measurement squarely within the same methodological domain as that which has characterized psychological test theory for some years.

In dwelling briefly on the function of educational tests in theory construction, I have called your attention to a function of tests that has always been recognized but has been more often a pretension than an actuality. My reason for emphasizing this use of tests derives primarily from the widely shared conviction that in the coming decade educational programs are likely to change in dramatic ways not only as a consequence of technological developments associated with the computer but also in response to revolutionary social forces aimed at restructuring social institutions. Greater attention to the problem of educational theory and the relationship of educational tests to theory would, therefore, seem to be a productive and stabilizing influence in the turbulent years that lie ahead.

Educative and Guidance Functions of Measurement

Some future social historian may well characterize this century as one of great optimism, in which the increased perfectability of the human condition was assumed without question, and where science and technology were seen by civilized men as the means for bringing about continuous human advancement. Education has, if anything, embraced this view with even greater ardor than other social institutions, with the ultimate consequence that something like one quarter of a billion tests are administered each year to students enrolled in our educational institutions. In many respects this enthusiasm for tests seems to rest more upon blind faith than upon observable benefits to the consumer. In nearly every school, for example, the yellowing pages of unused but carefully stored printouts of test score rosters bear mute witness to a problem that is characterized less as an information overload than as the totemism of test use.

However, the educational uses of computer-based information storage, handling, and retrieval systems coupled with the prospective establishment of education networks offer the possibility of extending the concept of a test to include more effective interpretations of the meaning of the test performance. For many years the need to provide better interpretative information than norms, grade equivalents, expectancies, or probabilities of success has been recognized, but despite the good intentions of testing agencies, the rate of misuse or disuse of test results has continued to mount.

The concept of educational measurement embedded in a system that incorporates a delivered interpretation of the meaning of the test performance will require some profound adjustments in testing and guidance practices. As such systems develop, questions arise as to what boundaries, if any, exist between the uses of tests in counseling and guidance and the technology of educational measurement itself. Conceiving of tests as components in computer-based measurement and guidance systems leads to the assumption that the boundaries must become blurred, if not nonexistent. It is, therefore, exactly this kind of fusion of functions that I am pointing toward in discussing still another use of educational measurement—namely, *the educative or guidance uses of tests.*

The educative use of tests suggests that evaluation of student achievement, attitudes, and values through educational measurement should carry with it the obligation to portray the individual in terms that will permit him to learn more about himself through an analysis of his own performance rather than primarily through comparisons with other groups of students. The uses of tests for institutional ends, such as selection, have fostered a notion that evaluation of performance is only possible under conditions of competition, or what B. A. Thresher called "adversarial" testing. But if the use of test information is centered on the processes of self-understanding and self-discovery by the student, achievement testing that is epistemological rather than adversarial would seem to be the natural result. A good example of this is provided in computer-assisted instructions where we see the development of criterion-referenced tests that are embedded in the interactive instructional programs.

Applying this viewpoint to systems of guidance information, such as those being developed by Martin Katz at Educational Testing Service and David Tiedeman at Harvard, raises issues concerning the function of most forms of educational testing in the future. For example, consider the familiar question of how test information can be interpreted "so that students can choose among various alternatives more realistically." As Katz points out, the question itself may be a false or misleading one, for a person rarely confronts a situation in which there is truly a fixed set of alternatives; rather "he often has some opportunity to construct or create his own options" (6). Furthermore, there is, as Katz eloquently describes, an enormous difference between seeing the task of guidance as that of helping students make "wise" or "realistic" decisions through choosing the best alternatives—in other words, those with highest probability of pay-off—and the task of assisting students to become wise in the processes of decision making. In this sense, there is an interesting correspondence between emphasis on *processes* in educational theory and a comparable emphasis in guidance theory.

Whether we are talking about computer-based systems of guidance that are student-centered, or about measures of problem-solving strategies as instruments of educational theory, the observations made by Jerome Bruner concerning the kind of education we need for the future may be relevant. Writing a year ago, Bruner (3) said:

> " . . we shall probably want to train individuals, not for the performance of routine activities that can be done with great skill and precision by devices, but rather to train their individual talents for research and development which is one of the kinds of activities for which you cannot easily program computers. Here, I mean research and development in the sense of problem solving . . . What this entails for education is necessarily somewhat obscure, but its outlines may be plain. For one thing it places emphasis on the teaching of interesting puzzle forms—ways of thinking that are particularly useful for converting troubles into problems . . . for converting chaotic messes into manageable problems . . "

If tests are to become integrally embedded in information systems designed to foster a sense of planfulness, orderliness, and continuity within the lives of students—or, in other words, if measurement is to assist in the process by which civilized men seek to convert into manageable problems the chaotic messes that ignorance of self and powerlessness seem to decree for them—the mission for educational measurement is indeed formidable. Whether educational measurement can become as proficient a servant of individual human beings as it has been the handmaiden of educational institutions is an important issue that deserves serious debate and creative energy.

Conclusion

In summary, we have identified three major functions of educational measurement that have developed in the past: selection, diagnosis, and evaluation. It was asserted that these uses of tests arose primarily from institutional needs of the educational system although their use by institutions may indirectly have also served the needs of students. Two functions of tests that deserve particular emphasis at this time are: first, the uses of educational tests in the construction and evaluation of educational theories, especially theories that give particular attention to processes or strategies of problem solving rather than outcomes alone; and second, the uses of tests in the service of individual students through systems of guidance that employ measurement as a means of fostering self-discovery and as a means for encouraging students to develop wisdom in decision making.

Development of these testing functions will require that educational measurement become integrally involved in both instruction and guidance, particularly in those approaches that utilize the unique capacities of the computer for interaction and objectivity. The search for means of expanding the functions of testing within the context of educational technology may also have the effect of reinforcing a humane use of modern technology rather than simply extending the mechanical efficiency of present functions of educational measurement.

Four hundred years ago Montaigne wrote of education in his day:

> We labor only to stuff the memory, and leave the conscience and understanding unfurnished and void. Like birds who fly abroad and forage for grain, and bring it home in beak without tasting it themselves to feed their young, so our pedants go picking knowledge out here and there . . . holding it out at tongue's end, only to spit it out and distribute it abroad.

Pedantry is not confined to the classroom; it can exist within the confines of a machine-scorable answer sheet as well. New uses for tests in the years ahead must not simply be passive responses to the needs of the educational system, but energetic efforts to extend and elevate the functions of educational measu :ment.

REFERENCES

1. Atkinson, Richard C. Computerized instruction and the learning process, *American Psychologist*, 1968, 4, 225–239.

2. Bloom, B. S. (Ed.) *Taxonomy of educational objectives.* Handbook I: Cognitive Domain. New York: Longmans, Green and Co., 1954.

3. Bruner, J. S. Culture, politics, and pedagogy. *Saturday Review*, May 18, 1968, 69–72, 89–90.

4. Dyer, H. S. The functions of testing—old and new. *Testing Responsibilities and Opportunities of State Education Agencies.* Proceedings National Testing Conference for State Education Agency Personnel. The University of the State of New York, The State Education Department, Division of Educational Testing, Albany, New York. 1966. 63–79.

5. Gordon, E. W. and Wilkerson, D. A. *Compensatory education for the disadvantaged.* New York: College Entrance Examination Board, 1966.

6. Katz, M. R. Can computers make guidance decisions for students? *College Board Review,* No. 72, Summer 1969. 13-17.

7. Krathwohl, D. R., Bloom, B. S., & Masia, B. B. *Taxonomy of educational objectives.* Handbook II: Affective Domain. New York: McKay, Inc., 1964.

8. *Let's look at children.* Princeton, N. J.: Educational Testing Service, 1965.

9. Loevinger, Jane. Objective tests as instruments of psychological theory. *Psychological Reports,* 1957, 3, 635-694. (Monograph Supplement No. 9).

10. Manning, W. H. The measurement of intellectual capacity and performance. *The Journal of Negro Education,* Summer, 1968.

11. Messick, S. *Evaluation of educational programs as research on educational process.* Proceedings of the American Psychological Association, September, 1969.

12. Wilson, L. *Merit and equality in higher education.* Presented at the Annual Meeting of the American Council on Education, Washington, D. C., 1969.

Social Consequences of Educational Measurement

Edgar Z. Friedenberg
State University of New York at Buffalo

Though its manifold unplanned consequences are probably more important, educational measurement has two generally recognized functions. In a society as hung up on competitive achievement as ours, testing emphasizes the assessment of individual competence. Then it assigns test scores, and with them the persons who made them, to ideologically acceptable social categories. The second of these functions is much the more important of the two, since individual competence is not generally esteemed in our society; nor is it, over the long haul, demonstrably and consistently the major factor determining relative success. It does, of course, become crucial for every individual on certain occasions; but educational testing is not very helpful in assigning particular individuals to particular social roles because the norms on the basis of which their scores might have been interpreted have been grossly contaminated by the factors that have been operating in the meanwhile to keep competence from obstructing the social process.

Preoccupation with individual test scores, though understandable in the individuals being tested, today I believe serves chiefly the ideological function of convincing the young that the American social system recognizes and rewards individual competitive achievement. This has induced them to cooperate in the testing program because they expect it to serve as the gateway to opportunity, based on a precise assessment of their individual merits. But, for the managers of the social system it has a contrary function. Testing is their means of stocking their various manpower pools with compatible varieties matched in predacity and adaptive characteristics so that they will not be troubled later by conflict or random variation. Greater precision and subtlety, in the interests of a just appraisal, scarcely concern them.

Paradoxically, as the ideological impact of testing on students declines, the real rewards for successful test performance may increase. As higher-status, more sophisticated incumbents reject the social system the schools serve, and the concept of success that prevails within it, they will be replaced by lower-status youth who are still hungry for the rewards it offers, and willing to agree that they *are* rewards. It may well be that the ends of social mobility and equality of opportunity are best served by a system of rewards so sickening that the successful are ultimately forced to hasten, tight-lipped and intoxicated, from the arena to make way for those who press upon their heels. Certainly, our educators are fortunate in having still available a pool of some twenty million black people most of whom, having been denied access to these rewards, remain more firmly convinced of their value than their oppressors are. It is understandable that beleaguered college presidents find the angry demands of blacks less threatening and humiliating than the derision of radical whites who treat them like a bad joke that has run on too long.

It is frustrating, therefore, that just when blacks are desperately needed as the last available, large, and untapped pool of candidates for socialization into the American middle class, a serious question should have been raised as to whether they are as capable, on the average, of making it. The issue raised by Jensen (2) in the *Harvard Educational Review* last winter was perceived as a threat by liberals from the moment his article went into galleys.

Jensen's argument, to be sure, is not really very startling. The chromosome being what it is, there are certainly clusters of genetic characteristics that come to be socially defined as racial; and it is surely plausible that certain of these characteristics should be related to cognitive functioning. The proposition is rendered more plausible, moreover, by the fact that ideological inhibition has impeded scrutiny of this possibility by American social scientists. Kurt Vonnegut Jr. is not really being funny when he observes in *Slaughterhouse-Five* (6):

> I think about my education sometimes. I went to the University of Chicago for a while after the Second World War. I was a student in the Department of Anthropology. At that time, they were teaching that there was absolutely no difference between anybody. They may be teaching that still.

They may, indeed, but if they are, Jensen is not the only scholar to think they might be wrong. Thus, Gerald Lesser and Susan S. Stodolsky (5), found consistent differences in patterns of mental ability among Chinese, Jewish, Negro, and Puerto Rican first graders in New York that were quite independent of social class. Their findings that Jewish children ranked significantly better than all the other ethnic groups in verbal ability, though as damaging, in view of the prevailing social stereotype, to the Jewish image as anything Jensen has to say about Negroes, caused neither astonishment nor consternation.

Where Jensen is on weakest ground—and ground onto which Lesser and Stodolsky were never tempted to precede him—is in his inference that the differences he cites are not merely ethnic but genetic in origin. In a society as permeated by discriminatory practice and perception as ours, this is hardly an empirically testable proposition, since the possible effects of racism penetrate every situation—certainly in the form of health and nutritional factors, they penetrate the womb—and color every parameter that might be observed. In our society, being black works the way Abe Martin said being poor did, 40 years ago: It ain't a crime, but it might as well be. The experience of stigmatization is simply so pervasive that even attempts as conscientious as those Jensen makes to factor-out its effects remain unconvincing—especially when all he has to work with are data gathered for other purposes than to test his hypotheses and hence incapable of being well-controlled for his purpose. I was hung up, for example, by his citation of evidence that Amerindian children, though more disadvantaged than black children on all citable environmental indices, still do significantly better in school until I realized that this was a classic example of empiricists' folly; and of exactly the kind psychometricians should beware: assuming that what you can assess on some scale, if you repeatedly get consistent results, must somehow be critical even if it isn't exactly relevant. For the Indian children on whom such comparisons are based are precisely those who still live in communities that, though squalid and poor, retain some supportive sense of a tribal tradition that, in our society, is romanticized now rather than denigrated. To be denigrated, you have to be a "nigra;" the difference is qualitative, and there is no use mucking about with scales to measure it.

Neither Jensen nor his opponents can escape the ethnocentrism implicit in their respective positions by the purifying rituals of science, for those rituals are themselves central to their common ideological position. Both start with the assumption that our society, with its dominant values, is the given—the reality with which one must come to terms—and that those who do not accept its terms or meet its demands have no just basis for demands upon it. Within this context, Jensen is,, of course, threatening because he infers that blacks are systematically less able, on the average, to meet those demands than whites; and that efforts to compensate for this difference by education could not be wholly successful, even if racial discrimination could be eliminated. Since both informal observation, like Jonathan Kozol's (3) and James Herndon's (1) and formal studies like Elenor Leacock's (4) and the Bundy report suggest that racial discrimination could only be eliminated from our urban school systems by genocide, the policy implications of Jensen's work seem rather remote. But if success, as society defines it, is what black people want—and most doubtless do—this is bad news, though hardly tragic. Jensen specifically insists that his conclusions cannot justifiably be applied, pejoratively or otherwise, to the cognitive possibilities of any individual. Most people, regardless of ethnic differences, have far more potential ability of every kind than they ever get to use; wherever the mean may be there is room to stand erect under almost any part of the curve. Only a very bad, or disingenuously racist, statistician would infer from his argument that any particular black student would be incompetent to meet the demands of the educational system, though one might quite properly conclude from it that those demands, if consistently inappropriate to the cognitive style most comfortable for black students, do indeed constitute a form of de facto discrimination. But a much more relevant question arises concerning the nature of those demands.

If I read Jensen correctly, he implies that in any large random group of blacks one would be likely to find fewer persons than among a comparable group of whites—and for purely ethnic reasons, leaving aside the question of opportunity—capable of developing the abilities of, say, Robert McNamara, or John McCone, or John Gardner, or Roger Heyns—to name four men whose excellence and consistent devotion to rational cognition undistracted by excessive passion or subjectivity have become a matter of public record and have brought them international distinction and groovy positions of public trust, if that is quite the right phrase for a director of the Central Intelligence Agency or the World Bank. This may be true; I fear it is. There may *never* be a black man with the kind of mind needed to produce a report like that of the McCone Commission on the disorders in Watts a few years ago. But a more fundamental question, surely, is whether and to what degree a person, white or black, must possess such a mind in order to live in this society with a reasonable guarantee against insult and prospect of satisfaction. Must one master the technique of affect-free cognition in order to succeed? Must one succeed in order to retain any shred of self-esteem? Must success be defined as the power to dominate and manipulate the life-styles of others, masked in the rhetoric of pluralism and equality of opportunity? Poverty, brutality, and exclusion are hard to bear; and no race can be so different—or so irrational—as to choose them. Jensen, however, suggests that there may be groups of human beings who cannot become quite like the dominant class in America, even if their lives, or more precisely their life-chances, depend on it.

If this is true, the process of educational measurement cannot have much relevance to their aspirations. Educational measurement is an inherently conservative function, since it depends on the application of established norms to the selection of candidates for positions within the existing social structure on terms and for purposes set by that structure. It cannot usually muster either the imagination or the sponsorship needed to search out and legitimate new conceptions of excellence which might threaten the hegemony of existing elites. Educational measurement is at present wholly committed to the assumption that legitimate forms of learning are rational and cognitive and that such learning is the proper goal of academic process. This is an ideological, not a technical, difficulty. It is perfectly possible to detect and appraise, in critical though not scalar terms, poetic skill, humane sensitivity, breadth and subtlety of human conpassion and the like. Educational Testing Service has, in the past, already done some of this in experimental revisions of testing programs in the humanities and in personality assessment, while the analysis of profiles of student activists prepared by the American Council of Education is very useful to college admissions officers in defining criteria by which potential militants among candidates for admission might be identified. Like chemical and biological warfare in relation to public health, this is merely humanism in reverse and a classic example of our commitment to ethical neutrality combined with devoted service to old customers and the power structure. But it is enough to prove that the technology of testing could serve humane goals in a more humane society.

But a more humane society might have little use for large-scale educational measurement because such a society would perforce be less competitive and universalistic and more generous and genuinely pluralistic than ours. I must stress that the reason is ideological, not technical. Educational measurement is technically capable of as great or greater service in helping people find out what kinds of knowledge, or what kind of job, or what college, or even what life would suit them best as it is in serving the competitive ends of society. For a people already ankle-deep in moonshit and happy ever after in the marketplace, even those computer-arranged courtships may be a form of salvation, or at least no joke. But society is not about to buy such a diagnostic use of testing on anything like the scale that it buys a competitive use because under its current ideology, only competitive testing can assign subjects to ideologically defensible categories for social action.

The major premise of the American system of social morality is that every individual should have an equal opportunity to compete for the prizes offered. The less frequently stated, but probably more crucial minor premise is that, if he does, he has no other legitimate basis for complaint. The contest may be destructive or banal, the prizes worthless, and the victory empty or pyrrhic; but to complain of these things is to be a bad sport and perhaps even an elitist, and such complaints are not honored in our system. It is most important, however, that every contest be objectively judged, as impersonally as possible, with no favoritism, nepotism, or any other kind of ism. To make this objectivity evident, access to preferred categories should, wherever possible, be granted on the basis of scaled scores that a machine can handle.

This is a very important dynamic in maintaining the American illusion of objectivity, since it permits the biases useful in maintaining our status structure and our institutions to be hidden beneath several levels of abstraction. Beneficiaries of the system, like middle-class college-bound students and their parents, or even its naive victims, like the more old-fashioned students in "general" or "commercial" tracks, may assume that there is no bias in the tests since there can be none in the scoring. Educators and more sophisticated students and parents, including a growing number of those of lower status, are aware of the problem of bias in the tests themselves, and may demand a "culture-free" test or the use of different norms in grading "disadvantaged" groups or the suspension of the testing program itself. But even they are unlikely to recognize the bias inherent in the very practice of basing judgments that may determine the entire life of a youngster—and, in view of our selective service policy, his death as well—on the display of a narrow range of cognitive behavior, quite apart from any question about the content of the test items. And virtually none recognize the value judgment involved in the monstrous, though familiar, decision that the welfare of any human being in a society with the means to nurture all its members should depend on any test score at all.

The widespread use of educational measurement, in short, reinforces our commitment to universalism, and our conviction that equality is the core of justice and that, moreover, equality is to be assessed by quantitative measurement in presumably scalar units. The value of this set of assumptions in the process of domestic counter-insurgency can hardly be overestimated. Every high school principal, college admissions officer, selections board for fellowship, or employment recruiter lives under continual menace from losers ready to accuse him of favoritism and the impediment of an egalitarian ideology which prevents him from making effectively the obvious rejoinder that favoritism, in the sense of allowing himself to be guided by his human and subjective perception of the needs and qualities of others, is a part of his professional responsibility. Partly for this reason, the use of testing has proliferated far beyond any expectation that the data it yields are needed to make any rational decision. They are needed, rather, to justify decisions for which no data were needed and to get administrators off the hook for having made them by showing that, however insensitive, uptight, and uncritical they might have been, and however willing to serve a corrupt master in a dubious cause, they are fair and impartial and play by the rules and have really nothing to answer for.

Fundamentally, the reason Jensen is so disturbing to liberals is, I believe, because his analysis threatens this basic stabilizing function of educational measurement and, with it, universalism itself. For if he is right, no amount of fairness and psychometric ingenuity can afford equality of treatment. Instead, one must choose between fairness and justice, and if a commitment to justice is to be preserved and finally implemented, the educational system must manage, against the pressure of so-called backlash, the difficult political decision to be generous and, as they say in the South, "partial." There is no reason to suppose that we find ourselves on this earth for the purpose of performing well on tasks involving abstract cognition or, indeed, for any purpose at all except our own. Educational measurement can serve those who want help in making an estimate that is more precise than one they could make themselves of what purposes might be realistic for them in view of their actual characteristics and the actual scenes they might make. It can also help them in legitimating their demands that they be permitted to make those scenes against possible social opposition; Educational Testing Service could, for example, provide SAT and Achievement Test scores in support of applications for college admission by high school dropouts, pushouts, and troublemakers who have poor recommendations and possibly no degree. It can do all this and more quite as proficiently as it can continue to assist in the grand process of channeling by which our society meets its manpower needs, sustains its status system, and brings peace of a kind to Southeast Asia. What it cannot do is get that society to authorize this process and pay for it.

REFERENCES

1. Herndon, James. *The way it spozed to be.* New York: Simon and Shuster, 1968.
2. Jensen, Arthur R. How much can we boost I.Q. and scholastic achievement? *Harvard Educational Review,* Winter, 1969.
3. Kozol, Jonathan. *Death at an early age.* Boston: Houghton Mifflin Co., 1967.
4. Leacock, Eleanor. *Teaching and learning in city schools.* New York: Basic Books, 1969.
5. Lesser, Gerald and Stodolsky, Susan S. Learning patterns in the disadvantaged. *Harvard Educational Review,* Fall, 1967.
6. Vonnegut, Kurt Jr. *Slaughterhouse-Five.* New York: Delacorte Press, 1969. P. 7.

DISCUSSION

Chester W. Harris
University of Wisconsin

As a discussant, I intend to draw on these papers as a source of issues or potential points of dispute and simply offer these issues to you in my own words. If it succeeds, this retelling of what you have only now heard will prompt new propositions, rejoinders, and perhaps questions. I assume that all, including the speakers, will be eligible to participate.

The nature of educational achievement cannot be analyzed successfully without considering the purposes of education and of the society (singular) or societies (plural) within which that education is conducted. It may not be possible to define achievement in a fashion that does not conflict sharply with certain humanistic and humane values; if so, the purposes of education must be thought of more as promoting discovery than as stimulating learning. If it is possible to define achievement as becoming as well as being and experiencing, then the range of types of desirable achievement becomes an issue—or a set of issues. A wide range permits the inclusion of non-cognitive or affective as well as cognitive or rational types of intended outcomes of the educational process. Such a wide range is suggested in the quotation from Lindquist which Mr. Coffman gave us. A narrow range—particularly one emphasizing "rational cognition undistracted by excessive passion or subjectivity"—seems to characterize the schools of our time, and some would alter this. However, others see a narrow range, with a marked emphasis on cognition, as the only realistic and appropriate one.

The question of who decides what should be the proper set of achievements looms larger today than it did only a few years agr. The local community presses to influence the answer to this question; students press for a "relevant" education in which the intended achievements are those they regard as appropriate. Closely associated with this is the questioning of the kinds of evidence of achievement that are to be gathered and communicated. The gathering of any kind of evidence itself has effects, not all of which may be anticipated The nature of educational achievement cannot be analyzed successfully without also considering the nature of that which is to be admitted as evidence of achievement.

Testing, which can be considered to include modes of systematic observation of behavior that go somewhat beyond choosing among written answers to written questions, is neither a villain nor a hero. It is a pawn and can be employed to serve the establishment, as both Mr. Manning and Mr. Friedenberg suggest. This may be the chief criticism leveled at testing. But our testing practices clearly tend to neglect the non-rational aspects of achievement, to neglect the detection of unintended and harmful outcomes—the unexpected side effects —and to neglect the observation of the process aspects or modes of behavior as opposed to products of behavior. There are old purposes of testing and possibly some new unrealized ones. Mr. Manning seems to be an optimist when he assigns testing an important role in the construction of educational theories. How one derives educational theory from data—which is not the same thing as testing theory by an appeal to data—is not yet very clear.

I have now succeeded in mentioning the name of each of the three speakers. I shall stop here.

Session II

Theme:
The Measurement of Educational Achievement

Validation of Educational Measures*

LEE J. CRONBACH
Stanford University

I am taking this occasion to introduce you to main ideas from a chapter prepared for the forthcoming Thorndike-edited *Educational Measurement.* Having that invitation to reexamine validity theory was rare good fortune: One does not often get a chance to revisit the sins of his youth and make up for omissions.

Almost 20 years ago, a group of us were asked by the American Psychological Association to prepare standards for psychological tests. Shortly thereafter, the National Council on Measurement in Education and the American Educational Research Association proposed to set committees to work on standards for *educational* tests. Having two or three sets of standards seemed likely to nullify the whole effort, so Paul Mort, then AERA president, organized a collaborative committee structure. This structure produced the 1954 Technical Recommendations for Psychological Tests, and, in 1955, an achievement-test version. The latter elaborated the recommendations, but did not look educational measurement square in the eye. In retrospect, I cannot say that we were wrong to push the educational problems aside—the committee had quite enough already on its plate. Validity theory for achievement measures probably had to wait until the proposals on aptitude and personality tests were assimilated.

When I say now that the committees failed to think through the logic of validation in education, Professor Ebel has every right to say "I told you so." He was in the unfortunate position of being added

*The material in this paper is taken principally from the author's chapter in the forthcoming revision of *Educational Measurement,* edited by R. L. Thorndike, to be published by the American Council on Education in 1970. Copyright 1969 by American Council on Education, used by permission.

to the original joint committee after the bobsled had already picked up speed; he had little alternative save to complain briefly about the path we were on, and then throw his weight into helping us make as good a trip as possible. He, Professor Lindquist, and a few others have consistently maintained that the *Standards* place too much emphasis on empirical validation and not enough on judgment (5). Another school of critics (1) has objected to the departure of the *Standards* from a strict operationism. I think perhaps now I see how operationism, empirical validation of construct interpretations, and judgments by educators fit under the same tent (though not into the same ring).

My proposed formulation owes a good deal to seminars with staff members of Educational Testing Service and especially to penetrating questions from Professor Coffman. It owes to numerous readers of draft manuscripts and to the study of decision making and generalizability in which Dr. Gleser and I have collaborated. I hope, then, that this synthesis comes close to the view of many wise colleagues. The statement is not a new and competing set of "standards." It supports and perhaps illuminates the existing *Standards* while pointing out crucial questions that no *Standards* and no effort by test publishers alone can cope with. *Validation is the task of the test interpreter.* Others can do no more than offer him material to incorporate into his thinking.

The logic of validation for educational tests is not different from that for psychological tests. Construct validation applies to many achievement tests, especially those of higher mental processes. Content validation applies to many psychological measures, notably attitude scales and observations of behavior. How one is to validate depends not on the test but on one's purpose in using the test. Since virtually no test is confined to a single purpose, it is illogical to speak of test validity. What one has to validate is a proposed interpretation of the test; for any test, some interpretations are reasonably valid and others are not.

One further preliminary remark: The testing movement has given too much attention to comparative interpretations (to individual differences) and too little to absolute, content-referenced measurement. Comparison (competition) is a theme straight out of John Stuart Mill and Charles Darwin. But evaluation of social programs and self-direction by individuals call for absolute judgments. Regarding a training program, what fraction of the graduates can perform the tasks they should? Regarding the student choosing a college major, what are the fields in which he has an active, sustaining interest? To answer such questions, tests must make absolute statements. Comparative ranks are irrelevant; in the ideal situation, everyone earns a high mark. The educator makes many absolute descriptive and predictive interpretations; the traditional, differential validity coefficients are not pertinent to these.

Table 1 outlines the formulation. What evidence is called for, and what judgments, depends on the nature of the interpretation. There is testing for decision making and testing for the purpose of describing a person or group.

While descriptions ultimately are used for decisions, any one description, such as that given by a beginning-of-year reading test, contributes to a great many decisions by the teacher and perhaps by the pupil. So the descriptive report should convey truthful impressions to the teacher, or to the pupil himself, or to whoever uses it.

DECISION RULES BASED ON TESTS

In this extract, only a little space can be given to the validating of decision rules. Decision making in education is best illustrated in the selection of applicants for advanced training and in the allocation of pupils to curricula or to different instructional schemes.

Validation of a decision rule logically requires an experiment in which, after being tested, persons are allocated to treatments *without regard to the scores* whose usefulness is being validated. The outcomes of the treatment are then appraised. There are usually many outcomes important to the decision maker, and a multidimensional criterion is preferable to a single one.

Every report of validation against a criterion is to be thought of as carrying the warning clause, "Insofar as the criterion is truly representative of the outcome we wish to maximize . . ." The report has to contain a clear description of the criterion and should contain a critique of it by the investigator. The reader must school himself to examine criteria with a hard eye, to convince himself that a test that predicts the stated criterion will also predict the outcome *he* is seeking. The tests that predict one outcome will often not be those that predict another, and prediction formulas that maximize one outcome may reject persons who would be outstanding by another criterion. In selection research one must continually resist the temptation to focus

Table 1

Summary of Types of Validation

Focus of investigation	*Question asked*	*Use made of student-response data*	*Use made of judgment*
1. *Soundness of descriptive interpretations*			
Content validity	Do the observations truly sample the universe of tasks the developer intended to measure (or the universe of situations in which he would like to observe)?	Scores on test forms constructed independently may be compared	To decide whether the tasks (situations) fit the content categories stated in the test specifications. To evaluate the process for content selection, as described in the manual.
Educational importance	Does the test measure an important educational outcome? Does the battery of measures neglect to observe any important outcome?		To compare the test tasks with the educational objectives stated by responsible persons.
Construct validity	Does the test measure the attribute it is said to measure? More specifically, the description of the person in terms of the construct, together with other information about him and the theory surrounding the construct, implies what can be expected of him in various situations; are these implications true?	Scores are compared with measures of behavior in certain other situations. Or, the test is modified experimentally and changes in score are noted.	To select hypotheses for testing. To integrate findings so as to decide whether the differences between persons with high and low scores are consistent with the proposed interpretation. To suggest alternative interpretations of the data.

II. *Usefulness for decision making*			
Validity for selection	Do students selected by the test perform better than unscreened students?	Regression of outcome measure on test score is examined.	To decide whether the criterion fully represents the outcomes desired, including outcomes more distant in time. To decide whether a new situation is enough like the validation situation for the results to generalize.
Validity for placement	Is performance improved when students are allocated to treatments according to their test scores?	Regression slope relating outcome measure to test score for one treatment is compared with that for another treatment.	To decide whether the criterion fully represents the outcomes desired, including outcomes more distant in time. To decide whether a new situation is enough like the validation situation for the results to generalize.

on criteria that are easy to predict. Attention should go to those that are most important.

With regard to selection decisions, modern thought places increased stress on local validation, validation on demographically distinct subgroups, and validity generalization.

A study that predicts school success by a statistical formula has direct significance when the formula is developed in the locale of the proposed application and the situation is sufficiently stable that the findings are representative of what will happen in succeeding years. Only if the supply of applicants and the curriculum remain much the same in character are the findings likely to remain directly applicable. Extrapolation is involved when a validity study is taken as warrant for continuing to use a test a decade later, after circumstances have changed. Far more hazardous extrapolation is involved in taking a published validity study made in a distant institution as warrant for one's local decisions. The legitimacy of an extrapolation to new conditions cannot be judged by statistical means.

It is good practice, where the sample size is sufficient, to treat separately the data for boys and girls, for whites and Negroes, and for subgroups differing markedly in previous preparation. Not infrequently the predictive significance of a score differs from subgroup to subgroup. But complex problems of policy arise in using subgroup statistics. If it were statistically valid to use different tests or different cutting scores for boys than for girls (for example), it would be difficult to convince applicants that sex-linked decision rules were not discriminating unfairly against one sex or the other.

A particularly satisfactory way of organizing input-output data is the "expectancy table," which reports the distribution of outcomes for persons having any particular pretest score. Decision theory requires emphasis not on a validity coefficient but on a regression slope relating the outcome measure to the test score. If the regression slope is great enough, outcomes for selected men are distinctly better than for unselected men, and the test is valid for selection. Its utility depends not just on the correlation between test and criterion, but also on the importance of the decision.

The placement model is the pertinent one when the school is concerned with the consequences of its policies for all the persons under consideration. Let treatments be labeled A and B and express the respective outcomes Y_A and Y_B on a common utility scale, since rational examination of a placement decision is not possible until the outcomes have been expressed in the same units. Again, let X be the predictor. There will be two expectancy tables, one for A and one for B, and corresponding regression functions. The regressions may align in various ways. The utility of the test is to be judged by examining the average outcome among persons distributed into treatments on the basis of test scores against, as a baseline, the outcome among persons who are indiscriminately assigned to the one treatment that is best on the average. A "validity coefficient" indicating that test X predicts success within a treatment *tells nothing about its usefulness for placement.* Comparison of regression slopes is the indispensable information. Placement decisions, I would argue, are more important as a use of tests than selection, but we seem to have no adequate examples of validation of placement procedures. Investigations of aptitude-treatment interactions are required, and the practical difficulties in that kind of research are great. As yet we know next to nothing substantive about which person variables interact with educational treatment variables. Hence, while something can be said regarding the logic of research on placement, actual validation of this kind is still over the horizon. A reasonably extensive discussion of some of the perplexities in research on interaction is given in a report available from ERIC (3).

DESCRIPTIVE INTERPRETATIONS

Three major questions arise regarding descriptive interpretations (see Table 1). 1) A description may be that and almost nothing more. ("Iames cɑn recognize 80 percent of the words found in freshman textbooks.") The only question is whether the test tasks are a proper sample of the domain referred to. 2) A description may include a value judgment. "James has done well in first-year Spanish" implies that the test is measuring what the listener wants to have taught. A certain printed test in Spanish may be an entirely valid sample of some stated domain, but the domain excludes auditory and oral skills some educators would want to develop. The second validity question is whether the right domain was selected. 3) Descriptions imply predictions and explanations. The interpreter who moves from task language to attribute language invokes constructs. To say that an examinee is anxious, or appreciates painting, or communicates clearly is to suggest what he is expected to do under various circumstances that may arise later. When the description is freighted with implications, the validity question is: Are the implications true?

Content-referenced Interpretations

A content interpretation refers to a universe of tasks or of observations. The universe description is an operational definition that restricts the admissible range of instruments, questions, settings, examiners, and so on; even the narrowest definition identifies not a unique operation but a class of operations. An operation is specified when one refers to use of "the Wechsler Block Design materials," but this is a class of instruments; it has thousands of members. The only indispensible requirement in a universe definition is clarity: Reasonable observers must agree as to what falls within the universe and what is excluded. (If the observation is a composite—a test covering several content categories—the requirement applies to the subcategories.)

Content validity has to do with the test as a set of stimuli and as a set of observing operations. The measuring procedure is specified in terms of a class of stimuli, an injunction to the subject that defines his task (what he is to try to do with the stimuli), and an injunction to the observer (rules for observing the performance and reducing it to a score). Judgments about content validity should be restricted to the operational side of testing—that is, to the explicit procedures of measurement. Interpretations regarding the subject's internal processes are to be validated not by judgment but by empirical studies. With regard to the Watson-Glaser Test of Critical Thinking, for example, it is a matter of content validation to have a qualified person judge whether the authors did indeed assemble problems of the sort they called for in their specifications. To ask the judge whether the problems actually elicit "critical thinking" is to solicit his speculations about construct validity.

In principle, validity of the selection of content is to be judged without considering at all the persons to be tested; attention is restricted to the test materials and the universe description. If the content fits the universe definition, the test is content-valid for persons of all kinds. From an absolute point of view the score on a task indicates that the person does or does not possess, in conjunction, *all* the abilities required to perform it successfully. A dictated spelling test is a measure of hearing *and* spelling vocabulary *and* ability to write. In terms of content, however, the spelling test tests ability to spell from dictation. The pupil who is deaf will earn a low score, but that score is a valid report of his inability to spell from dictation.

Professional constructors of achievement tests combine a content outline with a set of response-process categories, such as recall, reasoning, and application of principles. Such a specification has value in broadening the test, but it tends to confuse task operations controlled by the tester with processes presumably used by the subject. The usual content-by-process grid is not a universe specification in our sense. An item *qua* item cannot be matched with a single behavioral process. Finding the answer calls for dozens of processes, from hearing the directions to complex integration of ideas. The shorthand description in terms of a single process can be justified only when one is certain that every person tested can and will carry out all the required processes save one. Even to speak of "required processes," however, is misleading, since the task can perhaps be performed successfully in a variety of ways. In a universe definition, a proper response specification deals with the result a person is asked to produce, not the process(es) by which he succeeds or fails.

Content validity is necessarily limited by the inadequacy of the universe specification, which is usually couched in imprecise, everyday terms and can rarely mention every pertinent aspect of the task. Content is an ill-shaped and undifferentiated mass, hence there is a danger of vagueness in any reference to a content universe. Moreover, while there may be a definable domain of content, there is no existing universe of items. The only items in existence are likely to be those that constitute the so-called sample. It must be acknowledged that writing items to fit a content domain does not closely resemble the drawing of beans from an urn. But the central requirement is only that universe boundaries be well defined; this requirement of operational definition can be met. It is not essential that a universe be denumerable, or explicitly catalogued.

What, now, would constitute a rigorous validation of the fit between the operational definition of the universe and the actual test operations? To stimulate thought, one can suggest an experimental validation through duplicate construction. The construction would involve judgment, but the validation would employ completely hard data.

In principle, the rules for selecting test content can be described so fully that there is virtually no uncertainty as to what domain of tasks is to be sampled from. One would ordinarily make a test by a process of item writing, review, tryout, and revision. The experimental verification of a claim of content validity would call for a second team of equally competent writers and reviewers to work independently of the first, according to the same plan. They would be aided by the same definition of relevant content, sampling rules, instructions to reviewers, and specifications for tryout and interpretation of the data as were provided to the first team. In other words, they would work from the same operational definition of admissible procedures. If the universe description and the sampling are ideally refined, the first and second tests will be entirely equivalent. Any person's score will be the same on both tests, within the limits of sampling error. A favorable result, on a suitable broad sample of persons, would strongly suggest that the test content is fully defined by the written statement of the construction rules. An unfavorable result would indicate that the universe definition is too vague or too incomplete to provide a content interpretation of the test.

Test construction is never so logical as this. Ambiguity remains in many definitions of universes, and reviewing of draft items is an art not reducible to rules. No one has ever carried out the two-team study. It is not at all uncommon, however, for the test developer to claim validity by construction, bolstering the claim with a detailed account of the construction process. The test manual may list the textbooks from which content for items was chosen or may display the specifications given to the item writers. The reader is left to judge for himself whether this definition is explicit enough to allow two independent teams to arrive at approximately interchangeable tests.

Content validity is impermanent. The items or tasks in the test reflect social events, job descriptions, accepted beliefs about the world, decisions about what the curriculum should cover, and so on. These change with the passage of time, so that sooner or later the test becomes unrepresentative. The prospective user must be satisfied that a second team following the specified procedure *today* would arrive at a test reasonably like the original.

Correlations have nothing to do with content validation. Nothing in the logic of content validation requires that the universe or the test be homogeneous in content. The topics in the motor vehicle code are diverse: hand signals, right of way, reporting an accident, and so on. To make a decision about an applicant for a license, it is necessary to know whether he would pass a certain proportion of the items belonging to the universe defined by the code. If the items have low correlations (or if they vary in difficulty), it will take a larger sample of items to be confident that the subject's universe score reaches the required level. But, no matter how heterogeneous the universe, with enough items one can estimate the universe score as precisely as desired. Low item intercorrelations do not necessarily imply failure of the test content to fit the definition. Indeed, if the universe is heterogeneous, consistently *high* item intercorrelations imply inadequate sampling.

Correlations between tests are irrelevant to content validity (except in the construction experiment). Some critics are inclined to object to the creation of separate tests or scores for performances that correlate highly. But even if there is a large correlation between, say, a measure of acquaintance with chemical-bond theory and a measure of ability to apply chemical-bond principles, there is justification for keeping the measures separate. First, the absolute level of attainment of one objective might be much higher than that of the other; and this could suggest a need to modify the curriculum. Second, though the items correlate at the end of the instruction currently being given, some new instructional procedure might develop one competence while neglecting to develop the other. Keeping the categories separate in the list of objectives at least reminds all concerned to entertain such a possibility when evaluating the new program. This matter is discussed further in connection with construct validation, where correlations *are* relevant.

Evaluative Interpretations

When observations at the end of instruction are used to determine how successful some educational activity has been, the interpretation embodies value judgments. Hence the validity of an evaluative conclusion depends on the value question: Did the tests appraise the qualities I consider it most important to teach?

That question might elicit a positive answer from one educator and a negative one from another looking at the same tests. A content-valid test cannot satisfy decision makers who hold values unlike those of the test developer. Consequently, an ideally suitable battery for evaluation purposes will include separate measures of all outcomes the users of the information consider important.

The recommendation that the evaluation battery be comprehensive seems to run counter to the concept that an educational test should measure what has been taught. And students think a test "unfair" when it asks about topics not covered in the course. One can agree that it is unjust to let the fate of an individual be determined by a test that, through no fault of his own, he is ill-prepared for. But this only illustrates once more how a test valid for one decision can be invalid for another. Though it is unfair to judge the quality of a teacher's work by a test that does not fit the course of study he was directed to follow, that test may be a fair basis for judging the curriculum. If teacher-plus-course-of-study have left the pupil ignorant on some important matter, that is a significant fact about the adequacy of his education.

Sometimes a test can "fit the curriculum" entirely too well. The universe pertinent in summative evaluation is the universe of tasks graduates are expected to perform. To be sure, a curriculum developer who has a restricted objective can use a restricted test to determine how well he achieved *his* end. But if other educators considering adoption of the course desire broader outcomes that go beyond his aims, they will find such restricted studies inadequate.

Interpretations Employing Constructs

Whenever one classifies situations, persons, or responses, he uses constructs. Every time an educator asks "But what does the instrument really measure?" he is calling for information on construct validity. Constructs help us to interpret both measures used to appraise educational outcomes and measures to forecast response to instruction. The relevance to education of personality constructs such as authoritarianism may be granted readily. It is perhaps less obvious that construct validation is relevant for tests of subject-matter learning. Many phrases used to characterize commonplace educational tests appear to describe mental processes: "scientific reasoning," "reading comprehension," and so on. If such a term is amplified to specify a class of tasks, the interpretation can be limited to content interpretation. Interpreters, however, usually consider processes behind the score.

Consider reading comprehension as a trait construct. Suppose that the test presents paragraphs each followed by multiple-choice questions. The paragraphs obviously call for reading and presumably contain the information needed to answer the questions. Can a question about "what the test measures" arise? It can, if any counterinterpretation may reasonably be advanced. At least eight such counterhypotheses have to do with the possible effect on test score of motivation, style of work, speed, and other characteristics of the person. The test may be content-valid, in that it presents reasonable tasks; but perhaps it cannot be validly interpreted as measuring a comprehension skill, distinct from reading speed, vocabulary, and so on.

To validate an interpretation using a construct, one investigates the effect of each disturbing influence pointed out by the counterhypotheses. Construct validation is difficult to explain because so many diverse techniques are required to examine diverse hypotheses and counterhypotheses. Construct validation requires the integration of many studies (4).

Construct validation begins with the claim that a given test measures a certain construct. This claim is meaningless until the construct is amplified from a label into a set of sentences. When the test interpreter says, "John Jones is high on trait X," he implies many things about Jones. The sentences that generate those implications spell out the meaning of the construct. In principle there is a complete theory surrounding the construct, every link of which is systematically tested in construct validation. While something like this does happen as theory evolves through an endless succession of studies, investigations are far less systematic than this. The test developer (or some later writer) proposes a certain interpretative construct, explains at greater or less length what the construct means, and offers *some* evidence that persons scoring high on the test also exhibit other behavior associated with the construct. The initial report is usually far from convincing; the sophisticated reader will think of alternative ways to account for the test behavior.

If the construct interpretation is taken seriously by the profession, its validity is challenged over and over again. The challenge consists of proposing a counterhypothesis—an alternative construct to account for the test behavior in whole or part. While one could carry out construct validation by a plodding verification of every sentence written about the construct, the work would be interminable. It is the plausible counterinterpretation that directs research toward a possibly vulnerable part of the theory.

Procedures used to examine trait or process interpretations fall into three broad categories: correlational, experimental, and logical. Correlational studies determine how persons who score high on the test differ, in everyday life or in the laboratory, from those low on the test. Several types of correlational studies are mentioned below. The experimental study attempts to alter the person's test performance by some controlled procedure. If it can be shown, for example, that procedures designed to increase a child's confidence raise his score on an information test, this challenges the interpretation of the test as a measure of information alone. A logical analysis of the test content or the scoring rules may disclose disturbing influences in the score. A simple example is the observation that a certain outcome measure is invalid because the test has a low ceiling, so that pupils who do well on the pretest can gain only a few points at most.

Correlational Studies

A construct that can be measured by only one procedure is likely not to be very interesting. When we can invent several diverse procedures whose reports agree well with each other, the construct is significant (2). Thus, if reading comprehension is our construct, we would like to see convergence among tests with multiple-choice response, tests of recall, and tests in which the subject carries out acts for which the test paragraph gives directions. Convergence is shown by correlations across persons within groups, and by correlations across groups, whether the groups are demographic or are the product of experimental manipulations. Indicators of one construct should ordinarily have low correlations with measures interpreted in terms of other constructs. If two tests are very similar in the information they give, it complicates theory to retain two trait names for them.

Among techniques for studying convergence and divergence of indicators is factor analysis. A factor analysis of even a large number of measures of educational outcomes is likely to report only a few factors. This is too often interpreted as implying that the several outcomes "are not really different." Comprehension of physical laws will certainly correlate with ability to reason scientifically because, in a general population, those who have studied science will do better on both types of test. Even if the study is confined to persons who have studied physics, the correlation will remain high because the ablest students will have made greatest progress along both directions. If the high correlation means that there is no distinction between comprehension and reasoning, one could not criticize a curriculum for emphasizing the laws and making no effort to promote reasoning.

At first glance there appears to be a head-on conflict. The curriculum reformer argues that comprehension and reasoning are distinct attainments, and the correlational study proves that whoever is best in one respect is best in the other. The contradiction is resolved by a distinction between a within-group correlation and an across-groups correlation. Within a group completing the same course of study, the two variables correlate. But suppose the class averages for 50 classes are determined, and a correlation across groups is computed from these 50 pairs of values. The curriculum reformer who contends that some teachers neglect to develop reasoning is predicting that this across-groups correlation will be fairly low, that some groups will rank high on comprehension but not on reasoning. Even if the correlation across groups turns out to be high, the reformer has a tenable position to retreat to. If he can design a curriculum that concentrates on scientific reasoning, whose graduates score exceptionally well on the reasoning test while scoring at the norm on the comprehension test, he has proved his point. The high correlation across groups meant only that present curricula are holding the balance between reasoning and comprehension so nearly constant that the best programs (or those drawing the ablest students) get the best results on both dimensions.

Constructs as Educational Objectives

The formal rationale for construct validation sees a construct as defined by a network of relations, all of which are anchored to observables and so are testable. This rationale has been widely accepted in psychology, but its use in education needs further explication. The operationists who want to equate each construct with "one indicator" —rather, with a narrowly defined class of procedures—are advocating that we restrict descriptions to statements of tasks performed or behavior exhibited and are rejecting construct interpretations. Surely, however, the choice of interpretation is the prerogative of the investigator; a type of interpretation productive in one context may be sterile in another.

The writers on curriculum and evaluation who insist that objectives be "defined in terms of behavior" are taking an ultraoperationalist position, though they have not offered a scholarly philosophical analysis of the issue. The person who insists on "behavioral" objectives is denying the appropriateness and usefulness of constructs. The educator who states objectives in terms of constructs (self-confidence, scientific attitude, the habit of suiting one's writing style to his purpose) regards observables as indicators from which the presence of certain dispositions can be inferred. He will not, however, *substitute* "volunteers ideas and answers in class" for "self-confidence." From the construct point of view, behavior such as this is an indicator of confidence, not a definer. No list of specific responses-to-situations, however lengthy, can define the construct, since the construct is intended to apply to situations that will arise in the future and cannot be specified now.

Nearly all current philosophy of science, even the operationism of Bridgman, makes use of constructs embedded in networks. But one still encounters such statements as Ebel's: "If the test we propose to use provides in itself the best available operational definition, the concept of validity does not apply" (5). But this language gives the game away, for the "best available" definition is presumably not the best conceivable, and "How good is the operation?" remains a meaningful question.

The issue raised by the ultraoperationalists is possibly just a terminological one, since there seem to be few differences of opinion about how tests and test interpretations can and must be used (6). There is universal agreement that general propositions embodying descriptive concepts must in the end be verified by means of systematic observation, and that the procedures used to gather these observations must be given an adequate operational description in order to make the report useful.

The person planning instruction or choosing among courses of study has to think in terms of concepts that describe behavior in a broad class of situations. One of the tasks of social science is to seek the right breadth for its concepts (7). "Citizenship" is no doubt too broad; "ego-strength" is a good deal better, since it leads one to anticipate different behavior in situations all of which might be thought of as calling for citizenship. One cannot expect, at least in this century, to disentangle ego-strength from interacting traits and situational variables, and so long as each measure is subject to interactive effects, no one measure can be accepted as a standard.

One can retreat to very narrow concepts; citizenship could be broken down at least to the level of "participation in elections" and "obedience to speed laws." This would increase the number of variables beyond the point where they could be investigated, and would leave out of the discussion whatever behavior citizens exhibit in the less standardized aspects of their lives.

The most serious criticism to be made of programs of construct validation is that some of them are haphazard accumulations of data rather than genuine efforts at scientific reasoning. To merely catalogue relations between the test under study and a variety of other variables is to provide a do-it-yourself kit for the reader, who is left to work out his own interpretative theory. Construct validation should start with a reasonably definite statement of the proposed interpretation. That interpretation will suggest important counterhypotheses, and these also will suggest data to collect. Investigations to be used for construct validation, then, should be purposeful rather than haphazard. After collecting his data, the investigator is expected to integrate the hypotheses and findings with each other and to offer a final conclusion as to the soundness of the construct interpretation and the influence of impurities that have been identified.

CONCLUSION

Valic' tion of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate. The person validating a test should give thought to all questions suggested in Table 1, though the relative importance of the questions varies from test to test. The several kinds of study shed light on each other. Thus, criterion-oriented studies generate a theory of individual differences and a theory of tasks and situations. In the light of such constructs, one makes reasonable judgments about the design of new educational situations and the design of new measuring instruments. Since these judgments, in turn, need to be validated, the process of investigation, and therefore the growth of knowledge, never ends.

Responsibility for valid use of a test rests on the person who interprets it. The published research merely provides the interpreter with some facts and concepts. He has to combine these with his other knowledge about the persons he tests and the assignments or adjustment problems that confront them to decide what interpretations are warranted.

REFERENCES

1. Brodbeck, M. Logic and scientific method in research on teaching. In N. L. Gage (Ed.). *Handbook of research in teaching.* Chicago: Rand McNally. 1943. Pp. 44–93.

2. Campbell, D. T., and Fiske, D. W. Convergent and discriminate validation by the multitrait-multimethod matrix. *Psychological Bulletin.* 1959 56. 81–105.

3. Cronbach, L. J and Snow, Richard E. *Final report: individual differences in learning ability as a function of instructional variables.* Stanford University, 1969. Not available from the authors. May be ordered from ERIC Document Reproduction Service by requesting ED-029-001.

4. Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, 52, 281–302.

5. Ebel, R. L. Must all tests be valid? *American Psychologist,* 1961, 16, 640–647.

6. Hochberg, H. Intervening variables, hypothetical constructs, and metaphysics In H. Feigl and G. Maxwell (Eds.). *Current issues in philosophy of science.* New York: Holt, Rinehart and Winston, 1961. Pp. 448–456.

7. Nagel, E. *The structure of science.* New York: Harcourt, Brace, and World, 1961. Pp. 505–508.

Integration of Test Design and Analysis

Louis Guttman
The Hebrew University of Jerusalem and
The Israel Institute of Applied Social Research

Five years ago this Invitational Conference turned out to be the occasion for the development and first presentation of new results that were a surprise even to their author (2). Since then, these results have been replicated and extended in several published (7, 9) and unpublished studies, and indeed are true even for the data (Guilford's) used earlier to illustrate the use of our faceted definition of intelligence (1). It is gradually becoming common knowledge that if a battery of tests is constructed (or selected) according to two particular major facets, then the battery's intercorrelation matrix will tend to have a radex structure. One of the facets is the language of communication, with the three elements or varieties: verbal, numerical, and figural (geometrical). The other facet is the type of task imposed on the subject, with two elements: rule-inferring and rule-applying.

If other possible facets beyond these two are held relatively constant in the test construction or selection, then the empirical correlation matrix can be represented quite simply in a two-dimensional Euclidean space. Each test is represented by a point in the space, and the distance between two points decreases monotonely as the correlation between the corresponding two tests increases. Perhaps more important than the small dimensionality is the law of formation related to the facet design. Each of the languages of communication corresponds to a different direction from the origin, altogether partitioning the space into three wedge-like regions which can be labelled respectively "verbal," "numerical," and "figural." Similarly, there is a partitioning of the space corresponding to the second facet, but in a different manner. Points inside a circle around the origin correspond to rule-inferring tests, and points outside this circle correspond to rule-applying tests.

Thus, the two facets together provide a polar coordinate, or radex, framework for viewing the empirical space of the test interrelationships.

The latest published example (from which I have borrowed Figure 1) suggests adding a third element to the task facet—namely, "school achievement" (9).

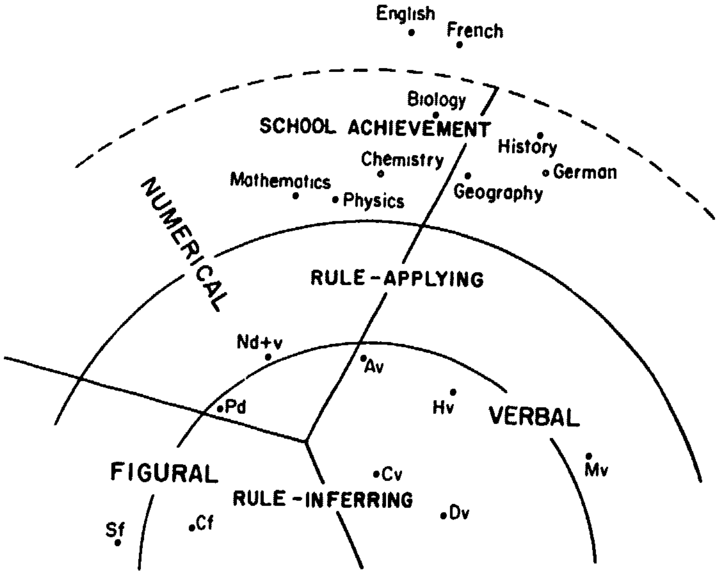
The three wedge-like regions are indicated in Figure 1 for the languages of communication, but three task bands around the origin are indicated rather than two. Rule-inferring is at the center, with rule-applying in the next band. A further, outermost band has been added for "school achievement." It turns out that school achievement tests are in the numerical and verbal wedges, according to the course matter being taught. We have not had the occasion to see these tests in order to inspect the content of the items, but it may be assumed that they are of the ordinary school variety. In light of the above results, these tests may be emphasizing learning a kind of rule-applying where the rule is taught formally in the school system. The kind of rule of the other two regions is ordinarily not formally taught in textbooks and is usually not in the ordinary classroom framework. It may therefore be useful to distinguish a further facet for classifying rules—namely, by the extent to which they are formally taught in school. Considering this additional facet, the battery of tests might be regarded as being defined by a mapping sentence of the following form:

The performance of student (x) on an item presented in $\begin{matrix}\text{verbal}\\ \text{digital}\\ \text{figural}\end{matrix}$ language and requiring $\begin{matrix}\text{inference}\\ \text{application}\end{matrix}$ of a rule $\begin{matrix}\text{exactly like}\\ \text{similar to}\\ \text{unlike}\end{matrix}$ one taught within one of his school courses ⟶ $\begin{matrix}\text{high}\\ \\ \text{low}\end{matrix}$ performance.

In this example, the third facet concerns levels of similarity of the rule to what is taught in school, and may serve to distinguish between predictors and criteria of achievement in school. The tests in the two inner bands of Figure 1 may be thought to be predictors of the tests in the outer band. Regardless, it is the rule-application tests (not exactly like those taught in school) which best predict the school achievement. The rule-inference tests correlate less with this criterion. Is this merely a result of the test construction for assessing school achievement, or

Figure 1

The 18 Variables of Höger's Study Portrayed in a Two-space



Variables Included in Höger's Study

Code	*Description of variables*
Rule-inferring	
Cv	Complete one missing word in sentence
Dv	Find which word is different from given set of words
Av	Word analogies
Hv	Give subordinates of two words (e.g., rose-tulip)
Pd	Numerical progressions
Cf	Find which of five geometric figures (circles, squares, etc.) can be put together from given parts of figure
Rule-applying	
Mv	Subject memorizes 25 words, each belonging to one of the following categories: flowers, tools, artifacts, birds, animals: then he is asked questions of the following form: The word beginning with the letter a was: . . . (a flower, a tool, a bird . . .)
Nd + v	Verbally formulated arithmetic problems
Sf	Match cubes presented in different orientations in space
School-achievement	German History Geography English French Mathematics Physics Chemistry Biology Arts Music

is this consonant with the purpose of the curriculum? Such a line of inquiry may well deserve close consideration in the future.

To comment briefly in another direction (in light of Lee Cronbach's preceding presentation), it has never been quite clear to me what is meant by "construct validity." And it is still not clear to me. Facet theory may help clarify the matter, as illustrated by the above example on the structure of intelligence tests. First, a *definitional system* is specified for the universe of content and observations on it in the form of a mapping sentence. Second, *specifications* are made about the facets of the mapping sentence. (For the intelligence example, the specification is that one facet will act like a polarizer—each of its elements corresponding to a different direction in the empirical space of the variables—and the other facet will act like a modulator of distance from an origin.) The definitions and specifications lead to a *structural hypothesis* (such as that of a radex) which is tested by the empirical data. If by "construct validity" is meant a correct hypothesis concerning a correspondence between a system of definitions and specifications and between empirical structure on data, then this is exactly what facet theory is about. If something else is meant by "construct validity," it apparently still needs to be spelled out.

In another project, which we hope ultimately will be funded so that it can be expanded in scope, an attempt was made to analyze the possible purposes of the curriculum and also the possible variations in teaching techniques to attain those purposes. Mrs. Hava Tidhar, of the Instructional Television Trust (Israel), and myself were led to the mapping sentence of Figure 2 for designing empirical research on the effectiveness of teaching methods. The fourth facet there incorporates one just discussed in the previous example, although this was not noticed at the time. The last facet overlaps the task facet of the previous example: inference versus application, with some further differentiation. Clearly, different teaching methods may be required for the different purposes, and the assessment of the "benefit gained" for each of these purposes certainly requires a differentiating battery of tests. It would be most interesting to know the complete structure of intercorrelations of such an enlarged battery.

When one goes beyond tasks of inference and rule-application to the generation of interesting experiences, it becomes an important problem to relate this new class of variables to the others. Should there be a high or a low correlation between interest in an experience and the success in inference or in application of rules, considering also the

Figure 2

A Mapping Sentence for the Design of Observations on the Effectiveness of Teaching Methods

$$\text{The level of} \begin{bmatrix}\text{investment}\\ \text{benefit gained}\end{bmatrix} \text{on the part of} \begin{bmatrix}\text{himself}\\ \text{the teacher}\\ \text{the school}\\ \text{Education Ministry}\end{bmatrix} \text{of student (X) in} \begin{bmatrix}\text{implementing}\\ \text{training teachers for implementing}\end{bmatrix}$$

$$\text{a teaching method that} \begin{bmatrix}\text{emphasizes concretization} \begin{bmatrix}\text{within school course}\\ \text{outside school course}\\ \text{in similar subject}\end{bmatrix}\\ \text{presents occurrences under natural circumstances}\\ \text{introduces innovations in curriculum}\\ \text{increases learning speed}\end{bmatrix} \text{through using as an instrument} \begin{bmatrix}\text{the student himself}\\ \text{the teacher}\\ \text{T.V}\\ \text{radio}\\ \text{slides}\\ \text{slides + sound}\\ \text{film}\\ \text{tape recorder}\\ \text{charts}\\ \text{supplementary readings}\end{bmatrix}$$

$$\text{integrated with} \begin{bmatrix}\text{the student himself}\\ \text{the class}\\ \text{the teacher}\\ \text{T.V.}\\ \text{radio}\\ \text{slides}\\ \text{slides + sound}\\ \text{film}\\ \text{tape-recorder}\\ \text{charts}\\ \text{supplementary readings}\end{bmatrix} \text{in school subject} \begin{bmatrix}\cdot\\ \cdot\\ \cdot\\ \text{biology}\\ \text{English}\\ \text{mathematics}\\ \cdot\\ \cdot\\ \cdot\end{bmatrix} \text{in grade} \begin{bmatrix}\cdot\\ \cdot\\ \cdot\\ 6\\ 7\\ 8\\ 9\\ 10\\ \cdot\\ \cdot\\ \cdot\end{bmatrix} \text{for the purpose of}$$

$$\begin{bmatrix}\text{turning the learning of subject into an interesting experience}\\ \text{providing basic concepts}\\ \text{providing instruments for solution of new problems in the same area}\\ \text{knowledge of details}\end{bmatrix} \longrightarrow \begin{bmatrix}\text{high}\\ \cdot\\ \cdot\\ \cdot\\ \text{low}\end{bmatrix}$$

Hava Tidhar
Louis Guttman

varying degrees of similarity to what is taught in school? There is every reason to believe that there should be no uniform size of correlation in this regard. For example, methods that introduce innovations in the curriculum may generate more interest and succeed in this regard; how this relates to the other criteria is a matter concerning which not too much systematic information appears to be available as yet. Similarly, teaching methods which emphasize concretization should yield better results for some criteria and poorer results for other criteria.

Regardless of the size of correlations, making learning interesting may be a criterion in its own right, on the same level as success in rule-application or inference. Ultimately, one may have criteria beyond the school experience—such as success in vocations or in other activities—and certainly the more varied the predictors the richer the possibilities of obtaining higher multiple correlations. For the school situation, interests and achievement may be concurrent criteria which, in turn, may usefully serve as joint predictors of further criteria outside the school.

Further features of importance to the assessment process are indicated by the other facets of Figure 2. Without assessing "investment" in the curriculum, it may be difficult to arrive at policy decisions concerning teaching techniques to be maintained or introduced in the future. The "benefit gained" may depend on the amount of investment in the particular teaching method and on the particular purpose. Simultaneous study of both input and output seems to be required for useful assessment. It is indeed difficult to think of an adequate testing program which is not coordinated with the design of the curriculum and with consideration of different criteria, each of merit in its own right (no matter how it correlates with other criteria).

An example of a detailed design for a curriculum was developed some time ago in preparing achievement tests for the first and second grades in Israel. The curriculum booklet distributed to all teachers in these grades was studied, and we found two major facets that could be used for classifying the subject matter, as listed in Figure 3. One facet was the content being presented, and the other was the relation between aspects of the content. The children were to be taught, for example, the names of members of the family: father, mother, son, daughter, uncle, aunt, grandfather, grandmother, and so on. Similarly, they were to be taught names of foods or of items of clothing. For some topics, emphasis was on mere listing of the elements; for other topics emphasis was on interrelations among the elements; and for

Figure 3

Facets for School Curriculum in Israel
1st and 2nd grades (January 1943)

A. *Content**

1. family (1)
2. home and yard (1)
3. food (1, 2)
4. toys and tools (1)
5. clothing (2)
6. Sabbath (1)
7. festivals (1, 2)
8. the immediate environment (1)
9. at school (1)
10. at the store (2)
11. working people in town and country (1, 2)
12. public services (1, 2)
13. country and state (1)
14. the weather (1, 2)
15. agriculture (1)
14. plant life (1, 2)
17. animal life (1, 2)

B. *Relations* (identification of)

1. element in set
2. the set
3. definiens
4. definiendum
5. place
6. time
7. source
8. quantity
9. characteristic quality
10. cause
11. effect
12. manner of use

*In parentheses—grade at which the topic will be taken up

still other topics, other emphases were made such as on definition or attributes (place, time, source, and so on). By constructing a two-way table from these facets and entering therein topics mentioned in the curriculum, it was possible to show where great emphasis was laid and where hiatuses occurred. Sometimes the emptiness of a cell could be unintentional and sometimes intentional—that particular combination of facets not being considered important for the curriculum. In any event, this two-facet design facilitated construction of items to study school achievement cell by cell.

The structure of the interrelations amongst the achievements for the various cells is very complex. Since this work was done several years before our present computerized nonmetric techniques were available (3), we hope to be able to reexamine these data, especially if a new project can be mounted and implemented with respect to the present and future curricula.

Test construction itself can take advantage of a facet design of the curriculum by using alternative elements of facets to create distractors systematically. Design of "wrong answers" in a systematic fashion around the "right answers" enables a test to become diagnostic, with no need for increasing the length of the test. To the contrary, evidence seems to show that even shorter tests than are customary can have higher than customary reliability when a clean facet design is employed for distractor construction. Two detailed examples are in the research reports of references (4) and (6). Two more illustrations will be given here briefly from earlier work, one of a verbal rule-inferring test and one of a numerical rule-applying test.

A verbal intelligence test of only 10 ıtems was constructed facetwise and found to be highly reliable, and also effectively valid for a prediction purpose for which it was first used. An item of this test consists of three sentences which are instances of a relation. A sentence which is an instance of the same relation is to be chosen out of three alternative sentences. A relation requires at least two facets for its description. The correct alternative exemplifies the relation—i.e., incorporates the appropriate elements of these facets. One of the distractors does not satisfy the relation in that it deviates from it in respect to one facet, and the other deviates in respect to the other facet.

EXAMPLE:

The teacher examined the pupils.
All pupils listened to the new teacher.
The teacher did not permit the pupil to enter class.

a. The teacher decided to tell a pupil to leave the class.
b. The pupils left, running, and the principal said nothing.
c. All the parents encouraged the young teacher.

The relation exhibited in the first three sentences, as well as in alternative a, is:

Teacher *interacts* with *pupil.*

Alternative b is wrong in that there is no interaction, and alternative c is wrong in that the interaction does not take place with the pupil.

The first version of this test was given to applicants for the position of teacher-counselor. A substantial relationship was found to obtain between the scores on the test and the recommendations of the examining committee (which were made without knowledge of the test results).

The second example is from arithmetic in the second grade curriculum consisting of problems of subtraction. The types of problems to be included in the test were planned by first determining which factors would present special difficulties for problems, each such factor or combination of factors determining a type of problem.

The following list was arrived at (sums taught in the second grade never exceeding 100 and never leaving a negative remainder):

First row of subtraction sum:
a. number of digits (one or two)
b. units (zero or more)
c. tenths, if any (one or more)
Second row (subtrahend):
d. number of digits (one or two)
e. Units (zero or more)

REMAINDER:

f. units (zero or more or negative, requiring "loan" of tenths)
g. tenths (zero or more)

Of course, not all combinations of these factors are possible, the nature of the remainder being to a certain extent determined by that of the two rows (if units in both rows are zero, units in the remainder will also be zero). Some of the combinations uniquely determine one problem (10 = 0). Four problems were constructed for almost every one of the possible combinations—24 types of problems in all. These were administered in a multiple-choice test to 187 second graders.

Comparisons of types of problems with each other showed that all of the above factors might contribute to the difficulty of a problem. The effect of some would be so strong as to override that of others (for example, d and c would override f), while that of others would be relatively small (g). No interactions between effects of different factors were apparent.

The main object of the comparison was, however, to determine which two types of problems constituted distinct stages, such that mastery of one of these stages implied mastery of the other. For this purpose scattergrams were prepared showing the incidence of students obtaining scores of 1, 2, 3, or 4 correct answers on one type of problem with 1, 2, 3, or 4 correct answers on the other type.

A typical comparison is the following:

L \ S	0	1	2	3	4	*total*
4	3	4	14	17	93	131
3	4	2	4	7	18	35
2	—	1	3	1	—	5
1	5	1	2	—	—	8
0	6	1	—	1	0	8
Total	18	9	23	26	111	187

As is to be expected, a high correlation obtains between these two types of problems, 159 of the 187 cases being removed at the most one step from the diagonal, as indicated in the diagram. More important for our purpose, however, is the fact that all cases of larger deviations than these—with the exception of a single case—lie in the upper left corner. This implies that problems of type "L" are mastered before problems of type "S", there being only one case obtaining a low score on "L" and a high score on "S".

A look at these problems will serve to show that this finding is by no means an obvious one:

$$\text{Type "L":}\quad \begin{array}{r}60\\ \underline{-40}\end{array}\quad \begin{array}{r}90\\ \underline{-30}\end{array}\quad \begin{array}{r}40\\ \underline{-10}\end{array}\quad \begin{array}{r}80\\ \underline{-50}\end{array}$$

$$\text{Type "S":}\quad \begin{array}{r}31\\ \underline{-30}\end{array}\quad \begin{array}{r}75\\ \underline{-70}\end{array}\quad \begin{array}{r}43\\ \underline{-40}\end{array}\quad \begin{array}{r}98\\ \underline{-90}\end{array}$$

Type "S" has an advantage over Type "L" in that the remainder of tenths is zero (g above). On the other hand Type "L" has zero units in both rows of the subtraction sum (b and f above), this advantage being apparently more effective than that held by Type "S".

Scattergrams of other types showed deviations occurring in both the upper left and the lower right corners, indicating that they do not constitute subsequent stages in the above sense. An example is given by the following two types:

$$\text{Type "P":}\quad \begin{array}{r}98\\ \underline{-68}\end{array}\quad \begin{array}{r}75\\ \underline{-15}\end{array}\quad \begin{array}{r}86\\ \underline{-36}\end{array}\quad \begin{array}{r}53\\ \underline{-13}\end{array}$$

$$\text{Type "T":}\quad \begin{array}{r}56\\ \underline{-52}\end{array}\quad \begin{array}{r}89\\ \underline{-81}\end{array}\quad \begin{array}{r}29\\ \underline{-22}\end{array}\quad \begin{array}{r}78\\ \underline{-72}\end{array}$$

With Type "P" the remainder of the units was zero. Problems of this type were very slightly more difficult than those of Type "T" in which the remainder of the tenths was zero, but mastery of one type did not imply that of the other.

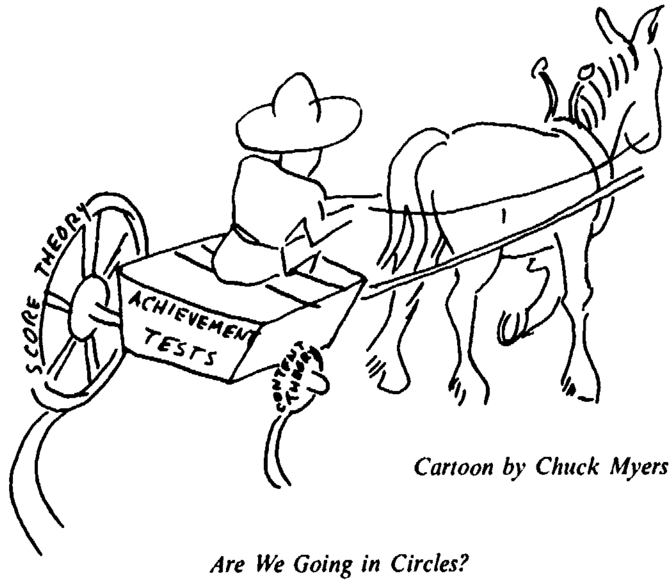
Occasionally two types correlated very highly, there being very few deviations in either the upper left or the lower right corner. This means, of course, that these types are indistinguishable from one another in terms of mastery by the students.

Having diagnostic distractors generally enables giving more than one score to the same test. The common type of "overall score" of number of right answers will, of course, still tend to be in order; it might be called the *concurrent* score on the items. But in addition to a concurrent scoring, there can be differential scores to indicate typical types of errors or deviations for pupils. Such diagnostic conditional scores can be useful for many purposes: improving teacher training, differentiation amongst kinds of pupils for whom different techniques may be appropriate, and individual guidance to the pupil. Tests of this kind could actually be given for classroom use by individual teachers, without need for norms based on larger groups. For an example of concurrent and differential (disjoint) scoring in another (attitudinal) context, see (8).

Having the content of the curriculum defined as clearly as possible by facets enables one to see what is not encompassed by the curriculum. If further criteria beyond the school are to be considered, a larger facet design should be established to incorporate both school and nonschool simultaneously. This will enable prediction and study of the larger correlation matrix involved. Knowledge of the larger structure will enable estimation of the potential and limitations of assessment of what goes on in the school system for predicting what will go on beyond the school.

In talking on a similar topic some time ago at Educational Testing Service—on the need for integrating test design with test analysis—one of the participants was inspired to make a sketch of the situation and it may be appropriate if we close with a look at his contribution, shown here as Figure 4. This also had some impact on at least one other colleague there (John Fremer) who has just informed me of his acronym F-A-C-E-T-S for: "*F*acets as *A*ssets in the *C*onstruction of *E*fficient *T*ests *S*ystematically." In the wider context of the entire educational process, too, clarity of definition of the universe of content can help us differentiate among, and facilitate achievement of, the several goals.

Figure 4



Cartoon by Chuck Myers

Are We Going in Circles?

REFERENCES

1. Guttman. L. A faceted definition of intelligence. In R. Eiferman (Ed.), *Studies in psychology, scripta hierosolymitana.* Vol. 14. Jerusalem, Israel: The Hebrew University, 1965.
2. Guttman, L. The structure of interrelations among intelligence tests. In *Proceedings of the 1964 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service. 1965.
3. Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika,* 1968. 33, 469-506.
4. Guttman. L., & Schlesinger, I. M. *Development of diagnostic analytical and mechanical ability tests through facet design and analysis.* Research Project No. OE-4-21-014. Jerusalem, Israel· Israel Institute of Applied Social Research. 1966.
5. Guttman. L., & Schlesinger, I. M. *The analysis of diagnostic effectiveness of a facet designed battery of achievement and analytical ability tests.* Research Project No. OE-5-21-006. Jerusalem. Israel: Israel Institute of Applied Social Research, 1967.
6. Guttman, L., & Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement,* 1967, 27, 569-580.
7. Guttman, R. Cross-population constancy in trait profiles and the study of the inheritance of human behavior variables. In J. N. Spuhler (Ed.), *Genetic diversity and human behavior.* Viking Fund Publications in Anthropology, No. 45, 1°67.
8. Jordan, J. E. *Attitudes toward education and physically disabled persons in eleven nations.* East Lansing, Michigan: Latin American Studies Center, Michigan State University, 1968.
9. Schlesinger, I. M., & Guttman, L. Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin,* 1969, 71, 95-100.

Knowledge vs. Ability in Achievement Testing

Robert L. Ebel
Michigan State University

To construct a good test of achievement one needs, first of all, a clear conception of the nature of that achievement. The thesis of this paper is that, in most school subjects, the essence of achievement is command of useful verbal knowledge. If this is true, the schools should direct their primary efforts toward increases in the pupil's cognitive competence, not toward his personal adjustment nor toward reconstruction of the society in which he lives; toward the cultivation of resources for effective behavior, not toward the direct shaping of the behavior itself: toward structures of useful knowledge on various important sutbjects, not toward the development of general mental abilities.

Of the four alternatives just mentioned (personal adjustment, social reconstruction, behavioral change, and developed abilities), only the fourth will be considered in detail at this time. Each of the other three merits similar careful and comprehensive treatment, but the limits of this paper will not allow it. Let it suffice here to outline one brief argument, and to quote one opinion in support of command of knowledge as the primary objective of education.

The essence of the argument is this: Only two means are available to the schools in their efforts to help human beings attain the desired ends of personal adjustment, social betterment, and behavioral effectiveness. One is to foster the cognitive development of their pupils. The other is to use the processes of conditioning to establish automatic, subrational responses. Not only does the first means seem to work a good deal better than the second where human beings are involved, it also shows a more decent respect for the right of every free man to make up his own mind.

The following quotation (9) is from the contemporary English educational philosopher, Richard Peters:

> To liken education to therapy. to conceive of it as imposing a pattern on another person or as fixing the environment so that he 'grows'. fails to do justice to the shared impersonality both of the content that is handed on and of the criteria by reference to which it is criticized and revised. The teacher is not a detached operator who is bringing about some kind of result in another person which is external to him. His task is to try to get others on the inside of a public form of life that he shares and considers to be important.

Those who agree that schools should seek above all to develop cognitive competence do not all agree on the nature of that competence. Some say, as I have, that it consists mainly in command of useful verbal knowledge. Others contend that the essence of cognitive competence is ability to think reflectively, critically, and straight, and that to achieve this ability the student's mental abilities of analysis and synthesis. of reason and judgment must be cultivated specifically and directly.

When I recently asked a class of 107 prospective and practicing teachers to choose one of five alternatives that came closest to expressing their view of the essence of educational achievement,

46 percent said that it was to cultivate the higher mental processes of reason, judgment, imagination, and creativity;
34 percent said that it was to learn how to learn;
10 percent said that it was development of a favorable self-concept;
6 percent said that it was to learn how to work effectively with others;
and only 3 percent said that it was to gain command of useful verbal knowledge.

But if this course goes as previous ones have gone, by the end of it when the students respond to the same question anonymously, about 75 percent of them will choose the alternative of useful verbal knowledge. I wish I could hope to be that persuasive with this audience.

The difference between the view that education is essentially concerned with knowledge and that it should concern itself primarily with the cultivation of mental abilities is, I believe, more than a semantic confusion. It does make a difference, it seems to me, whether the essential educational task is to help the pupil build a structure of knowledge or to develop some general mental abilities that will function effectively over a wide range of informational contents. Our choice of one or the other of these alternatives makes, I think, a real and an important difference in how we teach and how we write test items.

A third of a century ago the persuasive voices of two important educational leaders, Edward L. Thorndike (13) and Ben D. Wood (14), were raised in defense of the position I now espouse. I cannot presume to their eminence, but I can raise my voice, which I now propose to do. In this paper I will attempt three things: to outline the case for command of knowledge, to consider briefly objections that have been raised to emphasis on knowledge, and finally to examine the cultivation of mental abilities as an alternative to command of knowledge.*

The Case for Knowledge

Whatever a person experiences, directly or vicariously, and remembers, can become a part of his knowledge. It will become so if it is integrated into his own structure of knowledge. But this the learner must do himself. It cannot be done for him. In this connection, the distinction Scheffler has made between information and knowledge may be helpful. He says, ". . . it does not follow that the student will know these new facts simply because he has been informed; . . . knowing requires that the student earn the right to his assurance of the truth of the information in question. New *information,* in short, can be intelligibly conveyed by statements, new *knowledge* cannot" (9).

Some words of Kenneth E. Boulding (1) seem apposite at this point:

> The growth of knowledge even in the individual is not a simple cumulative process by which information is pumped into the head and remains in a reservoir. Knowledge is a structure, and its present form always limits its possibilities of growth. Hence we get the phenomenons of 'readiness' for certain kinds of knowledge at different stages of life; of wasted information input, which cannot latch onto anything in the existing knowledge structure; of false knowledge development, as the result of the acceptance of authoritarian pronouncements and the failure of feedback.

What we mean by a structure of knowledge in this discussion is not

*A previous but rather different discussion of this issue was published under a similar title, "Ability Versus Knowledge in Testing Educational Achievement," in the May 1969 issue of *The National Board Examiner,* National Board of Medical Examiners, Philadelphia, Pennsylvania.

what some others have meant by it in recent publications. It is more than a "body of concepts." Its purpose is not primarily to "limit the subject matter and control research about it" (7). Its function is not to define a discipline and distinguish it from other disciplines. It is not essentially a theoretical structure which identifies problems for research and specifies appropriate research methodology (11).

The structure of knowledge we are talking about consists of knowledge, only knowledge, and all of the relevant knowledge. Every factual detail and every generalization that can be related to other factual details and generalizations becomes part of the structure. The function of this structure, if it can be said to have a function, is to give the facts some degree of coherence and thus to make them meaningful and useful. What Darwin presented in *The Origin of Species* was such a structure. What the student finds in any good textbook of physics or economics or history is such a structure.

The structure we are talking about has some of the characteristics of a network, with the nouns and their modifiers in our factual statements corresponding to the knots, and the verbs and their modifiers corresponding to the strands between the knots. It also has some of the characteristics of our bodies, in which the individual cells are organized into tissues, the tissues into organs, organs into systems, and systems into the whole body. But this analogy is far from perfect. Whereas a man's body is usually a complete biological structure, no man's structure of knowledge is ever complete.

At the beginning of this paper I proposed that the essence of achievement is command of *useful verbal* knowledge. The words *useful* and *verbal* deserve a second look. Quite naturally anyone who plans a course seeks to develop in it only those concepts and ideas and skills that he thinks will be useful to the students who take the course. And students, when they are free to do so, choose courses whose outcomes they believe will be useful to them. Thus the natural emphasis in schooling is on useful knowledge, and it is hard to find anything that is being taught to anyone anywhere that can be convicted on the charge of being totally useless. On the other hand, almost every student can find in almost every course some information that he considers useless to him. If it is truly useless, it will never get built into his structure of knowledge, and he will forget it soon.

In assessing the potential utility of some item of information, one must be on guard against the inclination to ask too much. Few facts or ideas or concepts that are not learned automatically are likely to be useful to everyone, or essential for everyone to learn. Emphasis on programs of general education, and on common tests of achievement for all students, have misled some of us into believing that unless an item of information is likely to be useful to everyone it is not likely to be useful to anyone. Individuals differ even more in what they do learn than in how readily they learn. In a free, pluralistic, individualistic society it is good that this is so. The differences are valuable to us. In neither our teaching programs nor our testing programs should we act as if differences in what is learned are unimportant or undesirable.

Next, a word about the emphasis we have placed on the importance of verbal knowledge. Most of our thinking is done by considering unspoken verbal statements. Without verbal symbols our thought processes would be limited and slow. For this reason, and because it can be communicated and stored so easily, verbal knowledge is a very powerful form of knowledge. What distinguishes man from other primates and other mammals is primarily the facility with which he can produce and use verbal knowledge.

That the statements we issue and receive are not always laden with meaning, and that we sometimes use empty verbalisms as substitutes for knowledge must be recognized and guarded against. But no one in education practices rote learning of meaningless verbalisms as a means of learning, for it is next to impossible to build a structure of knowledge by that means. Thus the dangers of overemphasis on *verbal* knowledge in learning is easy to exaggerate. The inadequate structures of knowledge we so often observe are probably due not to too much learning of meaningless verbalisms, but to too little learning of any kind.

Objections to Emphasis on Knowledge

Consider now four objections that have been raised to emphasis on knowledge in the educational process. One is that knowledge alone cannot guarantee wisdom or goodness or happiness. That is true. But it is also true that the relevant knowledge a person commands can contribute substantially to the wisdom of his choices, the goodness of his behavior, the happiness of his life. What else that the schools are capable of doing can contribute more?

A second objection is based on the vast extent and rapid growth of the store of human knowledge. The task of getting command of any significant part of it, and of keeping up with new developments in even that part, seems hopeless. So teachers are told that what pupils should acquire is not knowledge but ability to learn, though this course of action would seem merely to postpone a solution to the problem of a hopelessly large amount to be learned. So too they are told that their aim should not be to increase the pupil's store of knowledge, but rather to increase his ability to cope with it, whatever that may mean.

Surely it is true that no one in this age can hope to achieve command of all knowledge. But in a society that is organized so as to make good use of specialized talents, no one needs to. Surely it is true that in some specialized areas knowledge sometimes grows with almost explosive speed. But this is not the knowledge you and I need to do our jobs well and to live good healthy, happy lives. That kind of knowledge seems to grow with distressing slowness. What troubles us is not an excess of knowledge but a deficiency. The fact that one cannot achieve command of all knowledge is a poor excuse for not trying to get command of some of it. And next time someone tells you that our store of knowledge is doubling every 10 years, ask how he knows. Where are the data? I strongly suspect that they do not exist, and that the kind of useful knowledge that ought to interest us is increasing more nearly at a rate of 10 percent than of 100 percent in 10 years.

A third objection asserts that knowledge is too ephemeral in validity, that truth changes too fast, to make it a sensible focus for our educational efforts. But is it really the facts that change? Or is it our interpretations of them? Despite Einstein, Newtonian mechanics is still essentially true as far as it went. When I taught high school biology we didn't know about DNA, but the blood still circulates through the same pathways we knew about then, and the crayfish looks the same now as it did then. New curricula differ from the old not so much because new knowledge has been discovered as because it has become fashionable to take a different approach to a different segment of that knowledge.

Finally, the obvious facts of forgetting are raised as an objection to emphasis on knowledge in schools. But an item of useful information that is well integrated into a structure of knowledge is not likely to be forgotten. Here again the distinction between information and knowledge is useful. Much of the information a person receives is quickly, and often quite appropriately, forgotten. But the new knowledge that a person succeeds in developing, on the basis of his old knowledge and his new information, tends to be much more permanent.

Alternatives to Knowledge

At least three alternatives have been proposed to command of knowledge as a focus of educational efforts: thinking ability, general mental abilities, and general educational development. The Educational Policies Commission (8) in one of its last major statements declared that the central purpose of American education is to develop in students the ability to think. College catalogs and college presidents often express the same idea. But no one ever thinks content-free thoughts. One always thinks about some kind of information. Further, he is always thinking, at least while he is conscious. Do students need to be taught to do something they do inevitably and automatically? Do they even need special skills to enable them to think reflectively, or critically, or straight?

Beyond question, knowledge of the principles of logical inference, deductive and inductive, of fallacies and semantic errors, of rational argument and proof does have some general utility. But what one needs to know to be able to think effectively in such different fields as statistical inference, or computer-assisted instruction, or military tactics tends to b largely specific to the field. Richard Crutchfield (2) has succeeded in developing in his students a "master thinking skill" using mini-mysteries and unsolved social problems as content. But will skills thus developed in one content area transfer to such other areas as the solution of word problems in algebra, the design of an experiment, or the diagnosis of a medical difficulty? On the basis of past experience and research one is entitled to be skeptical. Note too that even in such an endeavor one is dealing with *information* about a process, and seeking to build a structure of knowledge out of it.

Consider next the case for separate mental abilities or faculties. From Aristotle onward, and perhaps even before, efforts have been made to identify and define these faculties (5). These efforts led in the nineteenth century to the development of faculty psychology, to the doctrines of formal discipline, and to the imaginative hypotheses of phrenology. But early in the twentieth century negative research findings undermined these theories. Then the development of mental tests and of the techniques for the factor analysis of the test scores led to renewed efforts to discover the "essential traits of mental life" (4).

They have not been discovered, and what Aristotle himself suspected is gradually becoming clear to all: That they probably do not exist as a small number of separate, independent biological or psychological entities. In the words of Godfrey H. Thomson (12), the English factor analyst. "My own belief is that the mind is not divided up into 'unitary factors' but is a rich, comparatively undifferentiated complex of innumerable influences." Thus the various patterns of factors that various investigators have "discovered" are essentially clusters of somewhat similar, and hence related, traits. They do not exist to be discovered. but only to be defined by the tests involved and the methods of analysis used. They have the same essential characteristics as one of the groupings of organisms in a biological taxonomy, or one of the categories in a library classification system. Hence they cannot be developed as abstiact essences but only by enlarging and strengthening the elements in the cluster. And this can only be done, may we repeat, by building structures of knowledge.

The term "abilities" is used in many contexts with diverse meanings. To avoid misunderstanding we need to specify clearly what we have in mind when ·ve use the term. The kind of abilities we think do not exist are those which have been conceived as basic thought processes, few in number and elemental in function, developed by practice, essential to reflective thought and to the processing of information, but otherwise independent of knowledge. This kind of ability, we believe, has no demonstrable existence and properly belongs, therefore, only in the realms of mythology, ancient or modern.

Some other kinds of ability obviously do exist. There are specific cognitive abilities. like the ability to find a square root, or the ability to bid properly a hand at bridge. But such abilities are not independent of knowledge. On the contrary, they consist mainly of the command of knowledge relevant to a process. There are much more general cognitive abilities like mathematical ability or verbal ability. But these too consist mainly of more or less well defined clusters of more or less closely related specific abilities. They involve more extensive, less well integrated structures of knowledge.

Other abilities such as musical or surgical ability may be only partly cognitive. The other main ingredient in these is probably muscular skill. But in the case of all of these we are dealing with complex networks of elements and relationships. None of them can be properly considered as indivisible elements of behavior whose combinations and interactions produce all of the complex manifestations of behavior we observe.

Finally a word concerning the kind of general educational development (6) which is sometimes offered as a superior alternative to command of knowledge, and which has these two characteristics:

1. It consists mainly of a relatively small number of broad generalizations and involves little or no knowledge of specific facts.
2. It allows the same ends (i.e., generalizations) to be reached by using quite different content means.

I find it difficult to accept either the reality or the virtue of these characteristics. Valid generalizations are useful elements in any structure of knowledge, though they are often hard to come by and of limited scope, accuracy, and meaningfulness. But, it seems to me, they must be part of a structure. I do not trust a man's generalization if he cannot give a specific illustration of what he means by it, or cannot present specific evidence in support of it. I agree with Ben Wood's observation that:

> "... there is much more to be feared from the cult of generalizations, and the faith in principles as open sesames, than from the dangers of rote memory."

Nor can I agree that what is essentially learned in a good course in literature is something general like an understanding of the human predicament, not an understanding of the predicament of a particular man like Holden Caulfield; something general like appreciation of poetry, not something specific like the image of a sleigh stopped by the woods on a snovy evening; something general like ability to read and understand good literature, and not something specific like an unforgettable journey down the Grand Canyon with John Wesley Powell. Is it really conceivable that those who read *Jane Eyre* learn essentially the same things about human beings as those who read *Portnoy's Complaint*? Is not the range and variety of human experiei ces, human problems, and human personalities too wide to encompass in any limited set of generalizations? Do we not need, instead of a set of generalized outcomes of study, an intricate, complex, detailed, but always incomplete and far too limited structure of knowledge?

Now, what does all this mean for teaching and for the testing of educational achievement?

> It means that we should state most of our educational objectives in terms of achieved knowledge or specific abilities, not in terms of desired behavior, general abilities, or adjustment.
>
> It means that we should recognize ability to think—reflectively, critically, and straight—for what it really is in essence; not the exercise of some general ability, but the application of specific factual knowledge to specific problems to reach sound conclusions.
>
> It means that we should discontinue our search for "the essential traits of mental life," or for "the dimensions of achievement" (3), since they probably are not there to be found.
>
> It means that we can stop even lip service to that second, always troublesome dimension of our two-way-grid test blueprints—the one labeled "abilities" or "objectives"—and that we should concentrate more on the other dimension—the one labeled "subject matter" or "content"—to achieve an adequate sampling of the most useful knowledge.
>
> It means that we should give up the notion that problems of wide-scale testing of educational achievement, in the face of diverse curricula and content, can be solved adequately by using tests of general educational development which seek to emphasize general abilities and to be indifferent to mastery of specific content.

Professor Wendell Johnson of the University of Iowa used to tell his students that they ought to ask their teachers two questions incessantly: "What do you mean?" and "How do you know?" To these I would add a third that a student ought to ask *himself* incessantly: "Why is it so?" If he gets credible answers to these three questions, his structure of knowledge will grow, and he will be in command of it. He will understand what he knows. He will be able to use it to explain, to predict, and to solve problems. And in that field of knowledge his higher mental processes of reason and judgment, of analysis and synthesis, are likely to function effectively.

The cultivation of mental abilities is not an alternative to development of command of knowledge. If a mental ability can be developed, the best way to develop it is through command of knowledge relevant to the task. The soul of thought exists only in a body of knowledge.

REFERENCES

1. Boulding. Kenneth E. The uncertain future of knowledge and technology. *The Education Digest*, 1967. 33. 7-11.

2 Crutchfield, Richard S. Nurturing the cognitive sk:lls of productive thinking. In Louis J. Rubin (Ed.) *Life skills in school and society*. Yearbook. Association for Supervision and Curriculum Development. National Education Association, Washington. D. C . 1969.

3. Dyer. Henry S. Educational measurement—its nature and its problems. In Harry D. Berg (Ed.). *Evaluation in social studies.* 35th Yearbook of the National Council for the Social Studies. 1965.

4. Kelley, T. L. *Essential traits of mental life.* Cambridge, Mass.: Harvard University press, 1935.

5. Kolesnik. Walter B. *Mental discipline in modern education.* Madison. Wisconsin: The University of Wisconsin Press. 1958.

6. Linquist, E. F. Preliminary considerations in objective test construction. In E. F. Linquist (Ed.) *Educational measurement.* Washington, D. C.: American Council on Education. 1951. Pp. 127 134.

7. National Committee of the NEA Project on Instruction, Planning and Organizing for Teaching, Washington, D. C., 1963.

8. National Education Association. *The central purpose of American education.* Washington, D. C.. 1961.

9. Peters, Richard., *Education as initiation.* London: Evans Brothers, Ltd., 1963.

10. Scheffler, Israel. Philosophical models of teaching. *Harvard Educational Review*, 1965. 35, 131-143.

11. Schwab, Joseph . The structure of the disciplines: meanings and significances. In G. W. Ford and Lawrence Pugno (Eds.) *The structure of knowledge and the curriculum.* Chicago: Rand McNally, 1969.

12. Thomson, G. H. The factorial analysis of human abilities. *Human Factor,* 1935, 9, 180-185.

13. Thorndike. Edward L In defense of facts. *Journal of Adult Education,* 1935, 7, 381-388.

14. Wood, Ben D. and Beers. F. S. Knowledge versus thinking? *Teachers College Record,* 1936, 37, 487-499.

DISCUSSION

GOLDINE C. GLESER
College of Medicine
' uversity of Cincinnati

"To construct a good test of achievement one needs, first of all, a clear conception of the nature of test achievement." This statement, with which Professor Ebel began his presentation, is one with which the other speakers of this session and probably everyone here would agree. The speakers in the first session were also concerned with the nature of achievement, in particular emphasizing that in our society the conceptual model of competitive achievement affects both educational measurement and the educational process. Each of the present speakers in the context of his particular interest has contributed something to this point of view. Professor Guttman would provide both the school curriculum and the achievement-test items with a common definitional framework organized according to facets in order to assure relevance of testing to the educational process. Professor Cronbach speaks of the content domain whose specification forms the basis for test items. He would judge the educational relevance of the test as a description of achievement by having responsible persons compare the test items or tasks with educational objectives. Professor Ebel devotes his discussion to what he considers to be the nature of achievement in most school subjects.

Beyond this point the three papers of the present session diverge considerably in their scope and purpose. Professor Ebel has limited himself solely to a discussion of the complex question of what is the essence of achievement. Professor Guttman has been primarily concerned with the construction of tests that appropriately and systematically embody the curriculum and educational goals so that, in turn, the test results and their correlations with other ability tests can be used to determine the extent to which the goals are being achieved.

Professor Cronbach's oncern is with the validation of an interpretation or other use to which test scores may be put. From this standpoint he tends to assume the educational goals and the test itself as a given in the mathematical sense.

So much for the broad outlines of this provocative session. To get down to particulars, I should like to take another look at the suggestion made by Professor Ebel to the effect that the essence of achievement is "useful verbal knowledge." At first thought this is a simple and attractive thesis, perhaps too simple. In trying to analyze it, I find that my concern centered upon two main points.

First, I began to realize that the term "knowledge" itself is a construct almost as broad and difficult to translate into operational terms as the term "achievement." Professor Ebel points out that knowledge is more than facts or information; that it embodies an understanding of what we know and how we know it. It is structured information, organized in such a way that certain new information is readily absorbed; whereas other information, whether valid or invalid, is quickly forgotten. Does this imply, then, that only information which is remembered for a period of time is knowledge? Over what period of time must it be remembered? Must it be available for recall over this period or is it enough that it be recognized? Is it not necessary to designate the type of tasks which are appropriate to test knowledge as opposed to information? If so, and I have understood Professor Ebel's distinction between knowledge and information, then I cannot agree with his conclusion that we need concentrate only on content (putting aside those annoying questions of abilities) in order to achieve an adequate sampling of knowledge. Perhaps my difficulty here is in distinguishing what he terms specific abilities from broad general abilities. But surely if knowledge of a field implies that one can use information to explain, predict, solve problems, and reason in that field, then something more than content must be specified. Or, following Pro sor Cronbach's argument, we would need to engage in constr . validation for each achievement test to determine the appropriateness of designating the score a measure of the testee's knowledge in the field since this is an interpretation regarding the subject's internal processes.

My second concern centers upon the modifiers "useful" and "verbal." Ever since I had the good fortune to encounter a grown man of average intelligence who had a severe reading disability and discovered from him the extent to which our written tests discriminated against his obtaining jobs for which he was quite well qualified, I have become quite skeptical of our emphasis on verbal knowledge. I may be able to explain the principle of a gasoline engine; he knows the appearance and interconnections of the functioning parts in my automobile and can spot why it is not working properly. I could pass the examination; he can repair the car!

There is evidence that the overwhelming emphasis which has been placed on verbal achievement in our schools has been a contributing factor to children dropping out of high school or graduating with no feeling of self-worth or assurance of their ability to contribute to their own welfare or that of society. It may be reassuring to educators to equate verbal knowledge with achievement since this has been their mode of instruction, but surely there are other equally desirable and useful forms of knowledge. Is it possible our emphasis on verbal knowledge is a bit smug?

As for the word "useful," it indeed raises more questions than it provides answers. It is quite impossible to speak of "useful" unless one can specify "useful for what." There are many scientists who believe that what we "know" today may have little to do with tomorrow's world. What knowledge can we use to halt international aggression and insure world peace? Is our storehold of knowledge and past experience adequate to determine an economy that will give everyone a living wage? The young people today are doubting the relevance of much that is taught them, and even asking us "How do you know?" Worse yet, many of the facts taught in our youth are today considered myths. Which of the "facts" we now hold true will be the next to go?

I would like to turn now to Professor Cronbach's paper to make one or two comments. In my opinion, he has done an excellent job in clarifying the relationship of validity to the proposed interpretation of a test. However, his distinction between interpretations requiring only content validation and those for which construct validation is necessary is still not completely clear to me.

Any description of the capabilities of the subject, other than his actual score on the specific test, carries some implication. To say "James can recognize 80 percent of the words found in freshman textbooks" is not quite the same as to say "James recognized 80 percent of a representative sample of the words found in freshman textbooks in use in 1965." The implication in the first statement is that if all the words currently used in textbooks were presented to James, he would recognize 80 percent of them. Furthermore, the same test items would still fit the test specifications in the year 2000, but the inference would become less valid. Empirical evidence would certainly be of assistance in determining how good an estimate of the universe score is obtainable from the test. It might also indicate to what extent textbook vocabulary is changing over the years.

Professor Cronbach indicated at one point in his paper that content validity has to do with a set of operations, not with the responses. It is determined by the test designer, not the user. I should like to carry this further and suggest that it is not a parallel concept in any way to construct or predictive validity. It is not sufficient evidence for any type of interpretation. The term "validity" for this concept may be confusing, putting us in a semantic bind. Perhaps a better name would be "content relevance."

Session III

Theme:
Measuring the Performance of Systems and Programs

Systems Analysis of Education

THOMAS K. GLENNAN JR.
Office of Economic Opportunity

In the very small space allotted to me, one cannot hope to describe such an ill-defined body of techniques and procedures as systems analysis with any clarity. Hence. what I will try to do here is to create an impression of what I mean by systems analysis in education. I must begin my discussion with two disclaimers. First, systems analysis is not a well-defined set of techniques that can applied in routine fashion. Several books on systems analysis exist, but the best of them do not provide rules or formula but rather present their subject matter anecdotally and impressionistically. Second, systems analysis does not necessarily, or indeed usually, deal with an entire system although this is frequently held to be its distinguishing characteristic. Instead, it deals systematically with a component of a problem that either can be legitimately isolated from other components of the problem or which is within the purview of the decision maker for whom the analysis is being done.

Origins of systems analysis lie in the activities of British and United States scientists during World War II. At that time, many individuals from disparate backgrounds were thrown into the war effort and asked to help to solve problems facing military planners. In many instances, their solutions to these problems were original, imaginative, and unique—quite different from what the military analysts themselves would have come up with. In large part, this could be traced to patterns of thought that were quite different from the patterns that had been developed in the military themselves. Frequently, the rigor of scientific inquiry has been brought to bear upon operational problems that had previously been solved by relatively nonrigorous approaches. This work came to be called operations analysis or operations research and was generally concerned with ways in which to utilize existing weapons and weapons systems in a more effective manner. There are many terms that have been applied to this type of analysis since the Second World War. Operations research. systems engineering, management science. cost effectiveness analysis, and systems analysis are somewhat synonymous, although some analysts spend a great deal of time trying to distinguish between them. There are, I think, consistent differences between people who call themselves systems analysts and operations researchers, but as far as a decision maker is concerned, most of these activities should be viewed as relatively similar. If there is one distinguishing characteristic of systems analysis, it might be that it tends to treat a broader range of problems than other approaches do. Indeed, analysts at RAND have become wary of the term "systems analysis" and have tended in recent years to use the phrase "systematic analysis" to play down the notion that somehow they are dealing with total systems in all of their work.

In connection with these analytical activities, many methodological tools have been developed. These are sometimes referred to as a part of systems analysis, although they are definitely neither sufficient nor necessary tools. Activities such as cost-benefit or cost-effectiveness analysis, gaming, simulation, experiments, planning, program budgeting, and economic analysis are all used in systems analytic work but they should not be viewed as a necessary concomitant of systems analysis. Since a specific "definition" may still be of some use in a forum such as this, let me quote one given by E. S. Quade (1), a member of the RAND staff and editor and organizer of several books on systems analysis:

> In light of its origins and its present uses, systems analysis might be defined as that form of inquiry into problems of decision that aim to suggest a course of action by systematically examining the relationships between possible objectives of the action and alternative policies or strategies for achieving these objectives: comparing their costs, effectiveness and risks; and *formulating additional ones if those examined are found wanting*. Systems analysis represents an approach to, or way of looking at, complex problems of choice under uncertainty such as those associated with national security. In such problems, objectives are usually multiple, and possibly conflicting, and analysis designed to assist the decisionmaker must necessarily involve a large element of judgment.

Perhaps the best way to convey something of the flavor of systems analysis, as I understand it, would be to list attributes of systematic analyses.

1. *Comparison of real alternatives:* If there is one common element of systems analyses, it is the creation of alternative policy solutions and the development of moderately objective information that will support choices by a decision maker or planner among these alternatives. In the context of a school system, this may be a choice between alternative curricula, alternative uses of media, or perhaps more generally, alternative programs of instruction. It might also include an examination of alternative locations for new school buildings, alternative organizations of administrative support functions, or alternative organizations of educational functions among school buildings within a community.

2. *Examination of costs and benefits of alternatives:* It is traditional that these alternatives are examined in terms of both their costs and their benefits. If a school system is facing the problem of attempting to develop or to improve the quality of its reading instruction in the early grades, it is able to use a great variety of approaches to achieve such ends. Some of these might involve the use of additional media or such gadgetry as computers. Others might simply involve the adoption of a new text or a different style of alphabet or many other things most of you at this conference are far more capable of naming than I. Each of these potentially has a series of effects. Reading is improved along a number of dimensions, probably somewhat differentially by each of the programs. Programs may be better for certain types of youths than for others. What is important is that they presumably do have different degrees of effectiveness.

They also have different costs. Some of the programs are likely to require new equipment or new textbooks. Some will require more retraining of teachers than others. Some may require more experienced or better trained teachers than others. Therefore, there is associated with each program a group of costs, most of which can be measured in monetary terms and particularly in increments of cost over current programs. There may also be intangible costs such as *difficulties in organizing*, managing, or scheduling such a program. Systems analyses will compare the benefits and costs associated with each of these programs. The economist has developed a fairly well-proven framework of analysis and can give guidance as to how to treat costs and to a lesser extent, benefits or effects.*

3. *Interaction with policymaker:* I would argue that the first two attributes must be present in any study that purports to be a systems analysis. This attribute—interaction with policymaker—seems to me to be a quality associated with good systems analysis. The alternatives to be compared must be real alternatives. They must be alternatives that are at least potentially feasible and which can possibly be adopted by the educational system. Moreover, the results of any of these analyses are going to be based upon many simplifications, as in any scientific inquiry, and the credibility and utility of these results to a policymaker will depend upon his understanding of the nature of the assumptions and the implications of these assumptions. Thus, the interaction between systems analysts and policy makers is mutually beneficial. What the analyst knows about the policy maker's problem will help him define what we will call "feasible alternatives" for comparison. They may not always be alternatives that the decision maker finds acceptable, but they are ones which, in light of the analyst's knowledge of the school system, are feasible. At the same time, the user of systems analysis—the policymaker—should gain an understanding of what has gone on in the analysis by this interaction and, hence, will be in a better position to assess the reasonableness and reliability of that analysis.

4. *The process is an iterative one:* When one reads a good ·vstems analysis, it typically appears simp!., straightforward, and reasonable. The reaction is frequently "Why hasn't that been done before?" These appearances are deceiving. Typically, the analysis begins with an attempt to define or determine objectives, an attempt which was not often very satisfactory. Then alternatives for meeting tentative objectives are developed. As these alternatives are analyzed and examined,

* It should be noted that an investigation may well be required to determine differential effects and indeed in some instances, it may be that the relative differences in effectiveness are so small between programs that for all practical purposes there are no differences. In this instance, it may be that the choice of a particular reading program will be based solely upon its costs. In my judgment, the capacity to apply systems analysis to education, particularly to instructional problems in education, is and will continue to be, heavily dependent upon the capacity of evaluative techniques to determine the level of effectiveness of different programs and that systems analysis in education is thus heavily tied to, and interact ve with, 'he process of evaluative research.

some of the objectives are modified or changed in light of a clearer perception of the problem or a recognition that certain of the objectives are either trivial or absolutely impossible to obtain. Alternatives are then redefined and the process continues.

I think that if you had the opportunity to participate in systems analysis as it is carried out in an organization such as the RAND Corporation, you would find that it is far from the orderly process characterized even in the methodological books put out by that corporation. In passing, I might note that it is commonly felt at RAND that a very substantial proportion—perhaps more than half of the effort involved in systems analysis—is really directed at the problem of defining the problem.

5. *Interdisciplinary research teams:* There is nothing inherent in the conduct of a systems analysis that requires that it be carried out by an interdisciplinary team. Presumably, individuals of many different disciplines can pick up sufficient knowledge about the problem at hand to carry out something that can be called a systems analysis. In my judgment, however, most of the really path-breaking and significant systems analyses have been carried out by interdisciplinary teams. The reason for this, I think, is not that the different disciplines bring particular kinds of knowledge to bear upon the problem, although this is probably a positive benefit. Instead, I think the real benefit is that they bring a differing style of problem solving to bear on the problem. They bring different value structure to the analysis. They bring different analytical techniques. Interdisciplinary research is incredibly difficult to initiate and sustain. and I think that there is a strong tendency for the utility of the interdisciplinary team to decrease with the length of time that they have worked together.

6. *How large a system?* One of the arts of carrying out a systems analysis is to limit the problem. Suppose the problem is to improve the teaching of reading to disadvantaged kids in urban slums. It is possible to consider the problem of the best way to teach reading to first, second, and third graders in individual classrooms. It is also possible to consider it in the context of the entire school where new options will be open to allow the youngster to go to special remedial classes. Or you can deal with it in the context of entire school systems where special schools or special programs can be set up. You can expand the reading program outside the school system to utilize community resources. You

can work indirectly on a youngster's reading through his family. You can decide that the definition of the reading problem as one that is particularly important in the first. second. and third grades is inappropriate. and that emphasis should be placed upon early childhood development.

The definition of any system is arbitrary. The one chosen for a systems analysis should be the one which, in light of our knowledge, appears most useful to that analysis and the policy-making tools available. I would suggest that the primary criteria for the choice of such a system is that the effect of influences outside of the defined system upon the effectiveness of the programs that are being compared is small. Other criteria for the definition of system can be used. One of them may be that the system should include all of those elements which can be effectively manipulated by the administrative unit for which the analysis is being done. In the case of a school, this would include all of the in-school elements. This definition of a system should be tested, however. It may be that over a period of time, the analysis would show that there is a need to redefine the system and ultimately to redefine the functions of the administrative unit.

7. *Outside vs. inside analysis:* The best systems analyses, I think, are done by organizations that are outside the immediate operating unit. I should, I think, admit to the strong possibility of bias in view of my previous association with such an organization. RAND never, I must say, was very successful in doing systematic analyses of its own operations. I think that the objectivity and the freedom from constraints imposed by mores and traditions of a particular administrative unit are very important. I must reemphasize, however, the necessity of there being a close interaction between whoever is conducting the analysis and that administrative unit.

8. *Common sense:* The most important quality of any systems analysis is common sense. A concomitant, I suspect, of common sense in good systems analysis is simplicity.

So much for a description of attributes of good systems analyses. These are derived not so much from observations of systems analyses in the field of education as from observations of the use of systems analysis in fields other than education, both domestic and military. In my judgment, the [illegible] of systems analyses for education is yet to be demonstrated, although this point of view might be disputed by many here, and certainly by many of the toilers in the vineyards. The educational system in the United States poses many significant problems in the conduct of systems analysis. In particular, the highly decentralized, generally small administrative units of the system make it difficult to implement systems analyses on any broad scale. The policy levers controlling education are so widely distributed among the local, county, and state jurisdictions and, to a lesser extent, the federal level that the admonition that one should work closely with the policy maker becomes an extraordinarily difficult one to follow.

Another major, and perhaps even more fundamental, problem is the absence in our educational system of good, reliable, agreed-upon and quantifiable measures of educational output. This inability to adequately characterize educational outcomes has meant that such analyses as have been done have often concentrated solely on cost, choosing a minimum cost system that appears to reach some narrow set of objectives. Obviously, the problem of defining educational objectives and measurement of educational outcomes is of considerable concern to this conference, and I will say no more.

In a purely political sense, the multiple dimensionality of the output and, to a lesser extent, the costs poses considerable problems. As Quade's definition cited earlier notes, multiple dimensionality of outcomes is a normal attribute of the problems treated by systems analysis. Nonetheless, there is a tendency to reduce those outcomes to a relatively few measures and, for the kinds of problems treated in the past, this has been considered reasonable. In the case of education, such narrowing of measures is likely to be less defensible. Consequently, systems analyses, if they are honest and relevant, are going to fail to come up with any very clear-cut answers. This will frequently be a disappointment to both the analysts themselves and to the policy maker who commissions the analysis, for it does not give him the kind of clear-cut guidance that he had hoped for from this much-vaunted technique.

Finally, the lack of good and accepted theories of learning will mean that the whole style of systems analysis will have to change. Far more emphasis upon experimentation and upon evaluation will have to accompany systems analyses than has been the case in military analyses. Military analyses have been able to utilize physical laws to predict the performance and effectiveness of particular weapons designs. Such laws do not exist for education and as a consequence, it is impossible to make an a priori prediction with any confidence that a particular sequencing of activities or learning tasks carried out by a particular class of teachers with a particular kind of youngster will lead to a specific result. This kind of information must be determined empirically, a task discussed by most of the participants in this conference.

In closing, then, I would like to leave you with several thoughts. First, systems analysis is not a set of well-developed techniques which provides a quick and easy answer to the problems of choice among educational programs. Second, systems analysis must change very substantially, ooth in terms of the talents of the members of the analysis teams and the style of work which makes up the analysis. Nonetheless, it is my judgment that the style of analysis that has been developed is applicable to educational problems. If measurement techniques that adequately characterize educational outcomes can be developed, the application of this analysis should have great benefits for the nation's schools.

REFERENCES

1. In Quade, Edward S. (Ed.), *Analysis for military decisions.* Chicago: Rand McNally, 1964. P. 4.

Dr. Edward Suchman, whose paper "The Role of Evaluative Research" appears on pages 93-103, died shortly before the Invitational Conference was held. Dr. Louis Guttman, a colleague and close friend of Dr. Suchman for many years, paid the following tribute.

When I was invited to present a paper at this conference and received a copy of the program, I was very glad to see that Ed Suchman would also be appearing. I hadn't seen him for several years and I thought this would be a good opportunity for us to talk about old times and new times, but not a time when I would be asked to say a few words about Ed Suchman.

It is not an easy thing for me to do. I just learned about his passing away last week and I haven't had a chance to adjust to it.

Years ago Ed and I worked very closely together. The background, of course, is very simple. He got his M.A. at Cornell, his Ph.D. at Columbia. In between he was involved in radio research with his mentor Paul Lazarsfeld, and then came to the Pentagon during World War II to work with Sam Stouffer and his Research Branch on the morale problems of the American army. After the war, he joined the Cornell faculty, then became director of social science ativities for the New York Department of Health. He was associated with the injury control program of the U.S. Public Health Service, and was professor of sociology at the University of Pittsburgh when the untimely end came. His last book was on evaluative research, which is very relevant to our program today. I was closest to the work he did for Volume IV of the American Soldier series: *Measurement and Prediction,* on which we collaborated.

Those of us in the Research Branch of the American army formed a rather close group in those days, and I believe we all still feel a very warm bond—we feel this very deeply.

We will remember Ed as a person who was always lively and inventive. I still remember those nights when we were about to leave the

Pentagon—an enormous structure with no windows for the inner circle we occupied. It being a long walk to go out to one's car, he would telephone the Weather Bureau to ask if it was raining outside.

But he was also the kind of guy who would work day and night in the Stouffer tradition. He had boundless energy for seeing data through the mill. This is a tradition which I do not think has been maintained at the same pace; now we have computers and somehow the atmosphere seems more leisurely.

When the time comes for a full evaluation, Ed will be remembered and recognized as one of the productive, energetic, stimulating, practical, and most likable contributors to social science and to our lives.

The Role ot Evaluative Research*

EDWARD A. SUCHMAN
University of Pittsburgh

The Current Demand for Evaluative Research

In times of rapid social change, traditional public service and social action programs are apt to find themselves under constant challenge from new and untested approaches. Which of the old should be revised or discarded completely and which of the new deserves a trial? How much change is necessary? Can the old be patched up and made to work or must wholly new policies and practices (16) be developed? The climate of such times is likely to be one of vigorous academic debate and public conflict concerning what steps need to be taken to meet the even increasing cries of dissatisfaction and dismay (5). The presence of such dissatisfaction in almost all areas of public service today is perhaps the key to the current demand for evaluation. Evaluation feeds on dissatisfaction and change. Action, almost any kind of action, is frantically sought as a means of alleviating the discontent or, at least, of postponing open conflict. And, almost as an apology for too precipitous action, evaluation is often proposed as a means of maintaining rationality and control (7).

The field of education today, like that of health and social welfare, is under pressure to change its traditional programs and organization. Industrialization, urbanization, civil rights and minority movements, changes in educational technology, new occupational demands have created a strong public and professional search for new educational approaches. In many cases, trial and error, rather than carefully planned change, has characterized these educational innovations. For

* Dr. Suchman's paper was presented by Edmund Ricci of the University of Pittsburgh.

the most part, these new programs have been developed without any appreciable relevant theoretical basis (2). The less obvious the theoretical justification for a program, the greater will the need be felt to evaluate its success or failure. In fact, a major rationalization for a trial-and-error approach is that one can evaluate the trials and determine the errors. If the program can be made to work through "rank empiricism," the theoretical reason (if any!) for its success can be determined "after the fact" (1). Thus, too often, the situation becomes one of introducing innovation for innovation's sake alone, and perhaps it is exactly the highly trial-and-error nature of so many of these educational innovations that has led to the current demand for more intensive evaluation.

In short, the need for evaluation in education today is so great because we have lost faith in our traditional programs and are uncertain as to what we are trying to do with our new programs or why they should work at all. What the field of education (and, we might also add, health and welfare) lacks most today is an understanding of how to evaluate the relevance of its current fund of knowledge for social needs, and then utilize this knowledge in the development of new programs more suitable to these needs. This inadequacy stems, to a large extent, from the underdevelopment of what we might call "educational practice theory." Teaching as a profession has developed over the years without sufficient attention to the underlying rationale for its many instructional practices. Based largely on untested assumptions, these practices have been handed down from one generation of students in teachers' colleges to another (8). While supposedly based on the psychological theories of learning, only recently have we become aware that a wholly different type of "instructional" theory is necessary for the practice of teaching as opposed to the psychology of learning (9).

The relevance of this distinction between basic theory and professional practice theory for evaluative research has recently been underscored by the report of the Special Commission on the Social Sciences to the National Science Board entitled *Knowledge Into Action: Improving the Nation's Use of the Social Sciences* (14). According to this report. "The professions are among the main social institutions through which social science knowledge can be translated into day-to-day practice" (15). Evaluative research, it is pointed out, provides one of the major avenues whereby the various fields of professional practice can clarify, develop, and test the underlying rationale for and theory of their professional practice. Program evaluation is the *sine qua non* of professional efforts to translate knowledge into action.

A Conceptual Approach to Evaluation

At the most general level. evaluation refers to the process of determining worth or value. This process may range from a subjective assessment based on personal experience to a highly controlled experiment based on the scientific method (11). The object of evaluation may be as broad as national policy or as narrow as some individual need. In most cases, however, the purpose of such evaluation is to direct future action. Thus, we can locate evaluation in the area of decision making whether such a decision involves some personal act or some major planned social change (13).

Given such a wide range of application, it is not difficult to understand why so much confusion and disagreement exists today as to what constitutes the "proper" method of evaluation. Different purposes will require different evaluation designs. At the moment we lack any agreed-upon systematic scheme for classifying or ordering different forms and types of evaluations. Furthermore, it is our opinion that the field of evaluative research is changing much too rapidly to make any high degree of formalization desirable at the present time. It is our hope, however, that the following remarks will at least indicate some of the major theoretical and methodological problems in this area.

First, it seems unwise to conceive of evaluation studies as requiring some special theoretical or methodological model of their own. The basic differences between evaluative research and nonevaluative research are the same as those between any form of applied research and basic research. These differences have received detailed discussion in a number of articles and need not be repeated here (10). Nevertheless, a great deal of the current controversy over both the "theory" and "method" of evaluative research tends to lose sight of the fact that the major purpose of most evaluative research is administrative. The primary goal is usually to aid the decision-making process concerning some social problem or policy. Even in the conduct of evaluation studies, administrative considerations will often have precedence over scientific ones (12). Therefore, it does not seem productive to continue arguments over the degree to which the classical experimental design is applicable to evaluative research (3). As is so often the case in such controversies, the answer would be "It depends on the purpose" (4). In the early stages of evaluating a new program, it would probably be more profitable to utilize a rather fluid, clinical case study, "anthropological" type of design. Increased understanding of the problem and more detailed specification of the type of activity to be carried out could then be followed by a survey evaluation design which would provide preliminary evidence as to the effectiveness of one's program on an *ex post facto* or longitudinal basis. Finally, when the stage is reached for a definitive test of some particular program, it would then be possible to proceed to a more rigorous experimental design (18).

One cannot really argue in the abstract as to which approach is "correct." We have found it useful to distinguish between "pilot" projects where the main objective is to try out different approaches and where a flexible, anthropological approach provides the greatest amount of information and "model" projects where the emphasis is upon testing a program under ideal conditions and where a more rigorous experimental design is indicated. "Prototype" projects call for an operations research design whose main emphasis is upon the feedback of information for program improvement (19).

These kinds of practical considerations do not, of course, rule out the need to view evaluation research within a broader theoretical or methodological perspective. In our work, we have found it most helpful theoretically to link the evaluative research model to that of the independent-intervening-dependent variable sequence of multi-causal analysis. In terms of this model, the program activity becomes the independent variable which is to be manipulated or changed. The intervening process represents the "causal" factors which promote or inhibit the development of the valued goal. The dependent variable then becomes the desired objectives or changed conditions. Given this formulation, the nonevaluative hypothesis "If A, then B" becomes the evaluative hypothesis "By changing A (through a planned intervention program), the 'causative' process B will be effected in such a way that the probability of producing effect C (which I judge to be desirable) will be increased." Underlying this hypothesis are several key theoretical issues: 1) What are the "causative" factors affecting the achievement of the desirable outcomes? What changes must one produce in the underlying process in order to bring about a change in the result? 2) What activities, programs, or techniques can one develop for deliberate intervention into this "causative" process so as to increase the probability of the desired outcome? What are the social, economic, political constraints limiting one's ability to utilize a particular intervention? and 3) What is the total picture of changes produced by the intervention? What negative (unintended) as well as positive (intended and unintended) effects take place? What other aspects of the social system (20), besides the one within which one is working, are affected?

In our opinion, these three sets of questions represent the key "theoretical" issues underlying evaluative research. They point very decisively to the need for the development of theoretical models which will link professional activities to desired social outcomes. Basically, this is what we mean by professional practice theory. As the report to the National Science Board has documented, this area has been thoroughly neglected in most professional school training programs. Specific activities as well as entire programs are carried out largely on the basis of tradition without sufficient attention to the rationale for believing that they are capable of producing the desired results. On the other hand, "academic" disciplines continue to develop basic theories which fail to tie into professional practice. Finally, national policy or goals are set by social and political forces without due consideration of the professional consequences as determined by professional groups. The model we propose inherently requires a close linkage between professional practice and academic theory.

This formulation of the evaluation problem also permits one to distinguish between what we might call a "technical" versus a "theoretical" failure. Professional service or social action programs may fail to achieve their desired goals either because they are operating according to an invalid underlying theory of process (a theoretical failure) or because, even though the rationale is correct, they cannot succeed technically in developing and implementing a program that successfully engages the underlying change process. Confusion of technical with theoretical failures underlies much of the current controversy regarding programs in almost all fields of health, education, and social welfare. For example, does the apparent failure of programs such as Head Start stem from an invalid theory concerning the effect of early environmental deprivation upon intelligence or can it be more correctly explained in terms of a failure to institute the proper type of early intervention programs?

The preceding formulation of the evaluation problem, incidentally, also indicates the crucial contribution which evaluative research can make to underlying theoretical understanding. Professional practice and social action can serve as a highly significant crucible for the testing of academic theory.

Levels of Evaluation

Evaluative research may further be viewed as taking place on a number of different levels, each requiring a somewhat different conceptualization of the evaluation problem. We may order these levels from the broadest to the most narrow as dealing with (a) social systems; (b) organizations or institutions; (c) programs or projects. The major emphasis of the social-system level of evaluation will be upon policy or ultimate objectives, while an organizational evaluation is more likely to be concerned with intermediate administrative objectives and a program evaluation with immediate service objectives. On the social-system level, the objectives most likely would deal with ideological or value questions and be aimed at policy decisions of a social-political nature. Evaluation at this level challenges the goals and assumptions of the major societal subsystems in terms of their ability to affect such fundamental social values as the public's health, education, and welfare. Evaluative research strategy at this level tends to be more descriptive and philosophical than empirical and scientific, involving "great debates" over national goals and the means towards these goals. Objectives and criteria are usually formulated in terms of gross social indicators of "progress" towards overcoming disease, ignorance, and poverty. Social "bookkeeping" provides the basic evaluative data on effort and accomplishment, with political judgments being made concerning the adequacy and efficiency of comparative systems. Examples of evaluation studies at this broad level would include such comparisons as the "socialized" versus "free enterprise" medical care system, the educational approaches of the progressive versus the traditionalist schools, a welfare system based on need or entitlement.

At the next lower level, organizational evaluations have as their objective the evaluation of the structure and operation of the major institutional arrangements whereby the broad social goals of a society are to be pursued. On this intermediate level of administrative evaluation, we are most concerned with a systems approach for evaluating the overall operation of major organizations such as the Department of Health, Education, and Welfare. This type of evaluation would stress the objective of providing "accountability" information for organizational management. This level of evaluation is becoming increasingly important as massive social programs are mounted involving major federal, state, and local public service agencies. Evaluation at this level has as its major objective staff policy formation and change involving the allocation of resources and the assignment of priorities; it is apt to have the greatest impact upon planned social change. Long-range objectives are more likely to be affected by organizational or system changes than by specific program innovations. In fact, one of the major criticisms of much current evaluation aimed at the program level is that this form of "tinkering" is too specific and symptom-oriented to really ameliorate any major social problems.

Finally, on the level of immediate objectives, we have what is most common in evaluative research—program evaluation. Evaluation studies at this level are apt to concentrate on the "effort" category. Specific programs become evaluated largely in terms of input, or the amount of effort expended rather than the actual accomplishments of the programs. The objectives at this level stress such criteria as the quality and quantity of personnel, performance ratings, and amount of service rendered. It is usually taken for granted that such activities, if successfully mounted, will *ipso facto* produce the desired outcomes. Such operational programs are usually fairly easily evaluated by means of an experimental or quasi-experimental design. In fact, this is probably the most appropriate level of evaluation for the use of such designs. It is extremely important, however, to remember that an evaluation at this level should not formulate objectives which involve more than a measure of the accomplishment of the program being evaluated. Such single programs are not likely to show any appreciable impact upon intermediate or ultimate objectives. The inclusion of such higher-level objectives in the evaluation of service programs is probably the main reason why so many evaluations appear to indicate failure. It is a very rare service program indeed that can show any impact beyond the immediate level of its specific operational goal.

Evaluation and Measurement

Since the major focus of this conference is upon measurement, perhaps a brief word is in order concerning the relationship of measurement to evaluation. Our approach to evaluation stresses the testing of some hypothesis concerning the relationship between planned activities and desired objectives. This relationship has been referred to in terms of a broader input-process-output model. Stated in this form, it becomes apparent that measurement is an inherent aspect of evaluation research. Most critically, perhaps, such measurement refers to the conceptualization, isolation, and measurement of specific criteria representative of the relative degree of success or failure in the attainment of the desired output. In general, this has been referred to as the problem of criterion measurement and has a long and interesting history. Some would maintain that until one has developed reliable and valid criterion measures for the desired objectives, evaluation cannot take place in any rigorous, scientific way. The time-honored question "Validity for what?" presents a critical problem for evaluation research. All programs will have some effect and unless such effects are specified in advance and clearly related to the desired values or goals, evaluation research stands in danger of simply showing that some kinds of effects took place. Similarly, it would seem important that the measurement of effect embrace the magnitude and duration of the effect both in terms of the costs involved and the degree of effect necessary to make a program worth undertaking.

The problem of criterion measurement, however, is much more complicated than would appear from a review of the technical aspects of criterion validity. As formulated in this paper, the major objective of an evaluation study is to aid in the decision-making process. The rightness or wrongness of such decisions are only slightly amenable to evaluation by current educational tests and measurement (6). How, for example, does one measure the effectiveness of a social policy concerning welfare? The development of so-called social indicators which reflect the well-being of a population go far beyond the standard techniques of achievement tests and measures (21). Social goals and objectives are not easily translatable into specific criterion measures. It is much more likely that such evaluations will relate more to progress toward an ever-changing goal than to achievement of a specific goal.

This point is also relevant to the other major components of our input-process-output model. It seems to us that for many types of evaluation, a much more meaningful and realistic form of measurement would concern itself with the developmental process itself rather than with the final output. Thus, we face a wholly new set of measurement problems in terms of process. An important element in the definition of such process criteria is an understanding of the underlying sequence of events leading toward the desired ultimate objective.

Given such a sequence, it is possible to then determine certain crucial points along a time line which could be used as evaluative measures of progress (17).

Finally, we come to the measurement of the "input" component of the evaluation process. Measurement in this case becomes essentially one of defining, and varying, the essential program components that constitute the activities of the intervention or action program. The crucial question is "How does one measure the degree to which one has successfully introduced the type of program that is being hypothesized as the desired treatment?" As we have stated previously, unless one has a reliable and valid measure of input, it is impossible to determine whether any future failure was due to the inadequacy of the underlying theory or to the technical inability to put into effect the desired program. In its most desirable form, it would be possible to vary the input and to relate such variation to differing degrees of process involvement and output production. Thus, the problem of measurement of program components becomes an extremely important one for evaluative research.

In summary, measurement may or may not involve evaluation; however, all evaluation implies some form of "measurement" of all three components of the input-process-output model.

REFERENCES

1. Bloom, Benjamin. Some theoretical issues relating to educational evaluation. In Ralph W. Tyler (Ed.), *Educational evaluation: new roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 26-50.

2. Brickell, Henry M. Appraising the effects of innovations in local schools. In Ralph W. Tyler (Ed.), *Educational evaluation: new roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 284-304.

3. Campbell, Donald T. and Stanley, Julian C. Experimental and quasi-experimental designs for research on teaching. In N. C. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 171-246.

4. Chapin, F. Stuart. *Experimental designs in sociological research.* New York: Harper and Bros., 1947 (revised 1955).

5. Conant, James B. *Shaping educatic tal policy.* New York: McGraw-Hill, 1964.

6. Flanagan, John C. The uses of educational evaluation in the development of programs, courses, instructional materials and equipment, instructional and leaving procedures, and administrative arrangements. In Ralph W. Tyler (Ed.), *Educational evaluation: new roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 221-241.

7. Freeman, Howard and Sherwood, Clarence C. Research in large-scale intervention programs. *Journal of Social Issues,* January, 1965, 11-28.

8. Glaser, Robert. Ten untenable assumptions of college instruction. *The Educational Record,* spring, 1968, 154-159.

9. Glaser, Robert. Theory of evaluation of instruction: changes and trends. From the *Proceedings of the Symposium on Problems in the Evaluation of Instruction.* Los Angeles: University of California, December, 1967.

10 Hemphill, John K. The relationship between research and evaluation studies. In Ralph W. Tyler (Ed.), *Educational evaluation: new roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 189–220.

11. Herzog, Elizabeth. *Some guidelines for evaluative research.* Washington, D. C.: U. S. Department of Health, Education, and Welfare, Social Security Adm[illegible]ation, Children's Bureau, 1959.

12. Hesseling, P. S. [illegible]gy *of evaluation research.* Netherlands: Van Gorcum and Co., 1966.

13. Hyman, Herbert H., Wright, Charles R. and Hopkins, Terence K. *Applications of methods of evaluation: four studies of the encampment for citizenship.* Berkeley: University of California Press, 1962.

14. *Knowledge into action: improving the nation's use of the social sciences.* Washington, D. C.: National Science Foundation, 1969.

15. *Ibid.* P. 21.

16. Miles, Matthew B. *Innovation in education.* New York: Bureau of Publications, Teachers College, Columbia University, 1964.

17. Provus, Malcolm. Evaluation of ongoing programs in the public school system. In Ralph W. Tyler (Ed.), *Educational evaluation: new roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 242–283.

18. Suchman, Edward A. *Evaluative research: principles and practice in public service and social action programs.* New York: Russell Sage Foundation, 1967.

19. Suchman, Edward A. Action for what? A critique of evaluative research. Paper prepared for symposium on *The Organization, management, and tactics of research,* Vocational Guidance and Rehabilitation Services, February 20-21, 1969. (In press).

20. Suchman, Edward A. A model for research and evaluation on rehabilitation. In Marvin Sussman (Ed.), *Sociology and rehabilitation.* Washington: Vocational Rehabilitation Administration, 1966. Pp. 52–70.

21. Sheldon, Eleanor Bernert and Moore, Wilbert E. *Indicators of social change.* New York: Russell Sage Foundation, 1968.

Controlled Experimentation: Why Seldom Used in Evaluation?

JULIAN C. STANLEY
The Johns Hopkins University

As Edward Suchman (13) and Michael Scriven (7) have indicated, there is a definite though by no means unlimited place in evaluation for controlled, variable-manipulating, comparative experimentation. Modern experimental design and analysis are about 50 years old. By 1923, Ronald Fisher (9) had devised the randomized-block design, probably the first setup in which the categories of one independent variable were fully crossed with the categories of another independent variable in a systematically randomized way. The general factorial design, which incorporated two or more fully crossed "factors," came soon thereafter. By 1935 most of the basic developments still used today had been completed. Five years later in a pioneering book, E. F. Lindquist (4) made them more readily available to educationists and psychologists. Experimental psychologists began quickly after World War II to use the new methods. Controlled experimentation has not, however, been employed in many school-based comparisons of teaching methods or curricula. Why not?

In its simplest form, experimentation of the kind mentioned requires the assignment of sampling units to several different treatments randomly and independently of each other. The unit may, for example, be an individual pupil or a single classroom. Treatments might be several ways to teach physics or biology or mathematics, several different preschool curricula, two or more levels of reinforcement, praise versus blame, and so on. Randomized assignment of units to treatments is crucial because it prevents *systematic* biases in the initial status of the groups. The expected mean of each group is identical with that of any other; of course, if the number of units is small the actual group means may differ considerably at the start of the experiment. (For further explanation, see 2, 8, 10, 11, 12 on page 108.)

Julian C. Stanley

Controlled experimentation in schools seems to have foundered on the requirements of randomization and independence. Most principals resist randomization as though it were an abdication of their obligation to be authoritative. Nearly all school classes are formed without random assignment of pupils to them, even when several teachers teach the same subject at the same period. Certain teachers are assigned the "better" pupils because those teachers are more senior or more powerful or presumed to be better teachers of such pupils. Some pupils or their parents prefer one teacher to another and insist on assignment to his class. Also. pupils must be in those classes they can schedule, and this depends on what other courses they are taking. For example, if physics comes only at the sixth period, those students who enroll for it cannot take American history that period.

These and other restraints on the randomized assignment of pupils are features of the school environment that cannot be ignored by the educational experimenter. Usually, the sampling unit for a school experiment must be the intact classroom rather than the individual pupil. Yet few experiments in educational settings are well enough planned, administered, and financed to involve enough classrooms to give much promise of detecting among treatments differences of the magnitude likely to be produced with the curricular resources and efforts used. That is, many educational experiments lack statistical power and therefore lean toward producing findings of "no difference."

Despite the problems of securing randomness and independence, however, much experimentation could be done in school settings but seldom is. For instance, how much rigorous published literature is there on the benefits of cursive versus manuscript ways of teaching handwriting, despite the fact that each year several million children in a given age group must be taught to write? Any school system could research this area well with its own resources, but practically none does. Instead, each new language-arts supervisor tends to plan the reading curriculum as she wishes on the basis of armchair speculation –not worthless, of course, but leading often to conclusions that fail to convince others.

In an article entitled "Reforms as Experiments," Donald Campbell (1) has pointed out how threatening to administrators scientific methods, particularly controlled experimentation, can be. Powerful procedures can yield results from which the administrator may have no place to hide. Weak methods yield results that can be interpreted to his advantage. To many administrators, weak methods seem better for survival, at least over the short term, than strong ones do.

A somewhat related political consideration is that frequently the prospect of experimentation in schools conjures up thoughts of Frankenstein's monster, inhuman physicians in concentration camps, and the Antivivisection League. A new curricular treatment is judged a priori to be either dangerous (such as esoteric-appearing modern mathematics that has no obvious relationship to balancing a check book) or so likely to be beneficial that to withhold it from any pupils would be educationally foolish (team teaching, classroom TV, or computer-assisted instruction). This widespread carryover from pre-Baconian days glorifies pure thought far beyond the limits found for it in science. Yet it reflects the attitudes of many administrators, teachers, and parents. The researcher may partially nullify this particular objection by promising to remediate members of one group if they are found at the end of the experiment to perform appreciably poorer than do members of another group.

Prejudging experimental results may be due considerably to the treatment-versus-control dichotomy that implies a comparison of something with nothing, whereas in actual practice various alternative ways are compared. An illustration may help: Does two years of Latin improve one's knowledge of English vocabulary more than twc years of direct vocabulary training? Couched in this fashion, the contrast is not between taking Latin versus not taking Latin, but taking Latin versus a potentially effective competitive method. One might find a considerable number of Latin-eligible students at the beginning of the ninth grade and assign at random half of them to study Latin while the other half get vocabulary training. If at the end of the tenth grade one group is superior to the other, members of the lower-scoring group may get special attention until they have caught up. Also, those who studied vocabulary may take Latin in the eleventh and twelfth grades, if they wish. In principle. this kind of experiment is feasible, but so far as I know it has never been done, despite arguments for many years about the vocabulary-building value of Latin.

Of course, most teachers are not equipped technically to plan an experiment and analyze the results, and surprisingly few school systems have research personnel who can help them adequately. Most of the typical educational psychologist's graduate training has been in testing rather than in designing experiments and analyzing ensuing data. Measurement competence is necessary to do good experimentation on most educational problems, but it is far from sufficient. Fortunately, a number of doctoral programs are now producing persons with both sets of competencies.

For a long time I have argued that controlled experimentation has not failed to be of use in evaluation, but that for various reasons, such as those just mentioned, it has seldom been tried. Long-term research in school systems is infrequent. When anything is done, it tends to be too little and not pursued long enough to allow both the novelty of new procedures and the disruption they cause to dissipate. Some nice examples of classroom experimentation such as the doctoral-dissertation studies by Ellis B. Page (5) and William L. Goodwin (3) do appear in the literature, but too often they are one-shot affairs without crucial follow-through.

Apparently, there is more lack of intent, money, and technical resources than of available, applicable methodology. Those critics of experimentation for evaluation who say that controlled, variable-manipulating experimentation may be splendid for stands of alfalfa and weights of pigs but inapplicable to education do not adequately appreciate the generality of Fisherian and neo-Fisherian methods. If, for example, 10 percent of the Physical Science Study Committee's budget from the start had been for rigorous evaluation, undoubtedly a great deal of experimentation could have been done along with other procedures. A decade ago, Walter A. Wittich (6) at the University of Wisconsin did much with far less money.

In conclusion, I agree with Dr. Suchman that the requisite methodological tools for educational evaluation are already at hand. One of these—controlled comparative experimentation—can be of value at any stage of a program, though most likely in the basic-research early phases and the field-experimentation phases. The powerful principle of factorial design can be used to structure the components of a program systematically in order to see which are effective in what combinations. Despite straw men to the contrary, educational experimentation can be as on-going, flexible, and sequential as the cleverness of the evaluators allows it to be. Inflexibility is more in the minds of planners, researchers, and critics than in the methodology itself. Of course, there is no royal road to new knowledge; it is not easy to experiment with human beings, whether they be medical patients or school pupils. In my opinion, however, controlled experimentation and some quasi-experimental designs are important methodological tools of the educational evaluator. Recent attempts to rule experimentation inapplicable because other methods are also useful seem misguided.

REFERENCES

1. Campbell, Donald T. Reforms as experiments. *American Psychologist*, 1969, 24, 409-429.

2. Campbell, Donald T. and Stanley, Julian C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.

3. Goodwin, William L. Effect of selected methodological conditions on dependent measures taken after classroom experimentation. *Journal of Educational Psychology*, 1966, 57, 350-358.

4. Lindquist, Everet F. *Statistical analysis in educational research*. Boston: Houghton Mifflin, 1940.

5. Page, Ellis B. Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology*, 1958, 49, 173-181.

6. Pelia, Milton, Stanley, Julian C., Wedemeyer, Charles A., and Wittich, Walter A. The uses of the White films in the teaching of physics. *Science Education*, 1962, 46, 6-21.

7. Scriven, Michael. The methodology of evaluation. In Robert E. Stake (Ed.), *Perspectives of curriculum evaluation*. Chicago: Rand McNally, 1967. Pp. 39-83 (American Educational Research Association Monograph No. 1 on Curriculum Evaluation).

8. Stanley, Julian C. Controlled experimentation in the classroom. *Journal of Experimental Education*, 1957, 25, 195-201.

9. Stanley, Julian C. The influence of Fisher's *The design of experiments* on educational research thirty years later. *American Educational Research Journal*, 1966, 3, 223-229.

10. Stanley, Julian C. Elementary experimental design—an expository treatment. *Psychology in the Schools*, 1967, 4, 195-203.

11. Stanley, Julian C. The design of educational experiments. In Lee C. Deighton (Ed.), *Encyclopedia of educational research*. New York: Macmillan (in press).

12. Stanley, Julian C. Designing psychological experiments. In Benjamin B. Wolman (Ed.), *Handbook of psychology*. Englewood Cliffs, N.J.: Prentice-Hall, (in press).

13. Suchman, Edward A. The role of evaluative research. In *Proceedings of the 1969 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1970.

Accountability in Public Education

LEON M. LESSINGER
U. S. Office of Education

There is widespread agreement that something has gone wrong in American education. Too many students are leaving school without the basic skills needed for a productive life. Voters are rejecting requests for increased school taxes. Parents are demanding more voice in educational decisions.

To deal with these challenges, our educational managers need what John Gardner has characterized as "a well-tested way out of the dizzying atmosphere of talk and emotion." Gardner's prescription is "to put one foot doggedly after another in some concrete, practical activity."

Application of a new process of accountability to public education is a "concrete, practical activity" which we can use to confront some of our most critical educational dilemmas, including reestablishment of public and student confidence in our education system.

To achieve these results, the emphasis of this new accountability in education must be on what has been learned. Too frequently educational managers attempt to explain their activities in terms of resources and processes used rather than learning results achieved. These explanations no longer are adequate. A more sophisticated public is demanding evidence that every Johnny can read and that he has been provided with the other basic skills necessary to employment and a useful life in an increasingly complex society. The public is demanding "product reliability" in terms of student capabilities and no longer will accept mere assertions of professional superiority in educational matters. If our educational managers wish to retain professional control of the processes of our schools, it is axiomatic that those processes must produce the results desired by the public, who pays the bills.

In its most basic aspect, the concept of educational accountability is a process designed to insure that any individual can determine for himself if the schools are producing the results promised. The most public aspect of accountability would be independent accomplishment audits that report educational results in factual, understandable, and meaningful terms. These independent accomplishment audits might be undertaken by groups drawn from universities, private enterprise, and state departments of education employed by local school authorities in a manner similar to the process now employed to secure and utilize fiscal audits. Such audits would serve our educational managers by telling them which educational processes are productive and which are nonproductive and by suggesting alternatives which are likely to be better.

Like most processes which involve a balancing of input and output, educational accountability can be implemented successfully only if educational objectives are clearly stated before instruction starts. One mechanism for insuring clarity in objectives is the performance contract.

An educational performance contract, as its name implies, would prescribe anticipated learning outcomes in terms of student performance. Unlike contracts which simply describe the work or service to be provided by one party and the payments to be made by the other, the educational performance contract would specify the qualities and attributes of the end product of the service or work performed. In other words, it would establish the quantity and quality of student learning anticipated rather than focusing solely upon the quality and quantity of effort expended by those providing the work or service.

Performance contracts make greater initial demands on both the purchaser and supplier, but they mitigate most postdelivery haggling because basically simple performance tests can be used to determine whether or not the product performs as promised.

If an air conditioning contractor promises that his installation will reduce interior temperatures 20 degrees below outside temperatures, it takes only an accurate thermometer to determine if the promise has been met. Similarly, if an educational manager promises that all children attending his school will be able to read 200 words per minute with 90 percent comprehension on their twelfth birthday, as measured by a specified test, simply giving the test to all children on their twelfth birthday will readily reveal if the promise has been fulfilled.

Neither example, it will be noted, concerns itself with the process by which the performance promise is to be fulfilled. Both do, however, specify the tests to be used as determinants. It is when these tests show shortcomings in performance that fundamental failures of the present educational system become evident.

If cooling were less than he had promised, the air conditioning contractor could reassess his decisions and quickly select alternative procedures and equipment offering high assurance of fulfilling his promise.

The educational manager can also select alternative procedures or equipment in an effort to correct learning shortfalls, but, unlike the air conditioning contractor, he has little assurance that these will be any more effective than those they replace.

Even rudimentary analysis will indicate that there are fundamental reasons for this variation in effective response capability. The factors relevant to air conditioning performance are predominantly physical and permanent, whereas those relevant to learning performance are largely human and constantly changing. Of equal importance, however, is the fact that the technology of air conditioning has been systematically developed in direct response to specific, publicly understandable, and commonly desired performance objectives. The rudimentary technology of education, by contrast, too frequently has resulted from the whims and fancies of the developers and has only accidentally proved successful.

In order to substitute knowledge for personal whims and fancies in the development of an effective educational technology, tools must be found that will identify specific strengths and weaknesses in instructional practice. The development of these diagnostic tools essential to accountability in education is the major problem facing us in the renewal of our education system.

By using such diagnostic tools, educators can begin the laborious process of developing a technology of instruction which will provide answers to the learning shortfalls that will be discovered as educational input and output are systematically assessed.

DISCUSSION

MICHAEL SCR'VEN*
University of California at Berkeley

The papers I am to discuss offer considerable practical wisdom on methods of measuring change, but I am afraid they say little with which I can disagree or on which I can elaborate significantly.

There are, of course, some people—even good people—who disagree with, for example, the emphasis on fully controlled experimentation in Stanley's paper; but the issue has little residual intellectual content, becoming at best a dispute about practicality and at worst an exhibition of misunderstanding. It is more difficult to imagine someone disagreeing with the systems approach, described at some general level, or with the desirability of accountability. I propose to go beyond these compendiums of practical advice and into adjacent disputed territory.

The papers in this symposium represent current scientific and administrative research methods in problems of education. They have in common, by comparison with what would have been written a decade ago, a very much broader concept of what methods can be useful and of how widely one can range in search of solutions to our problems. They exhibit a more flexible and yet more effective approach than we used to have, when Fisherian design and taboos on social evaluation were generally accepted. But they lead us inevitably on towards another confrontation. For every step towards handling the practical aspects of real problems in education beings us nearer to the lair of the value judgment itself. We need only take systems analysis to include society as a whole and we face the question of defending or attacking the values of society, for the values of a society are simply

*Michael Scriven was ill at the time of the confe.ence. His paper was read by Robert H. Ennis of Cornell University, who added comments of his own.

part of a systems solution to the problem of social interaction. One can already see in Glennan's paper one of the consequences of this kind of generalization of the systems approach, where he says that planners are going to be disappointed by the lack of definite answers they will get from systems analysis in this area by comparison with the military logistics area.

To put the matter bluntly, the more fundamental an evaluation is, the more philosophical it has to be. It is perfectly clear that philosophy—in particular, moral philosophy—has already become a crucial issue in educational evaluation, if one thinks about the questions of segregated schools, or age-grading vs. what is called ability-grading (with the latter's tendency towards de facto racial segregation), and so forth.

The time has come to balance the areas of moral and technical evaluation in education, to face the fact that there are always moral presuppositions in any educational evaluation, to improve our skills in identifying and assessing them, and to begin the process of teaching these skills. For one of the most frightening aspects of current attacks on traditional educational practices is that their irrationality is no less than that used to defend the status quo. *That* fact is the greatest indictment of our educational system, not the mere fact that we continue to employ practices for which there is little intrinsic justification, from the repetitive teaching of U.S. history to the refusal to discuss honestly most of the crucial social and moral issues of our time. We do not teach a rational approach to values issues; many of us do not even realize that such an approach is possible; and a great many of us aren't much good at it.

The message of the papers in this session is that an ever-expanding range of evaluation problems can be handled by rational means; yet each solution depends on value presuppositions. If these are really arbitrary, the detailed methodology is a waste of time, icing on a cake of air.

I suggest to you that ethics is itself simply the conjunction of the most fundamental decision theories of the social sciences, the fusion of policy economics, policy sociology, policy political science, and so forth. It is not something different, alien to reason; it is simply the most basic and hence the most difficult level of all management sciences. Good evaluation cannot stop short of examining its own values—at some point. We are being driven backwards into a tacit recognition of this point as we expand the range of evaluation—but we schizophrenically deny its possibility by paying lip service to the idea of the value-free social science, the alleged distinction between facts and values. We should forget that false philosophy and proceed to do what we must need do in order to decide on the merit, worth, or value of educational or mensurational projects. Correlatively, we should be teaching our students how to make such analyses instead of telling them such disastrous falsehoods as that morality comes from religion, or is no concern of the state, or is a matter of personal taste.

While the papers this afternoon did not. on the above account, go far enough in the direction of basic evaluation, from another point of view they began at too abstract a level. They contain no discussion at all of the basic method of educational evaluation, one whose use quantitatively swamps any other. I refer to the practice of grading. Like so many other everyday practices, grading has often seemed too humble to merit the attention of high-powered test and measurement people. My feeling is that it is far more important and in more need of help than anything else they work on. Moreover it admirably illustrates the point just made. that the new critics of bad practices are about as irrational as most defenders of the practices. When done in a defensible way—that is, validly and with supplementary analysis—grading has three essential functions. It provides feedback to the instructor, which helps him judge and adjust his own performance; feedback to the student for the same purpose; and a basis for the allocation of scarce resources to those who can use them best. Two minor functions are feedback to the instructor's peers for (partial) evaluation of his teaching performance and stimulus for the competitively inclined.

Why not use written comment instead of grades? Why not use grades that aren't turned in? Why not use the pass/fail? Every teacher, K—18, should have a perfectly clear answer to each of these questions because they refer to his basic evaluation methodology—but very few do. Every student should have answers to these and other questions like "If we abolish grades, do we use a lottery for graduate school selection?" or "Would five-dimensional grades be acceptable or feasible?" (By "five-dimensional grading" I mean a system in which a student is given a grade on each of five different dimensions of accomplishment within a course rather than only an overall and therefore less informative grade. Dimensions I have in mind are such things as amount of work, originality, quality of reasoning, quality of presentation, and grasp of the subject matter.)

In short, while we may be reaching some degree of accord in the middle level of professional evaluation practice, we have still to bring our minds fully to bear on the most important theoretical and practical problems under this heading; and we are still further from facing the need to teach all our students the skills they need in order to handle the many important evaluation tasks they will and do face within education as well as outside it.

(The following comments were made by Dr. Robert Ennis)

I am in agreement with the points made by Michael Scriven and would be glad to defend them later on. Before doing so, though, I should like to make some comments of my own. I shall raise questions about each paper except Mr. Suchman's.

With respect to Mr. Lessinger's thesis about accountability, I, like Mr. Scriven, have no complaint about the general idea, but I am concerned about the recommendation that we seek performance contracts now. I am concerned because I am worried about the adequacy of the tests that we now have. My concern about this problem is apparently greater than Mr. Lessinger's.

In my own area of critical thinking, for example, which Mr. Lessinger recognizes to be important in the schools, it is not an easy job to make up for the deficiency of a widely available good test. There are none at the moment. To make one requires more understanding of critical thinking than we now have and a great deal of careful test development.

Because we do not have good ways to check performance, we must not start making these performance contracts in critical thinking next year. I am afraid that someone might try to do so. A correlative danger, given the performance contract stance, is to neglect critical thinking because it is not amenable now to performance contracting.

Turning to Mr. Glennan's paper, I should like to express two concerns about cost-benefit analysis. The first deals with determining what is a benefit and what is not; the second deals with the difficulty of predicting important things. My first concern echoes Mr. Scriven's urging that more attention be given to basic value judgments in education. The danger is that the cost-benefit analyst will avoid responsibility for the basic judgments about what is a benefit, and instead will devote his attention to the intriguing technical problems of prediction, of quantifying inputs and outputs in economic terms, and of adapting the whole process to Queen Computer.

In a talk at Cornell last year, one of this morning's speakers, Mr. Friedenberg, held that the current student radical movement is elitist in nature. Although there are numerous exceptions, I think he is basically correct. One basic value question which must be answered intelligently before we put the cost-benefit analyst to work on the technical aspects of his approach is whether elitism is a benefit. Is it? That is a tough question.

My first point then is that we must not forget the value dimension of cost-benefit analysis.

My other point is that the predictions needed to project into the future on significant matters are generally not available to us. We just do not know how to make them.

For example, was the current wave of elitism predictable 10 years ago when James Conant's *The American High School Today* was becoming the standard for American education? Many people at that time claimed that elitism was implicit in the Conant report. Is today's elitism an outgrowth of the Conant era in American education? Who knows? Is today's attack on regimentation and conformity a reaction against the pressures induced by the implementation of the Conant report? Who knows? If we cannot even answer such questions after the fact, how could we have made justified predictions 10 years ago? And how could we have made a cost-benefit analysis of Conant's recommendations without the ability to make such predictions?

In sum, although cost-benefit analysis is at first sight an attractive idea, let us be very wary in introducing it into education. The value judgments required are tough, and the predictive prowess required does not exist.

With regard to Mr. Stanley's paper, I, like Mr. Scriven, generally agree with the desirability of controlled experimentation, and yet I think that Mr. Stanley tends to ignore the real problem accompanying controlled experimentation.

He denigrates armchair analysis, and yet armchair analysis, or some form of intelligent appraisal and judgment, is inevitable after controlled research is accomplished. I mean, we are generally *not* going to get the Stanley-emphasized random sample of the population *about which we want to draw our conclusions,* the population to which we want to generalize. We want to generalize to next year's students, and to students in schools in other parts of the city, county, state, or country. But random (or stratified) samples from these larger groups are impossible or not feasible. Take, for example, Mr. Stanley's Latin experiment. Maybe that will be done in one school, or even (with extreme difficulty) in scores of schools. We will still want to generalize to a much larger population and we cannot get a random (or other legitimate) sample from the larger population.

We have to use our own common sense based upon what we know, and controlled experimentation is not going to do this job for us. We cannot afford to avoid responsibility for that necessary extension of our conclusion to the population from which we have no sample. This extension is the more difficult and hazardous part of the job.

One final observation: Mr. Stanley mentioned the PSSC. He said that if they had spent perhaps 10 percent of the project budget on evaluation (emphasizing, I presume, controlled experimentation with random selection), they would have been in much better shape. I was not involved in the PSSC but I do know that the people who were would have felt very restricted by the requirement of random selection. What they needed for their work were enthusiasm and ability. I do agree that after a while controlled experimentation of some sort would have been desirable, but I wonder whether he realizes how difficult it would have been both to get any kind of random selection and secure the cooperation of top quality academic physicists.

I am not simply pointing to the difficulty of persuading such people of the truth of Mr. Stanley's views, although that is a significant problem. I am also, in continuation of my first point, suggesting that the academic physicists (and mathematicians, and others) are justifiably suspicious of heavy emphasis on controlled experimentation with random selection, for they realize how much it interferes with high quality, creative classroom work. And they realize that its contribution is minimal without the accompanying intelligent judgment that takes account of the existing situation and extends the results to new situations in which the participants are not members of the sampled population.