TOWARD A THEORY OF TESTING WHICH INCLUDES MEASUREMENT-EVALUATION-ASSESSMENT

Benjamin S. Bloom

University of Chicago

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation of Instructional Programs

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

CSEIP Occasional Report No. 9, October, 1968
University of California, Los Angeles

TOWARD A THEORY OF TESTING WHICH INCLUDES MEASUREMENT-EVALUATION-ASSESSMENT

Benjamin S. Bloom

In the 60 years since Binet first introduced his intelligence test, testing has become the pride and despair of psychology and education. Testing runs like a powerful minor theme through most of the research and the applied work in these fields. We take pride in testing because it is the one area which has shown clearest development and most widespread use in these two fields. Our sophistication has grown rapidly in testing, and we know what we know and we know what we don't know in such clear ways that we can take advantage of the former while we attempt to reduce the latter.

But, our despair arises from the overuse of testing, its tendency to dominate both psychology and education, and the negative effect it sometimes has on human relations. Especially in education, testing is a two-edged sword which can do incalculable good as well as great harm to the individual. The recent reaction against intelligence testing in the large city schools, although emotional and in many ways misguided, brings home to us that children are judged in terms of test results and that faith in one child's ability to learn or rationalizations of a teacher's inability to teach another child are both related to test scores.

To control the matriculation examinations of a country is to control its educational system, to develop tests which are widely used for selection and prediction purposes is to determine which

human qualities are prized and which are neglected, to develop instruments which are frequently used to classify and describe human beings is to alter human relations and to affect a person's view of himself.

It is no great exaggeration to compare the power of testing on human affairs with the power of atomic energy. Both are capable of great positive benefit to all of mankind and both contain equally great potential for destroying mankind. If mankind is to survive, we must continually search for the former and seek ways of controlling or limiting the latter. What is needed in testing is a clearer understanding of what we have been doing and a new synthesis of our disparate methods and concepts in testing. Perhaps I can describe a few terms necessary for such a synthesis.

What I propose to do is to describe briefly three very different approaches to the field of testing, indicate why a new synthesis of these is in order at this time, and suggest some of the directions such a new synthesis could take. I do hope that I can impress you with the great need for such a synthesis even though you may be reluctant to accept my suggestions for the synthesis.

Three Approaches to Testing

If we view testing as a systematic method of sampling one or more human characteristics and the representation of these results for an individual in the form of a descriptive statement or classification, we can discern three very different approaches to this problem. For purposes of convenience, I will refer to these

approaches as Measurement, Evaluation, and Assessment. I am sure that some of you will use other terms to describe these approaches. However, the problem is not the accuracy or meaningfulness of the terms, but how to discern the very basic differences underlying these approaches and the contrast among them in the assumptions they make about the world, about man, and about the nature of evidence.

Measurement*

Perhaps the first approach (historically) to testing human characteristics began with the work of Galton and Binet. Although they differed in many respects, what they had in common was the development of standard stimuli, tasks, and questions. The subject's responses to these standard situations were to be appraised in terms of speed and/or accuracy--where accuracy was to be judged in a standard way--by all trained testers. The results for each examinee were translated into some quantitative form (I.Q., raw score, time of response, etc.), which was then given further meaning by relating it to the normative data for a given sample of individuals.

Since testing under this approach usually involves a sample of the individual's responses at a particular point in time (and at a particular point in the individual's career) there has been a great concern for determining the error of the sample by means of methods for estimating the reliability and objectivity of the score assigned the examinee. The meaningfulness of the results

*Some illustrations of the measurement approach are Terman and Merrill (1959), Thurstone (1938), Strong (1943), Gulliksen (1950), and Hathaway and McKinley (1951).

has been usually determined by some form of concurrent or predictive validity. That is, the validity of a measurement instrument is usually approached in terms of its relation with another measurement or appraisal.

Although the measurement view has not entirely ignored the environment in which the individual has developed, the environment is generally ignored at the time of making the measurements. What a measurement specialist does is to attempt to take into consideration the environment as an error term, since he assumes that his measurements are accurate to the extent to which the examinees have had "equal opportunity" to develop the characteristics being sampled. However, the measurement approach does seek characteristics which are "in the individual." That is, the individual is the possessor of I.Q., ability, creativity, etc., and he is to be measured to determine the amount of each characteristic he possesses.

In measurement there is an assumption that the same characteristics (I.Q., memory, etc.) can be measured in all men--no matter what their background--and that the characteristics can be measured in an analogous way at different times and at different places. I.Q. is very similar in 1967 and in 1917 in the United States, France, or India.

The use of the tests under the measurement view is largely for classification, prediction, and experimentation. The major quest in measurement is for a small number of dimensions or measures which will completely account for the variance of a criterion when put together in some additive or summative combination.

The problems which are most alive in measurement today are the search for better units (hopefully with properties akin to physical measurement units), the search for a parsimonious measurement system which will account for the variance of a large number of variables or measures, and the search for improved methods of sampling characteristics and individuals.

The great power of measurement is in its great efficiency. Given a dimension or a criterion, psychometric procedures enable measurement to secure parsimonious procedures for measuring it and for describing it in terms of a small number of dimensions.

Evaluation*

Starting in the 1930's, Ralph Tyler (1934) proposed that educational testing be concerned with the changes in students produced by educational means. He used the term evaluation to refer to a set of procedures for appraising changes in students.

The stress on appraisal of change meant that, theoretically at least, testing had to be done at two or more points in time on each individual to determine the extent of change. Since it was necessary to limit the types of changes to be tested, Tyler suggested that tests be constructed to sample the changes in students specified by the objectives of instruction--that is, the changes which were intended by the instructors, instruction, or the curriculum.

While the evaluation approach is concerned with the reliability, objectivity, and efficiency of the tests used, these are secondary questions. Its primary concern is with the content

*Some illustrations of the evaluation approach are Smith and Tyler (1942), Furst (1958), Bloom (1956), Dressel and Mayhew (1954).

validity of the instruments developed. That is, there must be an adequate definition of the objectives or characterisics to be appraised and a search for ways of testing these characteristics which appropriate experts can agree are sampling the desired behaviors. Once it has been possible to construct a valid test of the objective, it is possible to use concurrent validity to determine more efficient and parsimonious instruments to test the same objective (using the valid test as the criterion). Reliability and objectivity can then be improved until they reach the desired standard.

It should be pointed out that evaluation is concerned with securing evidence on the attainment of specific objectives of instruction. As the objectives become more varied in nature, it is to be expected that a greater variety of types of evidence may be appropriate. Thus evaluation evidence may include products developed by students, processes in which they engage, and behaviors they manifest in a great variety of situations. The evidence may be qualitative as well as quantitative. This is a far cry from the standard stimulus-standard response evidence gathering in measurement.

Evaluation follows the objectives of instruction. Therefore, to the extent that objectives differ from teacher to teacher, school to school, or curriculum to curriculum, it is necessary to devise evaluation procedures appropriate to the specific situations. A single standard test may not be equally appropriate to all situations.

Although evaluation is primarily concerned with changes in individuals, it may be applied to evaluating the effects of a

curriculum, a course, a teacher, a method of instruction, etc. For such problems where the concern may be with group changes rather than individual changes, it is possible to utilize student-test sampling methods which will yield evidence about the group rather than the individuals.

Since evaluation attempts to appraise the changes in students, it is necessary to find methods to judge the extent to which the objectives have been met. The standard against which the evidence is appraised may be the usual type of normative data on particular samples, it may also include absolute criterion-referenced standards, and it may even include the student as his own standard--for example, the change in the student over one period of time as contrasted with the change in that student over another period of time.

Evaluation need not be confined to a summative combination of items or scores. Various patterns of responses may be interpreted to determine the types of changes taking place in the student, the types of errors he makes, and the reasons underlying his attainment or lack of attainment of the objectives specified for instruction.

In measurement, the environment is a source of error in the scores or attainments of the individuals being measured. In evaluation, the environment (instruction, class, school, etc.) is assumed to be the major source of the changes. Ideally, evaluation is as much concerned with the characteristics of the environment which produces the change as it is with the appraisal of the changes in the individuals who are interacting with the environment.

In practice, the evaluator frequently limits himself to a description of the environment while he appraises in detail the changes taking place in the individuals.

One major use of evaluation has been to classify individuals for purposes of grading, certification, and placement or promotion. Perhaps of equal importance is the use of evaluation to determine the effectiveness of a method of instruction, a specific course, curriculum or program, or a specific instructor. Evaluation may be used in education experimentation, and it can be used as a method for maintaining quality control in education.

Perhaps a major difference between measurement and evaluation is the recognition (and utilization) of the effects of testing on the persons involved. Characteristically, measurement strives to limit or control the effects of testing on the student performance. Measurement's concern with "equal opportunity" usually is directed to limiting or equalizing the opportunity students have to learn about the sample of problems on which they will be tested. In contrast, in evaluation there is a more explicit concern with student growth or change and with the <u>utilization</u> of the effects of testing to promote such change. Thus, it is recognized that both teachers and students can be motivated to teach and learn by the nature of the tests they anticipate will be used--this effect can be maximized or minimized as desired. Furthermore, the translation of objectives into testing situations has the effect of giving operational definition to the desired characteristics--and, in turn, such operational definition can focus and intensify the development by teacher and students of these desired characteristics.

Also, the frequency of testing and its use for feedback purposes can do much to enhance the development of the desired characteristics in students.

The major quest in evaluation is for the identification of learning experiences and educative environments which produce significant changes in individuals and for the creation of instruments and methods of testing which will best reveal these changes. The problems which are most alive in evaluation today are the search for better appraisal methods for a great variety of changes (cognitive, affective, psychomotor, etc.); the search for ways of determining the types of changes which are of greatest significance in contemporary societies; the search for more accurate ways of determining change indices; and the search for ways in which evaluation may be best utilized in the promotion of the desired changes (i.e., the use of formative in contrast with summative evaluation).

The great power of evaluation is in its concern for human betterment through a systematic process of relating testing to the development of desirable characteristics in individuals. Used properly, it does much to lead educators to a quest for desirable changes and the means for attaining them. Its means-ends approach has considerable implications for the growth of institutions as well as the growth of individuals.

Assessment*

While the term assessment is a very old one, its use, in the sense in which this paper is concerned, may be attributed to the work of Henry Murray (1938) in the book, Explorations in Personality,

*Some illustrations of the assessment approach are Murray (1938), OSS Assessment Staff (1948), Barron (1963), Stern, Stein, and Bloom (1956), Sanford (1956).

and the book Assessment by Men, by the O.S.S. Assessment Staff (1948) in World War II. As used here it refers to the attempts to assess the characteristics of individuals in relation to a particular environment, task, or criterion situation.

Assessment in this sense is as much concerned with the environment as it is with the individuals who interact with the environment. The need-press scheme of Murray has been useful in analyzing the individual and the environment in analogous terms. The use of role theory has been effective in relating the roles demanded or emphasized by the environment with the roles which the individual is able to "play."

Assessment characteristically begins with an analysis of the criterion and the environment in which the individual lives, learns, and works. It attempts to determine the psychological pressures the environment creates, the roles expected, and the demands and pressures--their hierarchical arrangement, consistency, as well as conflict. It then proceeds to the determination of the kinds of evidence that are appropriate about the individuals who are to be placed in this environment, such as their relevant strengths and weakness, their needs and personality characteristics, their skills and abilities.

The evidence collected about the individual in assessment is multiform in that many types of qualitative and quantitative evidence may be collected, some of it highly structured and some of a more projective or unstructured form. The assessor may use evidence from self-reports, observations by others, interviews, projective situations, situational tests, role playing, free association, etc. Relevant evidence on a particular characteristic may be

decade assessment has been used to analyze the characteristics of the environment or criterion situation in order to better understand how environments or situations differ and the kinds of demands they create or the ways in which they influence human characteristics. It is safe to say that, with a few exceptions, in the last 15 years there have been few contributions of assessment to new instruments or methods of testing individuals, while there have been major contributions to analyzing and testing the environment (Dave, 1963; Wolf, 1964; Hess and Shipman, 1965; Stodolsky, 1965; Pace and Stern, 1958; Pace, 1963).

The major strength of assessment is in the search for evidence on both individual and environment. The attempt to relate the two types of evidence has contributed more to the understanding of phenomena than it has to the prediction or control of such phenomena--although the first does give a basis for the second. The problems which are most alive in assessment today are the search for more effective and efficient instruments to understand both the individual and the environment, the improvement of methods of processing evidence from a variety of instruments, and the development of more adequate ways of securing evidence on the criteria to be predicted.

Conditions Which Make A Synthesis Possible and Necessary

It is the writer's opinion that a synthesis of these three approaches to testing is more possible at this time than ever before. Until recently, many of us were so concerned with the distinctive characteristics of each approach that we tended to

overemphasize the differences. We suggested that each of these approaches had its own value and that each could make a useful approach to those problems for which it was best fitted.

Each of these approaches can be considered to be a partial view of the nature of man, the world, and the nature and use of evidence. If this is so, then we must seek a more complete view of man, the world, and evidence which can best utilize what each of these approaches offers at present. It is the writer's belief that we are in desperate need of such a synthesis--if we are to avoid further narrowing of the field of testing. Further, that any attempt to bring these approaches together is likely to bring about a period of "hybrid vigor" in which new problems and new techniques will give the field of testing a period of unprecedented challenge and growth.

What are some of the conditions which make this synthesis possible or necessary at this time?

Experience

We now have had approximately 60 years of experience with measurement, about 30 years of experience with evaluation, and almost the same amount of experience with assessment. There is an extensive literature already available and it is possible to use this literature as the basis for building on the foundation already available. Furthermore, the limits and the uses of each approach are recorded so that one can start where these leave off.

The length of the history of each of these approaches suggests that the pioneers who helped to create and develop each approach

are now replaced by new generations of workers and students who are less committed to the differences (and the emotional involvement of the pioneers) and who may take a more dispassionate view of the problems and opportunities for which various combinations of the three approaches may be useful and even necessary.

Linked to this history is the development of a large number of instruments and techniques which have been used under a variety of conditions. This means that the original ideas of the pioneers have been given an operational meaning and illustration and that one is no longer left to deal with the original verbal formulations--but can deal with the operational consequences of these formulations. Furthermore, the instruments and techniques yield a reservoir of procedures that can be built on as new problems are identified. Of special value in this connection are the reviews, bibliographies, and collections of information represented by the Buros' Mental Measurement Yearbooks, the reviews of educational research, the yearbooks, and the journals, such as Educational and Psychological Measurements, Psychometrika, and Measurements Used in Education. The point is that much of the information about tests and testing is now in a codified form which is easily located.

Sophistication in Statistics and Data Processing

During the past twenty years there has been an unprecedented development in the sophistication of statistical methods and data processing. In part this growth has been responsive to developments in testing, but in large part it has been quite independent of this field.

Factor analysis, multivariate procedures, canonical correlation, path coefficients, sampling methodology, etc., have moved very far during this period such that problems which were difficult or impossible to attack before are now amenable to an efficient and effective solution. Factor analysis as a method of reducing the detail and dealing with a smaller number of variables (or tests) has been effectively used in testing. Other statistical procedures are increasingly becoming available which are likely to enable testers to deal with more complex problems than have been possible up to the present.

It is, however, the computer that is likely to make the greatest difference in the work of testers. The enormous amount of data that can be stored, the ease with which data can be analyzed and summarized in a great range of ways, and the storage of longitudinal as well as cross-sectional data should enable the tester to attack problems which have hitherto been out of his reach.

A synthetic theory or approach to testing can give direction and meaning to this increased sophistication and ease in data collection and analysis. Problems which were impractical to attack in 1960 are relatively easy to attack in 1968. Furthermore, the facility with which complex theoretical ideas can be dealt with in the day-to-day data processing of test constructors and test users make a synthetic theory of testing a highly practical concern--where hitherto it might have seemed to have a set of ideas that could be dealt with only by a few highly skilled persons in the field.

Development of Educational Methods and Learning Theory

Testing is most powerfully related to education and to learning, instruction, and research in the schools. This is not to say that it does not have great value for industry, human development, and psychology. However, the advances in education have been such in recent years as to create new needs in testing.

Some of the advances in education that create the need for advances in testing have to do with new curriculum developments, new problems in education arising from new tasks being assumed by the school, and basic changes in instruction and instructional technology.

The new curriculum developments in mathematics, science, languages, and social sciences have been of a magnitude not dreamed of previously. Large teams of experts and specialists have been involved in a systematic approach to new curricula in which basic changes in the content and structure of the subject matter have been accompanied by a variety of instructional materials and methods. In addition, many teachers have been provided with in-service training on the new content and instructional material. These changes in curricula have not always been supported by changes in testing procedures at the level required for evaluation of the cognitive and affective consequences. There is a need for a theory of testing which will be appropriate to the new curricula and the problems they pose for testing.

The schools are very rapidly taking on many tasks which have hitherto been assumed by the home, social welfare agencies, employers, etc. For example, pre-school programs are being devised

for the culturally disadvantaged children to compensate for presumed inadequacies in the home's preparation of these children for education by the schools. Problems of integration of ethnic and racial groups are being thrust on the schools as are some of the problems of providing food, medical care, and special instruction (and day care) for children of poverty groups. The special problems of youth in need of employable skills (aside from the regular academic instruction in the schools) are being assumed by schools and other educative agencies. As the schools assume tasks which represent a departure from previous practices, there is an especially urgent requirement that testing and related methods of gathering evidence be appropriate to insure that the task is well done and that the consequences of performing the new task are positive and desirable rather than negative and harmful to the individual, the schools, and the community. New tasks require the appropriate development of new testing procedures as well as new conceptions of the nature of testing.

There have been some major changes in instructional methodology, such as programmed instruction, T.V. and videotape, computer-assisted instruction, and even the widespread use of tutors. These, the development of discovery and inquiry methods of teaching, and the use of non-graded school programs raise many new problems for testing and related activities. In general, the development of approaches to individualized learning make it necessary to develop more effective ways of gathering and using relevant evidence.

Testing must serve not only to determine the effectiveness of these new procedures and programs, it must also serve to help us understand the nature of the phenomena involved in order that educational policy and practice may increasingly be based on such understanding. Furthermore, testing must serve to predict consequences of particular educational decisions as well as to provide quality controls on the implementation of these decisions in practice. Such tasks require a larger and more complex theory of testing than is presently available.

Effect of Testing on the Phenomena

During the past few years there has been a great deal of criticism directed against testing. These criticisms should remind us that testing cannot be completely separated from the phenomena it attempts to record and study. The act of testing in the social sciences affects the humans who are involved. Especially in education, what is tested influences the perception of students, teachers, parents, and others in the society. To measure intelligence, specific attitudes, and achievement is to influence the values of the society, the ways in which people value themselves and others, the nature of educational policy and practice, and the very ends of education. External examinations do much to influence the curriculum, the ways in which students view school, and the things students study and learn.

Try as we will to control the effects of testing, we find that the nature of the tests used may under some conditions do more to influence student learning and teacher practice than the

other educational procedures which we regard as the substance of education.

A full awareness of the consequences of our ways of testing must be an intrinsic part of our use of tests to understand, predict, or control human behavior. We must, in the development of test theory, give full recognition to the range of effects the tests may have on the society as well as on the schools, the students, and teachers.

The Interplay of the Different Approaches

One way of relating the different approaches to testing is to recognize the special qualities of each and to have them support each other in attacking specific problems.

Thus, we may approach a problem of prediction by making an assessment approach to the individual, the environment, and the criteria. Using a great variety of evidence-gathering instruments and a theoretical framework, we may complete a very comprehensive assessment approach to our prediction problem. Such an approach is very costly in terms of resources and personnel required and makes use of complex human judgments--frequently requiring rare clinical skills and a team of experts not only to collect the data but to make the clinical appraisals of the data and their implications. If such an approach is relatively effective, the insights, instruments, and criteria can then be systematically reduced to a more efficient and parsimonious set of procedures by the use of measurement methodology. And, with computer analysis of very complex patterns of data, it is quite likely that the measurements

and their processing can yield results not significantly inferior to that secured by the most costly assessment approach.

Or, the situation can be reversed. Through measurement approaches, it is possible to efficiently measure and predict certain behavior. Quite frequently, the measurement includes symptoms and behaviors which yield satisfactory levels of prediction, but which are not "understood" in terms of why the relations are what they are. Assessment may be used to probe more deeply into the underlying relationships and into the reasons which would help to explain the results. Thus measurement, sometimes blind to causation, can be supplemented by assessment to probe into the theory and underlying behavior to account for the measurement results.

Evaluation seeks to determine the extent to which change has taken place in students as the result of particular learning experiences. Once the learning experiences have been defined and described, evaluation attempts to account for the changes in students in terms of the effectiveness of these learning experiences.

Measurement may be used in relation to evaluation in finding more parsimonious procedures for describing and testing the changes made. Are there multiple changes or are the changes accountable for in terms of a single factor or very few factors? Measurement can do much to improve the techniques for determining the amount of change and for determining the characteristics in the students and/or selected characteristics in the environment which "account" for the changes. Measurement can be used to make the evaluation instruments more efficient.

and/or evaluation approach has been used. Nor is it likely that assessment would be used to probe more deeply into a problem that has first been attacked from the measurement or evaluation point of view.

What is needed is a more comprehensive approach to testing that fully utilizes what we have already learned and can do with each of the distinct approaches to testing. What is needed is a comprehensive theory of testing which will give direction to the training of testers in the future and which will show the testing tactics required in attacking any given problem. Such a theory should help us determine the kinds of specialized personnel, team efforts, instrumentation, and data processing relevant for any given problem. This paper does not attempt to provide a quick solution to theory building of this type.

What will be done in the remainder of this paper is to suggest some of the ways in which such testing terms as VALIDITY, RELIABILITY, and NORMS might be altered to take care of some of the problems posed by measurement, evaluation, and assessment. The expansion and redefinition of these terms could serve to enlarge the range of ideas and methods available to testers and as a result carry us one step toward a synthesis of these approaches to testing.

In addition, two problems in testing have been selected, the determination of <u>stability</u> and the appraisal of <u>change</u>. For each of these problems, the writer has suggested some of the ways in which the problem would be altered if the three testing approaches were used simultaneously.

It is to be hoped that further work along these lines will pose the issues and underlying assumptions more clearly. Hopefully,

out of such work will come theoretical developments which can effectively synthesize what at the moment appear to be very distinct and different approaches to testing.

We might begin with a brief definition of testing which encompasses all present theories of testing. Testing may be defined as the act of gathering and processing evidence about human behavior under given conditions for purposes of understanding, predicting, and controlling future human behavior. While such a definition leaves much to be further defined, it does help us delimit the phenomena for which a theory must account.

Validity

Perhaps the key problem in testing is to establish the validity of the instruments and techniques developed. While this is the most difficult problem to solve in actual operation, the place of validity in relation to measurement, evaluation, and assessment has largely been solved by the work of The American Educational Research Association (1955) and the American Psychological Association (1954) in their attempts to delineate the different types of validity. In these reports, the committees described four types of validity:

Content Validity
Construct Validity
Concurrent Validity
Predictive Validity

Any testing instrument may be validated by one or more of these types of validation. It would, of course, be the rare test which would be validated by all four types. As we review these different types of validity, it is striking that each of the

testing approaches has emphasized particular types of validity and has developed techniques and procedures for utilizing the preferred types of validation.

Thus, evaluation has stressed content validity and has developed the techniques of defining objectives and content in behavioral terms such that competent judges can determine the appropriateness of particular test problems and situations for the defined specifications. While evaluation does make use of concurrent validation when it seeks to make more efficient instruments which will yield results relating to the original and more direct instrument (validated by content validity), it is clear that the emphasis in evaluation is on content validity.

Assessment has characteristically emphasized construct validity, since it goes to more elaborate lengths to use theories or models to guide it in particular assessment situations. It is these theories or models which make it possible to use construct validity, and it is the very complexity of the data collection and analysis required in assessment which make it necessary to have a theory or model to guide the workers through the intricacies of an assessment process.

Measurement has characteristically employed predictive and concurrent validity. And, as was pointed out earlier, given a criterion, measurement has very powerful methods of developing instruments which will yield maximum predictive and concurrent validity.

The suggestion that emerges from the work to date on validation is that an approach to validation which makes use of the

best features of the particular test approaches leads to an inclusive approach rather than to any new concepts of validity. Thus, for our purposes, validity may be defined by four terms:

Validity = CONTENT Validity : CONSTRUCT Validity : CONCURRENT Validity : PREDICTIVE Validity

It is the task of the tester to determine which types of validity are germane to the problem at hand and to determine when he has exhausted the validation possibilities for the particular test problem. Hopefully, the tester who is carefully trained in the different approaches to testing can be more ingenious and creative than those of his predecessors who tended to rely on a single type of validation. Hopefully, also, content and construct validity would become more central in the initial approach to a testing problem, while concurrent and predictive validity would become more central to the development of more efficient testing procedures <u>after</u> criterion measures with high content and construct validity have been created.

Reliability

Reliability has been the one testing concept that has been most fully developed, although primarily from the measurement point of view. If reliability is to more adequately deal with the problems posed by each of the testing approaches, several additional terms and operations might be added to the more traditional approach to this concept. Perhaps a more comprehensive view of reliability might include the following:

Reliability = READER Reliability : INTERNAL CONSISTENCY : INSTRUMENT STABILITY EXAMINEE Reliability : SAMPLING Reliability : CONGRUENCE Reliability

READER Reliability as the agreement of competent judges on the meaning or value of a particular product, response, or process presents few difficulties. This type of reliability or objectivity was one of the first problems attacked by testers. However, the definition of a competent judge would vary for each of the testing approaches. For measurement, this could vary considerably depending on the human characteristics being measured; for evaluation, especially in education, this is likely to be a person with considerable competence in the subject matter and learning processes under consideration, while in assessment this is most likely to be someone with considerable training and experience in the use of dynamic theories of personality or clinical psychology. However, the main point is that the specific testing problem must determine the qualifications needed by the judges for reader reliability.

INTERNAL CONSISTENCY types of reliability are estimates of the extent to which a scale or test contains items which are getting at a common characteristic, trait, or factor. This type of reliability has been widely used in both measurement and evaluation. It must be recognized that internal consistency is both a function of the items in the test and a function of the subjects being tested. Thus, from the evaluation point of view it is quite likely that the internal consistency of a set of items may be higher after the subjects have had the relevant learning experiences than prior to these learning experiences. It is necessary in stating the level of internal consistency to give some indication of the nature of the subjects used in determining this form of reliability.

INSTRUMENT STABILITY indices are attempts to determine the error likely to be attached to a particular score as a result of fluctuations in the performance of the examinee from sample to sample. As the time intervals between samples increase, there is a shift from the usual concept of test reliability to the stability of the instrument or characteristic under consideration. There is increasing evidence (Bloom, 1964) that particular characteristics become more stable at some ages or stages of development than at others and that under some conditions the reliability of the test results over a five-year span may be as great as it is over a five-day span of time. This type of reliability is needed for problems encountered in each of the testing approaches. Especially when major decisions are being made on the test results (e.g., admission to a special school for the mentally retarded, admission to a particular educational program, guidance with regard to a vocational career), it is important that the tester be able to indicate the stability of the test results. This is developed more fully on pages 32 to 34. Here, it may be pointed out that a stability index would serve to caution the user of test results against long-term decisions where the stability is low, and it would caution the psychologist or educator against overoptimism with regard to changes in an individual (e.g., intelligence, values, problem-solving) where the stability of test results is very high.

EXAMINEE reliability. While reliability is generally attached to an instrument in relation to a particular group of subjects, the tester is finally interested in the reliability with which he

between performance on the two tests would form a type of sampling reliability.

This type of reliability is especially useful when one is attempting to extrapolate from the performance on the sample to the population being sampled. An illustration of this is Terman and Merrill's (1959) attempt to estimate the size of an individual's vocabulary from his performance on the vocabulary sample in the Stanford-Binet Intelligence Test. They selected a random sample of the words in a particular dictionary and then attempted to generalize from performance on the sample to the total set of words in the dictionary. While it is probable that this form of reliability would be appropriate to all three test approaches, it would be most useful in describing test results from the evaluation point of view where the tester is trying to give meaning to the individual's score as representing performance over the entire set of content and behaviors being sampled.

CONGRUENCE reliability is a difficult concept to explain. It arises especially in assessment where a variety of evidence is used to assess a particular individual or group. Thus, evidence on leadership qualities may come from self-reports, projective techniques, observations in special situations, and from reports of superiors or subordinates. Finally, the assessor must put all this evidence together in a descriptive statement or in a rating. What he needs is some way of estimating the congruence of all the pieces of evidence. With what certainty can a particular statement or rating be made? While it may be straining a point to call this reliability, this seems to me to be the most appropriate place to put this problem. Another illustration is the college admissions

officer's problem in determining the admissability of an applicant on the basis of previous grades, test scores, interviews, and letters of recommendation by secondary school personnel. When all the evidence is positive or negative, he has little difficulty--the results are congruent. When the evidence yields conflicting pictures of the candidate, the admissions officer has difficulty in reaching a sound decision.

The determination of congruence may require great insight on the part of the interpreter of the evidence, since it is quite possible for what on the face of it appears to be contradictory evidence to really be highly consistent in the light of a particular theory of personality or human behavior. Thus, contradictory projective test results and self-reports may be perfectly congruent for particular characteristics such as aggressiveness, anxiety, attitudes toward persons in authority, etc.

One may draw an analogy between the attempt to determine congruence in testing with medical diagnosis on the basis of a great variety of symptoms and evidence. The medical practitioner begins with the assumption that there is a medical explanation for the different symptoms--that is, he assumes that the evidence will be congruent, if he can find the appropriate ailment, cause, or condition. I have no clear suggestions for the form that a <u>congruent reliability</u> index might take, since it is both a qualitative as well as a quantitative problem. However, I suspect that we may find leads to it in one or more error terms based on multivariate methods which relate a variety of predictive indices to a variety of criterion indices.

Here again, it is the task of the tester to determine which types of reliability are relevant to the problem at hand. Relatively simple types of validity are required for many measurement problems, while more complex forms of reliability may be required for some assessment problems. However, it is the use to which the evidence is to be put which will be the primary determinant of the form of reliability that is appropriate.

Norms

Test results are usually given meaning in relation to normative data of some sort. Especially for aptitude and educational achievement tests, test makers have devoted a great deal of time and resources to the securing of normative data. Quite frequently more resources are expended on the development of norms than on the construction of the instruments themselves. Several types of norms are suggested for the problems encountered in three approaches to testing. A more comprehensive approach to the development of norms for tests might include the following:

Norms = DISTRIBUTION Norms : INTRA-PERSON Norms : CHANGE Norms : CRITERION REFERENCED Norms : SEQUENTIAL Norms

DISTRIBUTION Norms. These are the usual type of norms in which a well defined sample of individuals take a given test and the results for individuals and groups are related to the appropriate distributions. Such norms are indispensable to the measurement approach, while they are useful to the other test approaches.

INTRA-PERSON Norms. For some tests problems where there are several scores (e.g., Differential Aptitude Tests, Kuder Preference

Record), it is useful to have norms on the differences between pairs of scores as an additional basis for interpreting scores. This type of norm would appear to be most vital for evaluation and assessment problems. However, there are many problems of interpreting measurement test results for guidance purposes which could make use of this type of norm.

CHANGE Norms. Especially for evaluation of change as a result of therapy or education, it would be useful to have norms on change scores. Such norms could indicate the statistical significance of particular measures of change, and they could also indicate the frequency with which a particular change is found under given conditions. Thus, a measure of change in vocabulary or other language measures under various pre-school programs, reading and language programs in the elementary school years, etc., could do much to help in the evaluation of the changes produced by a given curriculum or learning strategy. It is likely that this type of norm would be most useful for the evaluation approach.

CRITERION-REFERENCED Norms. Glaser (1963) has advocated norms which indicate the attainment of a particular level of skill or competence. These are not norms in the sense of distributions. Instead, they make use of definitions of a task attainment or expert judgment to determine particular standards of performance. While Glaser was recommending this type of norm from the evaluation point of view, it is likely that procedures for determining such norms would be most valuable for the assessment approach.

SEQUENTIAL Norms. A somewhat different type of norm is suggested by the problem of evaluation. Given a set of scores at two

or more points in time, what are the expectancies for a third or later point in time? This type of norm would be especially useful in longitudinal studies of school achievement. Thus, Payne (1963) finds that achievement at grade 6 can be highly predicted from achievement in grades 1 and 2. Such predictions, put in the form of normative data, would help to determine the long term consequences of the changes taking place over shorter time intervals. The great value of such sequential norms is that they could alert educators, therapists, medical practitioners, and others to the consequences of present procedures. Such norms would make it possible to use the time interval between the prediction and the consequences to take those steps that may prevent consequences which are regarded as undesirable or to maximize consequences regarded as desirable.

It is quite likely that other types of norms would be useful, in addition to these named in the foregoing. The major point is that an expanded view of the scope of testing requires the development of a variety of terms and operations to deal with the changing nature of the problems in this field.

SOME APPLICATIONS OF A SYNTHETIC THEORY

Perhaps some ideas of the value of a synthetic theory of testing may be seen in the consequences it could have for several problems. The reader is invited to consider other problems in the special fields with which he is concerned. Here, we will limit ourselves to the consideration of two rather general problems: stability of human characteristics and the appraisal of change.

Stability

We have already referred to stability as one form of test reliability. However, the problem of determining stability of a human characteristic is one which goes beyond the long-term reliability of an instrument. Given a human characteristic which can be measured by one or more tests at a particular age or development, what will be the most probable state of that human characteristic at some point of time in the future? Thus, if we measure height, general intelligence, language competence, anxiety, etc., at age 6, what can we expect on similar measurements at age 7? age 10? age 18?

It is suggested that some basic terms in the determination of stability might be the following:

$$\text{Stability}_{X_1 \text{ to } X_2} = \text{INSTRUMENT Stability}_{X_1 \text{ to } X_2} + \text{DEVELOPMENT}_{X_1} + \text{ENVIRONMENT}_{0 \text{ to } X_1} + \text{ENVIRONMENT Future (Likely)}_{X_1 \text{ to } X_2} + \text{ENVIRONMENT (Future Ideal)}_{X_1 \text{ to } X_2}$$

Each of these terms is briefly described or explained in the following.

INSTRUMENT Stability $_{X_1 \text{ to } X_2}$. This is merely the long term reliability of the instrument over the period of time X_1 to X_2. This term was explained briefly on page 26. Thus, it might be the stability of the Stanford-Binet Intelligence Test for ages 6 to 10 or the stability of the Stanford Reading Comprehension score grade 2 to grade 5.

DEVELOPMENT $_{X_1}$. This would be some index of the level of development in the particular age or grade. Ideally, this should be on some scale of absolute development.

ENVIRONMENT $_{0 \text{ to } X_1}$. This would be some index of the relevant environmental characteristics (for the particular human characteristic) over the time interval birth to X_1. However, an approximation to this may be secured by an estimation of the environment at time X_1. (See Wolf, 1964, for the home environment index for general intelligence; Stodolsky, 1965, for maternal behavior influencing language development.)

ENVIRONMENT $_{\text{Future (Likely) } X_1 \text{ to } X_2}$. This can only be an estimation of what is likely to take place between times X_1 and X_2. While we need to know a great deal about the stability of environments over periods of time, it is possible to make some estimates of this for those age periods in which the home and the school are the dominant environments. There is some likelihood that, barring major crises, the home environment is not likely to be fundamentally altered over a three-year period with respect to these characteristics in it which influence general intelligence or language development. Also, it is likely that a particular school environment will not be fundamentally altered over a three- to five-year period insofar as it affects language development or reading--if studies of the school as a bureaucracy can be relied upon. In any case, an environmental study of the school over the previous three or four years gives some indication of what may be expected in the next few years.

ENVIRONMENT $_{\text{Future (Ideal) } X_1 \text{ to } X_2}$. This is an estimation of what is possible if the environment approximated some ideal: (a) if the home environment during period X_1 to X_2 approximated the best home

environment for these characteristics (e.g., general intelligence, language development), (b) and/or if the school environment during this period approximated the best school environment for these characteristics.

Where the ideal environment is similar to the likely environment, and the past environment, stability of the characteristic is likely to be greatest. Where the ideal environment is very different from both the likely and past environments, stability of the characteristic could be considerably decreased if some intervention measures are successful in the attempt to produce such an environment. The point is that the use of these different terms in the estimation of stability provides a basis for determining the conditions under which stability is likely to be maximal or minimal and thus serve as a basis for intervention if this is regarded to be in the best interests of the individual.

In a problem of this type, the distinction between measurement, evaluation, and assessment are no longer as clear as they were in the consideration of validity, reliability, and norms. Techniques based on all three approaches to testing would be used in the determination of stability. Each of the terms suggested for stability would profit from a synthetic approach to testing.

Change

Another problem to which a synthetic theory of testing might be applied is the appraisal of change--a problem to which the evaluation approach to testing has been applied in the past. Some of the possible terms for the appraisal of change might be the following:

$$\text{Change} = \begin{matrix}\text{FINAL}\\ \text{Status}\end{matrix} - \begin{matrix}\text{INITIAL}\\ \text{Status}\end{matrix} + \begin{matrix}\text{STRATEGY}\\ \text{Employed}\end{matrix} + \begin{matrix}\text{EFFECT OF}\\ \text{Instrument}\end{matrix} + \begin{matrix}\text{RELATED}\\ \text{Changes}\end{matrix}$$

FINAL Status. This represents the post measurements for each of the characteristics under consideration. Ideally, these instruments are parallel in content and form to those used in the initial status tests, with high validity and reliability. Both the initial and final status measurements should have high content and construct validity based on the specifications desired in the change.

INITIAL Status. This represents the initial measurements for each of the characteristics under consideration. Presumably each of these characteristics is tested with instruments which have high validity and reliability.

STRATEGY Employed. The specific learning experiences, therapeutic techniques, environmental intervention, or other strategy used must be described in sufficient detail to delineate it from other strategies and, if possible, the context in which the strategy is used must be included in this description. In the past, it has been common to label two or more learning strategies, which quite frequently turned out to be very similar in major respects.

EFFECT of Instrument. It is possible for the instruments used to appraise initial and final status to be as powerful as the learning strategy in producing the changes. There must be some way of distinguishing the effects of the strategy from the effects of the instruments used. Research design procedures represent one approach to this problem.

RELATED Changes. It is likely that other changes take place in addition to those specifically sought and appraised by the

initial and final status instruments. Some of these related changes may be regarded as desirable while others may be regarded as undesirable. It is possible that the constructs of the assessment approach may be useful in hypothesizing what related changes are likely to be produced in relation to the strategy employed, the instruments used, or in relation to the changes showing up in the comparison of initial and final status measurements. The basic problem here is one of limiting the number of related changes that are to be investigated to those which are most probable.

Here again, it is likely that a problem of the type suggested above could only be attacked by a combination of the resources and techniques of all three of the existing test approaches.

CONCLUSION

The main thesis of this paper is that testing is now ready for a major effort to create a synthesis out of what has hitherto been a series of unrelated approaches to testing. Such a synthesis is necessary if testing is to adequately deal with the very complex problems of describing, explaining, and predicting human characteristics. The attempt in this paper is to indicate the ways in which some of the powerful aspects of each testing approach may be brought together into a more complex way of handling test problems.

Perhaps the major weakness of this paper is that it approaches a synthesis of testing methods by adding terms to the more traditional ones. Hopefully this is only one step toward a more effective synthesis which creates an entirely new view of testing with fewer terms and clearer operational procedures than can now be described.

The value of work toward a new synthesis would be in its effect on the training of a new generation of specialists in this field as well as in opening up to a greater variety of attack those problems which have hitherto been regarded as the special province of a single approach--whether it be measurement, evaluation, or assessment.

REFERENCES

American Educational Research Association. Technical recommendations for achievement tests. Washington, D. C.: AREA, 1955.

American Educational Research Association. Technical recommendations for psychological tests and diagnostic techniques. Washington, D. C.: ADA, 1954.

Barron, F. Creative and psychological health. Princeton, New Jersey: Van Nostrand, 1963.

Bloom, B. S. (Ed.). Taxonomy of educational objectives: Handbook I: Cognitive domain. New York: David MacKay, 1956.

Bloom, B. S. Stability and change in human characteristics. New York: Wiley and Sons, 1964.

Dave, R. H. The identification and measurement of environmental process variables that are related to educational achievement. Unpublished doctoral dissertation, University of Chicago, 1963.

Dressel, P. L., & Mayhew, L. B. General education: Exploration in evaluation. Washington, D. C.: American Council on Education, 1954.

Furst, E. J. Constructing evaluation instruments. New York: Longmans, Green, 1958.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 17, 519-21.

Gulliksen, H. Theory of mental tests. New York: Wiley and Sons, 1950.

Hathaway, S. R., & McKinley, J. C. Minnesota multiphasic personality inventory. New York: Psychological Corporation, 1951.

Hess, R. D., & Shipman, V. Early experience and socialization of cognitive modes in children. Child Development, 1965, 36, 869-886.

Murray, H. A. Explorations in personality. New York: Oxford University Press, 1938.

OSS Assessment Staff. Assessment of men. New York: Rinehart, 1948.

Pace, C. R. College and university environment scales: Technical manual. Princeton: Educational Testing Service, 1963.

Pace, C. R., & Stern, G. G. An approach to the measurement of psychological characteristics of college environments. Journal of Educational Psychology, 1958, 49, 269-277.

Payne, A. The selection and treatment of data for certain curriculum decision problems: A methodological study. Unpublished doctoral dissertation, University of Chicago, 1963.

Sanford, N. (Ed.). Personality development during the college years. Journal of Social Issues, 1956, 12, 1-71.

Smith, E. R., Tyler, R. W., et al. Appraising and recording student progress. New York: Harper, 1942.

Stern, G. G., Stein, M. I., & Bloom, B. S. Methods in personality assessment. Glencoe, Illinois: Free Press, 1956.

Stodolsky, S. Maternal behavior and language and concept formation in Negro pre-school children. Unpublished doctoral dissertation, University of Chicago, 1965.

Strong, E. K., Jr. Vocational interests of men and women. Stanford: Stanford University Press, 1943.

Terman, L. M., & Merrill, M. A. Measuring intelligence. Boston: Houghton Mifflin, 1959.

Thurstone, L. L. Primary mental abilities. Psychometric Monographs, 1938.

Tyler, R. W. Constructing achievement tests. Columbus: Ohio State University Press, 1934.

Wolf, R. M. The identification and measurement of environmental process variables related to intelligence. Unpublished doctoral dissertation, University of Chicago, 1964.