DOCUMENT RESUME

ED 102 208 95 TM 004 509

AUTHOR Bradley, Robert H.; Caldwell, Bettye M.
TITLE Issues and Procedures in Testing Young Children. TM Report No. 37.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO TM-37
PUB DATE Dec 74
CONTRACT OEC-0-70-3797(519)
NOTE 16p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Affective Tests; *Children; Civil Liberties; Cognitive Tests; *Decision Making; Elementary School Students; Evaluation Criteria; Observation; Preschool Education; Psychomotor Skills; Test Bias; *Testing; Test Reliability; *Test Selection; Test Validity

ABSTRACT

Because of the developmental characteristics of young children, the potential user of tests for educational evaluation needs to be keenly alert to the kinds of decisions which can be made on the basis of testing and to the limitations of testing when young children serve as subjects. The choice of a test to be used will depend on the type of decision to be made. Several decision types are discussed: program planning and evaluation, screening, and administrative decisions. After considering the type of decision to be made as a factor in test selection, several test characteristics should be considered: practical criteria, including relevance, scope, timeliness, importance, efficiency, and credibility; validity--content, discriminant, criterion, content, and edumetric; test reliability; and test bias. Cognitive, affective, and psychomotor domains are discussed in relation to determining what kinds of test characteristics are important in assessing for a particular educational decision. Some strategies which offer effective ways to obtain more ecologically valid assessment data are presented along with comments on other testing methods, human rights, and achievement tests. (RC)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 37 DECEMBER 1974

ISSUES AND PROCEDURES IN TESTING YOUNG CHILDREN

Robert H. Bradley and Bettye M. Caldwell*

INTRODUCTION

Much ado has been made of the Age of Accountability in education. The assumption that educational programs should be accountable to their sponsors and that some objective evidence should be offered to prove that the programs actually accomplish what they claim to be able to accomplish has been one of the least popular ideas to impact professional education in recent years. There is no doubt that extravagant claims are sometimes made for unproven educational programs. For example, just a decade ago, proponents of early childhood education were suggesting that academic achievement and personality characteristics could be improved and subsequent dropout rates reduced by having children start school prior to the officially sanctioned ages of five or six. Since any new program, or any significant modification of an existing program, can cost the citizenry a great deal of money, and since the history of the field of education has been characterized by a progression of panaceas offered by one expert or another, it is not too surprising that in this era of greater scientific and administrative sophistication, clamors should be heard demanding that educators demonstrate the value of any large-scale endeavor which will consume a significant portion of public monies.

Implicit in the concept of accountability is the notion that the merits of educational programs can be demonstrated objectively and convincingly. Such demonstrations rely on formal evaluations, evaluation being defined as the process of delineating, obtaining, and providing useful information for making decisions.

In the field of education, the handmaiden of evaluation has been testing—sometimes of individuals but far more often involving groups of children. Although testing is only one of many sources of information useful for evaluation, it is the method which has been most widely employed, perhaps primarily because of the magnitude of the evaluative task when millions of individuals are involved. For this reason, it is important to make clear that testing and evaluation are not synonymous. Moreover, as the number of children to be evaluated increase, intensity of the testing effort must, of necessity, be decreased. Because of the greater verbal and literary sophistication of older children and adults, this diminution of intensity need not necessarily complicate the testing process. However, because of the developmental characteristics of young children, the potential user of tests for educational evaluation needs to be even more alert to the kinds of decisions which can be made on the basis of testing and to the limitations of testing when young children serve as subjects.

For many educators, the world of testing is confusing and frustrating. It is easy to be overwhelmed by the profusion of forms that tests take, the variety of content they contain, the proliferation of uses to which they are put, and the diversity of conditions under which they are administered. Terms such as projective tests, paper-and-pencil tests, achievement tests, diagnostic tests, personality tests, psychometric tests, edumetric tests, true-false tests, multiple-choice tests, norm-referenced tests, criterion-referenced tests, standardized tests, and teacher-make tests confront the educator frequently and produce bewilderment about the nature of the testing process. This bewilderment is easily compounded when one is concerned with the value or the limitations of testing very young children. With young children, the external criteria against which the usefulness of testing can be evaluated are either missing or less easily recognized. Furthermore, it is much more difficult to determine with young children the extent to which test results are distorted by factors other than the abilities or traits one is attempting to measure. Thus, educators planning to develop testing programs for young children need to be

*The authors' work is supported in part by Grant No. SF-500 from the Office of Child Development and by a grant from the Carnegie Corporation.

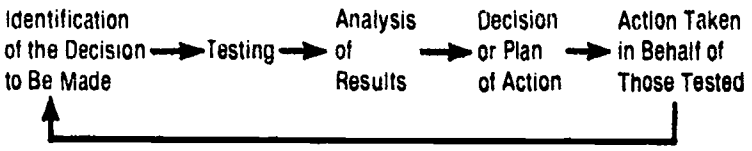
alert not only to the issues pertinent to testing with any age group but especially to the difficulties and limitations associated with testing of the very young.

All tests are micro-experiments (Caldwell, 1972) in that they sample skills or attitudes presumed to be representative of a larger repertoire of skills and attitudes hypothesized by the experimenter (tester) to be useful indicators of achievement or development. This is true in the case of a well organized and conceptualized collection of micro-experiments (test items) such as the Stanford-Binet Intelligence scales or in the case of a single micro-experiment test such as an attempt to determine if a child has attained the concept of conservation of mass. Tests are micro-experiments in which some capability, attitude, or behavior is examined to see if it is present or absent under a particular set of conditions. Such micro-experiments are administered for only one valid reason—the making of decisions about some course of action to be taken which in some way will involve the child who has been tested. Just as a medical test is given in order to help make a decision as to whether a certain form of medical treatment might be necessary or advisable, so are educational tests administered to help make educational decisions.

TESTING FOR DECISION-MAKING

The model below illustrates the typical manner in which tests are used as part of a decision-making process in a democratic system of education.

Model of Decision-Making in a Democratic System of Education



First, the educator identifies two particular decisions to be made. Does Angela Durkheim need special individual help due to a learning disability? Does the school district need more kindergartens? Second, a test is used to help make these decisions. In the case of the decision about Angela, a battery of tests aimed at diagnosing her specific learning disability and suggesting appropriate remediation might be administered. In the case of the decision about the additional kindergartens, each entering first grader in the district might be given a readiness test to determine the percentage of entering children identified as significantly below the expected achievement level. Third, the results of the tests are analyzed in relation to the decisions to be made. Angela has a rather severe perceptual deficit, and 40 percent of the students in the district are at least one year below entering first-grade level. Fourth, decisions are made about what to do. Angela should be given attention for her perceptual problem. New kindergartens are needed in areas of the city with a high proportion of children identified as not ready to begin first-grade work. Fifth, action is taken as a result of the decision. Angela is placed with a learning disabilities specialist who designs a special program for her perceptual problem, and the district opens four new kindergartens in schools with the highest percentage of poor achievers. Finally, after the passage of enough time to permit some results to appear, the entire cycle should begin again, with further decisions and testing. Has the remediation program helped Angela? Are the new kindergartens helping to improve scholastic readiness in entering first graders? These are new but related decisions which will require a second cycle of decision making.

These two rather oversimplified examples in no way exhaust the list of educational decisions that may fruitfully involve tests as part of the decision-making process. However, they do show the general manner in which tests should function in the making of educational decisions. They should seldom, if ever, be the only thing considered as these decisions are made, and they should not be given unless some clear benefit for children and the community can be anticipated as an outcome.

Types of Decisions

Tests provide useful information for making many different types of educational decisions. In a given decision situation, the choice of a test to be used and the utility of the one chosen will depend on the type of decision to be made. For this reason, we shall consider several different types of decisions relating to the education of young children.

Program Planning and Evaluation. A large percentage of the decisions made in educational settings involve decisions relating to planning and evaluation. Some of these decisions relate to individuals ("Is Tommy Harrison making the kind of progress expected in reading?" or "Should Clarence Bell go on to the next level of instruction in math?"). Other decisions relate more directly to programs ("Should the present language arts program be changed to better meet the needs of the children participating?" or "Is the present kindergarten program accomplishing its objectives?"). Bloom, Hastings, and Madaus (1971) have classified these types of decisions as formative and summative evaluation decisions.

The first major type of decision relating to program planning and evaluation is called formative evaluation. Once a child has entered an instructional program matched to his/her individual needs and capabilities, it is important to monitor the child's progress during the program and to check for errors in the program. These two functions are called formative evaluation and can be distinguished from the functions of diagnostic evaluation, such as classifying students and determining underlying causes of difficulty. Essentially, a formative evaluation is designed to allow an educator to decide whether modification is needed in instruction to enable children to develop in a satisfactory manner.

In general, the instrument that determines whether the child possesses certain prerequisite characteristics which make him/her a good candidate for a particular program is different from the instrument which locates where in the program a child is experiencing difficulty. In many instances, diagnostic tests are concerned with individual differences which predispose an individual to receive a certain kind of treatment. Formative tests, by comparison, are primarily concerned with changes within an individual in skills, attitudes, and so on, resulting from the instructional program.

Carver (1974) has introduced an important distinction about the function of tests which is helpful for the design and selection of formative tests. He says that a test may serve either a psychometric or an edumetric function. Psychometric tests, which have dominated the testing field for many years, focus on individual differences, whereas edumetric tests reflect individual growth from one time to another. Edumetric tests and formative evaluation are very similar concepts; when one is concerned with formative evaluation, one would be wise to choose an edumetric test. To be more specific, the test designer should choose items that are maximally sensitive to individual growth (edumetrically useful). The reason for choosing edumetrically efficient items is that education is a purposeful enterprise, one aimed at fostering individual development. Education should not be concerned merely with showing how individuals stand in relation to one another. Therefore, the best items for formative evaluation are not those that approximately 50 percent of the students "pass" and for which there is an approximately normal distribution. The best items are those which almost nobody passes prior to instruction and which, it is hoped, almost everybody will pass at the end of instruction (Carver).

The concepts of edumetric testing and formative evaluation are important because program planning should often attempt to determine whether some particular type of program shows the potential for meeting educational needs of children. Thus, if one were trying to plan a program to ameliorate learning disabilities in preschool children, one would need to assess the children's progress frequently and in fairly narrow segments of activity in order to estimate whether the educational program does appear to be accomplishing its objective. If not, then one would want to change certain aspects of the instructional program to try to get the program back on target.

The second major type of decision relating to program planning and evaluation is called summative evaluation. At the end of a program or a unit of instruction, educators are often called upon to make decisions related to the final performance of participants. Should the program be continued? What grade should an individual get? Should an individual be certified as having certain skills and knowledge? Should a participant enter the next level of the program? Did the program accomplish its objectives? What additional effects did the program have other than those intended? Such decisions are related to what Gronlund (1973) calls summative evaluation. In formative evaluation, the focus is on a test which measures improvement in all the specific skills to be achieved during the unit of instruction. In summative evaluation, the focus is more general—the educator wishes to obtain a representative sample of performance on program objectives and related variables.

A summative evaluation usually involves measuring level of mastery or attainment at the end of the total instructional period. Since individuals are likely to differ in terms of how much they have benefitted from a program, educators will often find it useful to include items which assess a fairly wide range of performance for each content area in a summative test. In this way, the educator can obtain a more accurate estimate of the program's effects on each individual. We might add that including items which have a range of difficulty is useful regardless of whether the individual's performance is compared to norms or to more specific performance criteria.

In thinking of evaluation, it might be useful to consider the reminder offered by Morgan (1971): In evaluation, the problem is determined by the situation. Many people may be involved in defining the evaluation, including those who participate in establishing the philosophy and objectives. Morgan also contends that the outcomes of an evaluation are not intended to be generalized to other activities. Findings apply only to the particular program evaluated. The goals and objectives of the program are used as criteria to determine the extent to which the purposes of the program have been achieved.

In the past, it has been common practice to use standardized achievement tests in summative evaluations. Tyler (1974) offers two major criticisms of this type of norm-referenced instrument when used for such a purpose. First, norm-referenced achievement tests include too small a sample of exercises appropriate for

appraising the learning of children who markedly deviate from the average. The tests include so few items at the level where most disadvantaged children are learning that changes in test scores due to improved capability cannot be distinguished from those due to chance variation in performance. Second, a teaching method or set of instructional materials is generally designed to improve learning but not to improve it spectacularly. A real improvement of 5 to 10 percent in student learning can often be considered a success. To measure such differences in the capabilities of students would require a precision of measurement that is rarely attained with a survey instrument such as a standardized achievement test. As a consequence, most achievement tests cannot provide information about the differential effectiveness of teaching methods or instructional materials unless the differences are large.

Screening. In day-care facilities, kindergartens, and elementary schools, considerable attention is devoted to screening for developmental difficulties. The screening decision in these settings is similar to the selection decision in industrial settings (see Cronbach [1969] for a discussion of selection decisions). Gallagher and Bradley (1972) point out that the essential decision is a dichotomous one—a child either goes into a program or he does not. Therefore, that instrument is most useful which does the best job of balancing errors in deciding whether a handicapping condition is present or absent for the individual child. To be more specific, a good screening instrument is one which has both a low percentage of false positives (those identified as having the condition who, in fact, do not have it) and a low percentage of false negatives (those identified as not having a condition who, in fact, do have it). The relative importance of the two types of errors will depend on several factors including the need for specialized programming, the opportunity for more intensive diagnosis, and the availability of programs for the condition.

Screening instruments for young children are likely to be concerned with identifying children who, if permitted to enroll in regular classrooms without some instructional adaptation geared to their particular disabilities, could be expected to experience frustration and failure rather than success. Examples are developmental delay or mental retardation, learning disabilities, or early behavioral dysfunction. The use of screening in a particular educational setting does not mean that the children are going to be identified and separated and in any way stigmatized. With the advent of mainstreaming (enrolling children with special learning needs in regular classrooms with other children), screening programs are more likely to be used to alert teachers to the special needs of some of the children and to encourage genuine individualization in classroom instruction.

Diagnosis. In many educational systems, it is standard procedure to follow screening tests with much more intensive diagnostic testing. This more careful examination of the individual leads to a decision which differs from the decision made in screening. That is, the aim of diagnosis is not to determine whether or not a person should be placed in a program, but what specific type of instruction or treatment he/she needs. To reiterate, the purpose of screening testing is to determine whether a treatment is needed; the purpose of diagnostic testing is to determine what kind of treatment is needed. Diagnostic testing, then, leads to a classification decision (see Cronbach [1969] for a discussion of this issue), not a selection decision. This distinction is important because a test which does an excellent job of identifying a child with a reading problem may be virtually useless in pointing out what kind of remedial program that child needs.

Since a diagnostic test is used as a basis for choosing between different programs or treatments, the differential validity of the test must be demonstrated. It must discriminate, for instance, between those whose reading difficulties stem from a visual perception problem and those whose reading difficulties stem from poor word-attack skills and must, therefore, imply appropriate directions for remediation of the specific pattern of disabilities identified.

Administrative Decisions. Hayman (1974) states that improvement in education clearly depends on an improved management practice. Improved management, in turn, depends on more effective information systems. Hayman goes on to say that the management decision process is quite complex even in a system of moderate size. Several levels of management, each with its own level of decision making, are always in existence. Moreover, since the needs of administrators at each level differ, there is often a real problem of articulation among information systems at the different levels. Because of this complexity, Hayman believes that most information systems will have to serve a single operational level. Thus, when choosing a test as a means of providing information for a specific administrative decision, it is probably best to select one that provides the kind of information useful at the particular administrative level where the decision is being made rather than a test which provides a more general kind of information.

Unlike other educational decisions, administrative ones are not primarily made with respect for a specific individual. Administrative decisions concern whether a new math program is needed, not whether Johnny Jones is showing adequate achievement within his present program. To put it another way, administrative decisions do not focus so much on the teaching-learning process per se as on the conduct of that process.

Since administrative decisions are made without ref-

erence to individual children, they can often be made without testing children at all. Instead, information may be gathered about other aspects of the educational situation (classroom climate, educational level of a teacher, and so forth) of the larger community situation. Even when individual children are assessed, however, the type of test chosen will be determined by how useful the information is in making that decision.

Given the broad range of educational decisions which might employ tests as a means of gathering information for that decision, very few generalizations about the types of tests needed to make educational administrative decisions can be made. Reliability and validity are certainly important for all tests. In addition to these considerations, the educational administrator may desire that a test provide information that is neither too broad nor too narrow in scope. He might also be concerned that a test provide information that is understandable to the individuals who must operate on the basis of that information. Finally, he might be concerned that the information provided by a test articulate with other sources of information used in making a particular decision.

CHARACTERISTICS OF GOOD TESTS

After considering the type of decision to be made as a factor in the selection of a test for young children, educators should consider several additional test characteristics before making a final selection.

Practical Criteria

Stufflebeam and his colleagues (1971) provide several useful criteria for judging the practical worth of evaluation data in making decisions. They include the following six criteria:

Relevance. Test data are collected to meet certain purposes and if they do not relate to those purposes, they are useless. The criterion of relevance asks whether or not the purposes are, in fact, served.

Importance. A great deal of information can be collected which is nominally relevant to some purpose; but, obviously, not all information is equally important.

Scope. Information may be relevant and important but lack sufficient scope to be useful.

Credibility. Credibility relates to the quantity of trust or belief one can have in the data. Not all users of test information are in a position to determine its validity, reliability, or objectivity. They need to be able to trust the information provided by the test.

Timeliness. The best test information is useless if it comes too late or too soon.

Efficiency. After test data have been examined for adequacy in meeting the practical criteria listed above, there are probably still alternatives which differ in terms of such requirements as time, cost, and personnel. The criteria of efficiency can be employed as a means of guiding the educator to the appropriate alternative.

Validity

When choosing a test for use in some educational setting, primary emphasis is placed on choosing an instrument that is valid. Broadly defined, validity means the extent to which a test measures what it purports to measure; and it is a necessary characteristic for all tests to have. Anderson, Ball, Murphy, and their associates (1975) have noted that measurement literature is replete with adjectives associated with various conceptions of validity. Several of the different types of validity will be discussed in the present paper, and their proper usage in educational decision making will be highlighted. In this context, it is worth remembering that all types of validity do not equally apply to all types of decisions involving tests.

Construct Validity. Construct validity may be defined as the degree to which scores on a measure permit inferences about underlying traits. For example, a teacher might ask his/her students to answer a series of questions about school in an effort to find out what their attitude toward school is. Very generally, the construct validity of a test instrument is determined by scientifically examining the relationship between test scores (measure of attitude toward school) and certain hypothetically related variables (attendance at school, behavior at school, grades). This type of validity is most important in certain diagnostic and summative tests.

Bloom, Hastings, and Madaus state that construct validity is a characteristic most frequently attributed to ability and personality tests. However, construct validation creates numerous problems when applied to measures in the affective domain (see discussion of affective domain below) since affective traits are often impossible to measure directly. As the conceptual foundation of many affective measures and some cognitive measures (see discussion of cognitive domain below) is shaky, many of these tests have been developed in an almost circular fashion. Construct validity is particularly hard to establish for instruments designed to measure young children.

It is doubtful that scores obtained from different populations and environmental settings have the same meaning even when the same testing procedure is used.

Discriminant Validity. Discriminant validity involves the accuracy with which an instrument can discriminate among different groups of individuals, such as those who can detect the difference between certain spoken sounds and those who cannot. This type of validity is imperative for diagnostic or classification type tests which must be able to indicate which individuals will benefit from various types of treatment or instructional programs.

Criterion Validity. Criterion validity involves comparing the results of a test to an outside criterion which is thought to be related to the test. For example, one might compare scores on a diagnostic reading test to performance in a reading program. As one might expect, criterion validity is principally applied to tests used for selection and for classification purposes.

There are two general types of criterion-related validity: concurrent and predictive. The concurrent validity of a test is established by correlating scores on the test with performance on some outside criterion measured at approximately the same point in time. Bloom, Hastings, and Madaus suggest that the notion of concurrent validity can be of use when one is inquiring into the relation between an indirect and a more direct measure of some behavior. For example, it is often critical to obtain concurrent validity data for affective measures when the variable being measured is an abstract trait or characteristic such as self-concept. By comparison, concurrent validity data is rarely needed for psychomotor measures since most psychomotor capabilities are readily observable. Bloom and his colleagues warn that it is not very fruitful to try to establish the concurrent validity of achievement tests by correlating them with things such as intelligence tests. It is not fruitful primarily because it is difficult to decide which test one is validating if both consist of samples of items from the same table of specifications.

Predictive validity is determined in the same manner as concurrent validity except that performance on the outside criterion is measured after—sometimes long after—scores on the test of interest have been obtained. This type of criterion validity is important in tests used for selection and classification purposes where performance on the test needs to be related to subsequent performance in a program. Bloom, Hastings, and Madaus claim that predictive validity is also important for summative tests if those tests are used for prediction purposes (such as predicting success in the next level of instruction).

Ebel (1972) makes note of the fact that it is often difficult to find a criterion which is adequate for judging the validity of a test. He states that efforts are sometimes made to determine the criterion validity of educational achievement tests by correlating the test scores of pupils with their grades or ratings from teachers. The problem with this procedure is that teachers' ratings are notoriously unreliable, particularly in the case of very young children. Furthermore, grades and ratings often reflect pupil achievement that the test either could not or was not intended to measure. Hence, for many achievement tests, the most important type of validation is direct validation (validity based on the test makers' judgment that an item is testing the desired concept at the desired level of learning).

Content Validity. According to Bloom and his colleagues, content validity generally refers to the correspondence between test items and curriculum. That is, the items in the test should cover what was done in the program. Content validity is important for both formative and summative types of evaluation, especially when one is assessing skill or knowledge. Formative tests should include all the important elements in the unit as detailed by the table of specifications. By comparison, summative tests need only include a sample of the range of contents and behaviors outlined in the table of specifications.

Ebel notes that the content validity a test possesses cannot be determined by looking at the scores the test yields. Instead, one must look at the test itself, at its rationale and specifications, and at the directions for administering and scoring it. The content validity of many tests of psychomotor and cognitive learnings in young children is rather easy to establish since the performance criteria for these learnings are relatively limited and unambiguous. For example, it is relatively simple to determine if a child can walk unassisted along a balance beam or classify objects. Nevertheless, some educators are misled into believing that a test is a valid measure of certain competencies simply because it contains appropriate content.

Tyler mentions that those who use achievement tests to assess the effects of school efforts have been criticized on the grounds that achievement tests rarely reflect the particular objectives of an education program. That is, nationally standardized achievement tests quite often are not content valid for the purpose of assessing the effects of a particular educational program. Tyler also contends that standardized achievement tests seldom contain appropriate content for formative and summative evaluations because of the manner in which items are selected for inclusion in such instruments. To be specific, the items found in most standardized, norm-referenced measures involve capabilities which are not usually the focus of instruction. More likely they involve capabilities learned incidentally in the normal run of a child's life

experiences since children are more likely to differ markedly on those capabilities learned piecemeal through life than on those capabilities systematically taught in school.

Edumetric Validity. Edumetric validity is important for testing instruments designed to measure individual growth. This type of validity usually cannot be determined by administering the test to one group of subjects at one point in time. Carver states that "ordinarily, the test must be administered in two situations or conditions wherein gain or growth is expected. Degree of actual gain may be compared to the degree of expected gain to estimate the degree of edumetric validity" (p. 514). Expected gain is difficult to estimate in preschool children unless one relies totally on developmental norms. Moreover, if one is testing in the affective domain, as in a study of aggressive behavior, norms of expected behavior are largely lacking. This limits the extent to which edumetric validity can easily be determined in tests for young children.

Validity is essentially a matter of degree (Ebel). Tests are not valid or invalid; they are more valid or less valid. Furthermore, the problem of interpreting whether a test is valid or not lies mainly with the user of the test (Cronbach). Only the user can judge whether the conditions necessary for validity have been maintained when the test was administered. Only the user will know the specified purpose for which the test is being used.

Reliability

A reliable measure is one that provides consistent and stable indications of the characteristic under examination. Reliability, in this sense, is critical for all tests, whether edumetric or psychometric.

As Carver defines it, psychometric reliability means that a test can be expected to consistently discriminate between individuals from one occasion to the next. There are two major types of reliability estimates used for psychometric instruments. The first is test-retest reliability. Bloom, Hastings, and Madaus contend that when very little time intervenes between two test periods, test-retest reliability reflects the consistency with which a testing procedure measures. This type of reliability is important for psychometric instruments; but its usefulness is restricted to those retest situations in which memory is not a great factor. When memory is a problem (such as with many cognitive tasks), one can often remedy the problem by using the alternate-forms method of estimating reliability. To be more specific, one can devise two tests, each composed of different items, to measure the same area of content. The tests can be administered to the same group within a short space of time. The correlation between the two sets of scores can then be used as the estimate of reliability. When the time interval between tests is long, test-retest reliability measures stability of a trait or ability in the individual. As one might expect, measures of young children often do not have—nor should they have—high reliability in this sense because the traits and characteristics themselves are not stable over time.

A second general type of psychometric reliability is internal consistency. Internal consistency coefficients (such as the Kuder-Richardson formulae or the Cronbach alpha) indicate how homogeneous the item content of a test is. Put another way, the internal consistency of a test is how consistently the items in the test measure the same trait or capability. Internal consistency estimates are useful when the tester is using a group of test items together (as in a scale) to measure a single area of cognitive content or a unidimensional affective trait. For broadly focused tests (such as many summative instruments), internal consistency is of relatively minor importance. The tester should remember that internal consistency coefficients are strongly influenced by the number of items in the scale. Therefore, when the scale is short, the obtained reliability coefficient may underestimate the test's true reliability.

Like psychometric tests, edumetric tests must be reliable if they are to be valid. Little attention has been given to developing reliability indices for edumetric instruments. Carver suggests that reliability for this type of test be defined as the consistency of gain or growth within individuals as reflected by the test. The reliability of edumetric instruments may be estimated by administering alternate forms of a test in both pre- and posttreatment conditions. The constancy of change scores between forms would provide a good indication of reliability. This type of reliability is important for most formative tests; but it is difficult to achieve where growth is hardest to measure as with many affective variables.

Test Bias

There are many different forms of bias which prevent a test from being valid when used with a particular individual in a particular instance. Test bias in its many forms can be categorized as emanating from three major sources: bias due to the situation in which the test is administered, bias due to the test itself, and bias due to the function of some characteristic of the person being tested. Some examples of bias due to external conditions are noise, unpleasant or strange surroundings, a tester who is unable to establish rapport with a child, and failure to make the child understand the nature of the task to be accomplished. Some examples of bias resulting from the child himself would include fatigue, boredom, and a response style incommensurate with what is needed in the test. Bias can result from the test itself not

only in terms of the normal sources of invalidity, but also from such things as a test filled with irrelevant difficulties, the need for test-wiseness, social-desirability effects, and unclear directions. Even when the validity of a given test has been established in terms of meeting the purpose for which it was designed, one must consider these additional sources of invalidity before administering such a test in a particular situation.

SELECTING APPROPRIATE TESTS

Selecting appropriate instruments to use in making educational decisions relating to young children is no simple task even for the experienced tester. Test titles and even test content are often deceiving. They are deceiving primarily because tests frequently are not direct measures of the characteristic in question. Instead, tests are experiments through which one makes inferences about an individual based on his/her performance.

In certain respects, trying to judge what a test measures is like trying to judge the shape of an object in a hall of mirrors. What the eye sees it may well see clearly; but clarity of perception in no way guarantees that the reflected image accurately represents the real object. It is knowing the shape of the actual object (such as one's body) that makes a hall of mirrors such fun. The tester, however, seldom knows the dimensions of what he is trying to measure. He must depend exclusively on what the test reveals. In a hall of mirrors, accurate estimation of the shape, size, and other dimensions of an object comes from knowledge of the distortion properties of the mirror. That is, the viewer must know if the mirror makes objects seem smaller, larger, more top-heavy, more rounded, at an angle, and so on, in order to have a correct concept of the object's shape.

The ideal situation, of course, would be to find a mirror with no distortion. In the absence of such a mirror, one would probably do better to select a mirror whose distortion properties he understands than one he thinks may have less distortion. It is in knowing precisely how the mirror distorts that one can "correct" the image one sees. For example, suppose we know that a particular mirror makes everything in the top half seem one third wider than it actually is and everything in the bottom half seem one-fourth narrower. We could, then, on the basis of that knowledge, draw a relatively accurate picture of a person if we saw only his image in this mirror whose properties we understood. Suppose, on the other hand, we looked in another mirror that had no distortion in the bottom four-fifths and a distortion of only one-tenth in the top one-fifth. Without knowledge of these distortion properties we could use these more accurate actual images and still draw a bunch of overly fat-headed people.

The available repertoire of educational tests has been classified by Bloom (1956) as falling essentially in three major psychological domains—the cognitive, the affective, and the psychomotor. These domains are not as neatly separated in the behavior of the young child as they are presumed to be in the older child or adult. For example, when a young child grabs a toy from another in the nursery school, can we conclusively infer that he is acting aggressively (affective domain) rather than attempting to solve a problem (cognitive domain)? The younger the child, the more a response in a testing situation is likely to involve components of all domains and the more the examiner will need to rely on cues from more than one domain to know whether a child has succeeded on a given item. Let us suppose one is testing a nine-month-old infant for object permanence, an extremely important psychological dimension within Piagetian theory. The examiner covers an attractive toy with a handkerchief and waits for the baby's response. If successful, the baby is likely to look puzzled or bewildered (cognitive), move away the handkerchief (psychomotor), smile with delight (affective) upon again seeing the hidden toy, and enthusiastically pick it up (psychomotor domain again). Essentially, all of these component responses are necessary to let the examiner know that the nonverbal child has indeed "succeeded" on this test item.

Although there is not complete agreement on the best labels for these major domains of human behavior, we feel that the Bloom taxonomy is useful for illustrating how one can determine what kinds of test characteristics are important in assessing for a particular educational decision.

Cognitive Domain

The cognitive domain as described by Bloom, Hastings and Madaus, is composed of six levels arranged in a hierarchy according to the complexity of the learning involved. As Bloom has noted, most of the instruction at the preschool and early elementary levels is focused on the first three levels within the cognitive domain. Therefore, we have provided some illustrations of test items for these levels of the hierarchy.

(1) *Knowledge.* Knowledge involves the recall of specifics and universals, the recall of methods and processes, or

the recall of a pattern, structure, or setting. A typical knowledge item for young children might be "What color is a banana?" or "Tell me what you do with a fork." For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material.

(2) *Comprehension.* Comprehension represents the lowest level of understanding—knowing what is being communicated and being able to make use of material or an idea without necessarily relating it to other material or seeing its fullest applications. An example of a comprehension item appropriate for young children might be "Which of these men is the tallest?" or "Put all the beads in the box."

(3) *Application.* The use of abstraction in particular and concrete situations. The abstractions may be in the form of general ideas, rules or procedures, or generalized methods. The abstractions may also be technical principles, ideas, and theories which must be remembered and applied. An example of an item appropriate for young children which examines application of principles or ideas might be the following:

This year we studied how to add and subtract.

Which would you have to do in order to:

(a) keep score in a ball game.

(b) figure out how much you would have to pay for three records.

(c) figure out how much change you should get when buying groceries.

(d) figure out how old you will be when you graduate from high school.

(4) *Analysis.* The breakdown of a communication into its constituent elements or parts so that the relative hierarchy of ideas is made clear and/or the relationships between the ideas expressed are made explicit. Such analyses are intended to clarify the communication, to indicate how it is organized, and to show the way in which it manages to convey its effects as well as its basis and arrangements.

(5) *Synthesis.* Putting together of elements and parts so as to form a whole. This involves the process of working with pieces, parts, elements, and so forth, and arranging and combining them in such a way as to constitute a pattern or structure not clearly there before.

(6) *Evaluation.* Judgments about the value of material and methods for given purposes, quantitative and qualitative judgments about the extent to which material and method satisfy criteria, use of a standard of appraisal. The criteria may be those determined by the child or those which are given to him.

Affective Domain

In developing the taxonomy of educational objectives, Bloom did not have in mind the separation of cognitive and affective capabilities. Indeed, the research indicates that these two domains both develop and manifest themselves in concert. Bloom argues that there is need to develop good measures in the affective domain so that educators can improve the effectiveness of programs in forming affective behaviors. He further contends that affective aspects of the curriculum will continue to be ignored until we give attention to evaluating its outcomes. Bloom states that "It is often desirable to evaluate a student's affective behavior formatively. Such an evaluation is diagnostic in that it can indicate to the student his progress toward the attainment of such outcomes; it can be educational, for example, when he is given a profile of his academic and vocational interest patterns. The point is, however, that feedback to the student, not the assignment of a grade, should be the purpose of making a formative evaluation of affective objectives." Krathwohl, Bloom, and Masia (1964) divide the affective domain into five ascending levels: For children younger than nine years of age, attention will be devoted primarily to the receiving and responding levels within the affective domain. Somewhat less attention will be paid to the valuing level, and almost no attention will be given to the organizational and characterizational levels. According to the developmental theory described by Piaget and Kohlberg, the final two levels within the affective domain would require a person who had developed formal operational thinking.

(1) *Receiving.* This category is defined as sensitivity to the existence of certain phenomena and stimuli—that is, the willingness to receive or attend to them. An example of a receiving objective appropriate for young children might be "Child is able to listen to teacher read an entire story."

(2) *Responding.* Responding refers to a behavior which goes beyond merely attending to the phenomena; it implies active attending, doing something with or about the phenomena, and not merely perceiving them. A possible objective at the responding level is "Child requests that adults read to him."

(3) *Valuing.* Behavior which belongs to this level of taxonomy goes beyond really doing something about certain phenomena. It implies perceiving them as having worth and consequently revealing consistency in behavior related to these phenomena. "Child urges other

children to read" is an example of an objective at the valuing level which might be appropriate for young children.

(4) *Organization.* Organization is defined as the conceptualization of values and the employment of these concepts for determining the interrelationships among values.

(5) *Characterization.* The organization of values, beliefs, and attitudes to an internally consistent system is called characterization. This goes beyond merely determining interrelationships among various values; it implies organization into a total philosophy or world view.

Bloom contends that it is more difficult to assess affective objectives than cognitive objectives. A large number of cognitive objectives can be assessed quite easily with traditional paper-and-pencil achievement tests. However, many forced-choice, multiple-choice, and other paper-and-pencil type instruments are of dubious validity when used to assess affective objectives. It is likely that a pupil taking such an exam would be concerned about how he will be evaluated by others (albeit this is less of a problem with young children than with older ones). Therefore, faking is probably less of a problem with young children; but it can never be completely ruled out as a possible determinant of response patterns. An additional problem in assessing affective outcome is that many human dispositions, attitudes and so forth are not as stable in young children as are cognitive competencies. Therefore, the test-retest reliability in many of the instruments is often low. This is true particularly if a relatively long time period intervenes between assessments.

There are several methods one might consider employing in order to get a more valid assessment of affective behavior. The first would be systematic observation in either natural settings or certain types of structured simulated settings. Also available are things such as interview techniques, open-ended question techniques, closed-item question techniques, and projective techniques.

The Psychomotor Domain

The psychomotor domain deals with observable voluntary human movement. These voluntary movements require use of the muscles, nerves, propriocepters, and the central nervous system. A taxonomy of educational objectives for the psychomotor domain was developed by Harrow (1972). Like the other two domains, this taxonomy is arranged in a hierarchy. Harrow warns, however, that in some instances there will also be continua existing within a particular classification level. There are six major levels within the psychomotor domain:

(1) *Reflex Movements.* Reflex movements or actions are elicited in response to some stimulus without conscious volition on the part of the learner. They are not voluntary movements but they may be considered as an essential base for movement behavior. One such reflex movement is the grasp reflex.

(2) *Basic Fundamental Movements.* Basic fundamental patterns occur in the learner during his first year of life. He builds upon the reflex movements inherent in his body. Common basic movements include such things as visually tracking an object, reaching, grasping, and manipulating an object with the hands, and progress through the developmental stages of crawling, creeping, and walking. The movements included in this classification level are those inherent motor patterns which are based on the reflex movements of the learner and which emerge without training. These movement patterns serve as the starting point for further and permanent perceptual and physical abilities and are essential to the development of skilled movement.

(3) *Perceptual Abilities.* Perceptual abilities assist the learner in interpreting stimuli, thus enabling him to make necessary adjustments to his environment. Perceptual abilities include such things as aesthetic discrimination, visual discrimination, auditory discrimination, tactile discrimination, and coordinated abilities.

(4) *Physical Ability.* Proper functioning of the various systems of the body enables the learner to meet the demands placed upon him by his environment. The physical abilities are, in fact, an essential part of the development of skilled movements. Physical ability includes such things as endurance, flexibility, agility, reaction response time, and dexterity.

(5) *Skilled Movements.* Skilled movements are the result of the acquisition of a degree of efficiency when performing a complex movement task. This classification level includes movements which require learning and are considered reasonably complex. Activities included in this classification level are those which involve some adaptation of the inherent movement patterns listed under Basic-Fundamental Movements. All sport skills, dance skills, recreational skills, and manipulative skills fall into this classification.

(6) *Non-Discursive Communication.* Non-discursive communication involves forms of movement behavior encompassing a wide variety of communicative movements ranging from facial expressions, postures, and gestures to sophisticated modern dance choreographies.

ELIMINATING SOURCES OF INVALIDITY

Educators can assess the validity of many tests using criteria such as those set down for internal validity (Campbell and Stanley, 1963) and external validity (Bracht and Glass, 1968). In a recent article, Snow (1974) discusses the external constraints on the validity of an experiment. He pointed out the need for experiments to be representative of the situations in which the findings of the experiment were to be applied. Representativeness makes the experiment ecologically valid. Representativeness is particularly important in educational testing since educators are usually more concerned that a child be able to demonstrate a competency in a variety of real-life situations than in some limited, often artificial, school-related context. On the following pages are some strategies which offer effective ways to obtain more ecologically valid assessment data:

Systematic Observations. Instead of always giving children a series of test items at the end of a unit or program, the teacher might be wise to systematically observe performance at various check points as learning progresses. It is often easy for the teacher to build into an activity a systematic procedure for observing the child's performance. Such on-the-spot observations are much less likely than cumulative exams to introduce irrelevant sources of difficulty for the child. They are also less likely to have some of the negative side effects associated with cumulative testing. Allen, Rieke, Dmitriev, and Hayden (1972) demonstrated how children's competencies in cognition, communication, and social and physical development can be systematically observed using a check list while the child is engaged in daily educational activities. They note that many of the recordings can be simple counts of particular behaviors or notations about how long children engage in certain activities. In addition to being an ecologically valid method of assessment, systematic observation of children has positive side effects for teachers. The teacher who is a good observer of children is more likely to be an effective teacher than one who is not and is likely to feel more confident that he/she can intervene in appropriate ways in the teaching-learning process. However, Allen and her colleagues warn that a teacher does not become a good observer of child behavior without some training and experience and that educators must constantly maintain caution when interpreting data gathered through observation. Behavioral data, like poems, are easy to misinterpret if one is inexperienced.

Systematic Replications. A second means of securing more ecologically representative assessments of children's competencies is by systematic replication of the testing situation. When a competency is assessed at only one point in time, it is likely that performance was influenced by some temporary motivational state. The performance of very young children tends to vary somewhat from one point in time to another. Thus, repeated assessment is more likely to provide a truly representative portrait of the child's capabilities.

Consonance of Test and Pattern of Experience. A source of ecological invalidity is a testing situation which is at odds with a student's previous experiences or preferred mode of problem solving. Hertzig, Birch, Thomas, and Mendez (1968) found dramatic differences between American middle class and Puerto Rican working class children in the behavioral styles with which they respond to cognitive demands. These differences held up even when the IQ levels of the two groups were comparable. Messick and Anderson (1970) state that "any qualities can be assessed in a context that is compatible with the student's previous experiences and thus does not introduce the irrelevant difficulty of 'strangeness.' This strangeness or the perceived irrelevance of the test to the life experiences of the examinee represents a kind of face invalidity, if you will, which poses a constant potential threat to the psychometric validity of the assessment and individual instances" (p. 24). Much of the recent concern relating to the bias of standardized tests can be related to the fact that they are not truly representative of the past experiences of individuals who take the test or the context in which most of those people gain their experiences. Many of the problems associated with this type of ecological validity of tests can be overcome by: (1) astute observation of children displaying certain competencies in a natural setting, or (2) imaginative construction of test situations in a context which is more natural and comfortable for the child.

Student Preparation. A fourth issue related to the representativeness of tests in preparation of the student. An educator is more likely to obtain an ecologically valid assessment of a child when the child is engaged in a normal classroom activity and especially when the child has had the opportunity to practice the activity for a while. Disruptions in performance due to novelty are less likely to occur once the child has become "tuned" to the task (Snow, 1974). Making a test ecologically valid in this way usually presents no problem to the teacher since teachers typically have children repeat activities as part of the regular teaching process.

Other Testing Methods

There are several other methods of assessing children which offer a potentially more valid measure of

children's competencies than do typical testing procedures. One such method is the micro-experiment. Caldwell (1972) has stated that each test item from an infant development scale is itself a test. That is, it is an experiment whereby one can judge whether a child can or cannot do a particular thing. A single test item by itself can often be more ecologically valid than the same item in the context of an entire scale. The main reason for this is that a single test item is not as likely to be encumbered with irrelevant difficulty as is an entire scale of items.

Messick and Anderson discuss four sources of irrelevant difficulty which lead to bias in tests. One such source cited by the authors is test format. Suppose, for example, that a test requires the child to read each item. Reading itself may be an irrelevant difficulty. When a child is confronted by a single item, it is much easier for the teacher to overcome this kind of difficulty. A second source of irrelevant difficulty is that items are sometimes more germane to one group than to another. When items are given individually, the teacher or educator can be selective in terms of giving some items to certain groups and some items to others. Testing conditions can also make some individuals feel anxious, threatened, or alienated. When test items are given by themselves, the test will seem more like everyday problems which generally occur singly rather than in groups. Thus, such a test is less likely to cause anxiety in children. Similarly, "test wiseness," a fourth source, would appear to be less of a factor of irrelevant difficulty with single items since the problem to be solved in an individual item is more like the kind of normal problem-solving situation that every child (regardless of experience) encounters. Many such micro-experiment tests are particularly useful in the investigation of conservation skills many of which can be assessed in gamelike situations that are more like the real world than the usual classroom testing situation.

The performance test is another alternative to typical testing procedures used in educational settings. In such tests, the individual must demonstrate a competency rather than answer questions about it. Performance tests have been used for many years, particularly to examine adults' ability to do a particular job after training. They have also been used to assess adolescents' skills in school settings. The performance measure also offers a potentially useful method of assessing the capability of young children, although it has not been used often for this purpose. Performance tests have the advantage of being more ecologically valid as a means of assessment than typical testing procedures. Boyd and Shimberg (1971) discuss the advantages of using performance tests as measures of classroom evaluation. They also discuss the means through which one can develop a performance test. To quote them, "Most of us recognize that there is a fundamental difference between knowing about a job and being able to do the job. Knowledge of a job is really an essential ingredient for doing a complex job correctly, but while it is a necessary condition, it is rarely a sufficient condition for doing a satisfactory performance. . . A person may be able to bluff on a written test, but he can seldom carry off a successful deception when a realistic performance test is required. One of the great virtues of the performance test is its impressive face validity and credibility, because the task one must do so closely resembles the job itself" (p. 3).

A performance test, conducted in an actual or simulated real-world setting, can be particularly useful for assessing certain kinds of capabilities in children. If one is interested in examining communication skills in children, for example, one might better assess this by placing them in a situation where they have to demonstrate their communication skills than by testing their usage of grammar in a written examination. The same is true with mathematics. Setting up situations in which one has to compare prices on food, compare percentages of nutritional ingredients in food, make correct change in monetary transactions, judge whether one is being hoodwinked in monetary transactions, and so forth, may be an excellent means of assessing certain mathematical capabilities in young children. Similarly, the kind of math proficiencies demonstrated in the playing of certain children's games may also be excellent indices of mathematical ability. Boyd and Shimberg (1971) state that "Whether one starts with a job analysis or with behavioral objectives which are originally derived from such an analysis, one must decide which elements are crucial to success. It is from these critical elements that one should select the task to be used as a measure of performance. Because performance is generally a slow and time-consuming process, only a few of the critical elements can be included. One must decide which ones are really crucial" (p. 5).

GENERAL CONSIDERATIONS

Human Rights

In considering the proper function of testing in education, one quickly realizes how often tests have been used improperly. One of the most lucid discussions of the kinds of abuses which occur when tests are not used properly appears in a recent article by Mercer (1974). She lists five rights of children that are frequently violated by present assessment and educational practices. In brief, those rights are as follows:

(1) *The Right To Be Assessed as a Multi-Dimensional Human Being.* Mercer found that many children are classified as retarded on the basis of IQ tests alone, even though IQ tests assess only a fraction of the total number of human skills and capabilities. For some children, retardation is school-specific. Their competencies in other situations are quite adequate.

(2) *The Right To Be Fully Educated.* For many children, educational assessment is a prelude to being stuck in a track or class which in no way meets their individual educational needs.

(3) *The Right To Be Free of Stigmatizing Labels.* The problems associated with being labeled as retarded, speech-impaired, learning-disabled, and so forth are clearly documented. Mercer argues that the learning environment must be carefully structured so as to avoid the negative stereotyping which accompanies such labels.

(4) *The Right To Ethnic Identity and Respect.* As a general rule, standardized intelligence tests and achievement tests assess those competencies valued in Anglocentric societies while ignoring many of the competencies valued in other cultures. Such tests often provide "a mechanism for blaming children and their families when the educational program of the school fails" (Mercer, 1974, p. 137).

(5) *The Right To Be Evaluated within a Culturally Appropriate Normative Framework.* Many assessment procedures employ a single normative framework for interpreting the scores of all children. Those tests ignore the fact that the experiences of many minority group children differ widely from those of the dominant cultural group. Thus, for minority group students, comparison to the norm can be misleading.

To this list we would add at least one additional right that is especially applicable to young children—the right not to be permanently labeled on the basis of testing done during the preschool period. Present assessment techniques may be even more hazardous with very young children than they are with slightly older children and adults. Problems of reliability and validity (all types) are much greater with the very young child. If a child is growing and changing, do we want a test that produces identical results across time, one that has high test-retest reliability? As fatigue builds up quickly and militates against a long testing procedure, should we expect high split-half reliability? As large representative groups of young children have seldom been assembled, as can be the case with public school samples of older children, norm-referenced tests seldom have norms based on statistically acceptable samples. Criterion validity is difficult to determine, as widely accepted outside criteria can seldom be identified. For example, suppose one wanted to develop a test to measure emotional disturbance in preschool children. Presumably one might use psychiatric diagnosis as an acceptable outside criterion against which to validate the test. But most psychiatrists would feel doubtful about identifying those children whose behavior might justifiably be labeled emotionally disturbed because immaturity is an outstanding characteristic of such behavior, and all very young children are, by definition, immature.

These hazards do not mean that one has no right to apply tests to young children; they do suggest, however, that the person who uses such test data for educational decision making must anticipate quick obsolescence for any test data obtained on young children. That is, a cardinal rule for the testing of young children should be frequent retesting, in that fairly large changes in behavior can be expected in short periods of time. No child should be labeled permanently on the basis of test data obtained during the early childhood period.

Many educators have voiced concern over the abuses made of standardized testing by school personnel. Their concern points up a more general issue regarding the use of any kind of information in educational settings. As educators we must ask ourselves: What am I obliged to do in terms of gathering and disseminating school-relevant information? In a democratic system of education, the one who orders that information be obtained is the agent chiefly responsible for the information. Take the case of a school administrator. If an administrator wants a test to be given to a group of students by their teacher, he/she must make provisions for informing the teachers of the correct ways of obtaining and interpreting the information from the test. The information must then be used exclusively for the benefit of the child and the community. One interesting recent legal ruling aimed at improving the use of information in educational institutions is the Family Educational Rights and Privacy Act of 1974. This law guarantees the parents' right of access to their children's school records. The courts have deter-

mined that parents not only have the right to know what is in the record but also the right to challenge the veracity and appropriateness of the information it contains. This landmark legislation may prevent a repetition of many of the abuses associated with testing and other school-related information.

Achievement Tests

Mercer's insistence that educators test in such a manner that the rights of children are protected offers one more strong argument in favor of deemphasizing the use of standardized, norm-referenced tests when making educational decisions. As discussed previously, Tyler contends that standardized achievement measures are seldom adequate as summative tests, especially when disadvantaged children are involved; and Carver convincingly demonstrates the weaknesses of many norm-referenced measures when used to assess improvement in capabilities. This is not to say that standardized measures should be abandoned. On the contrary, they are often a wise choice because of their meticulous construction and convenience. What is important is that educators be judicious in employing such instruments, with an eye toward their appropriateness for making a particular decision.

One additional suggestion we might offer to educators is that they not rely too heavily on a single test score when making decisions. A battery of tests often provides a broader, more precise base of information. Unfortunately, although much lip service has been paid to the wisdom of using multiple measuring instruments, little training has been given on how to use them and even less data collected on their actual effectiveness.

POSTSCRIPT

At the beginning of this paper, we indicated that accountability has become a watchword in education. Educators are more mindful of the need to demonstrate how educational programs are making good their claims. Testing, probably because of its long and rather successful history in education, has been the chief evidence-gathering vehicle for establishing accountability. This heavy dependence on the use of test information is currently being questioned both by those within the field of education and by those without. Complaints are being made that tests may be inadequate as a basis for making many educational decisions. It is perhaps fair to say that testing has entered its own Age of Accountability—educators can no longer be arbitrary when using tests. The person tested or his family may offer a rebuttal to what the tests have found. In the still largely mysterious world of young children, it is even more essential that educators heed these warnings.

REFERENCES

Allen. K.E., Rieke, J., Dmitriev, V., & Hayden, Alice H. Early warning: observation as a tool for recognizing potential handicaps in young children. *Educational Horizons*. 1972. Vol. 50. No. 2, 43-54.

Anderson. Scarvia B., Ball, S., Murphy, Richard T. *Encyclopedia of educational evaluation*. San Francisco: Jossey-Bass, 1975.

Bloom. Benjamin S. (Ed.). *Taxonomy of educational objectives handbook I: cognitive domain*. New York: David McCay, 1956.

Bloom. Benjamin S., Hastings. J. T., & Madaus, George F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

Boyd. Joseph L., Jr., & Shimberg, B. Developing performance tests for classroom evaluation. *ERIC/TM Report # 4*. 1971.

Bracht. Glenn H., & Glass. Gene V. The external validity of experiments. *American Educational Research Journal*, 1968. 5, 437-474.

Caldwell. Bettye M. Practical developmental testing. *Pediatric Portfolio*, 1972.

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

Carver. Ronald P. Two dimensions of tests, psychometric and edumetric. *American Psychologist*, 1974, Vol. 29, No. 7. 512-518.

Cronbach. Lee J. *Essentials of psychological testing*. New York: Harper and Row. 1969.

Ebel. Robert L. *Essentials of educational measurement*. Englewood Cliffs. N. J.: Prentice-Hall, 1972.

Gallagher. James J., & Bradley, Robert H. Early identification of developmental difficulties. *Yearbook of the National Society for the Study of Education*, 1972. Part II.

Gronlund. Norman E. *Preparing criterion-referenced tests for classroom instruction*. New York: Macmillan. 1973.

Harrow. Anita J. *A taxonomy of the psychomotor domain*. New York: David McCay. 1972.

Hayman, John L. Educational management information systems for the seventies. *Educational Administration Quarterly*. 1974, 10, 60-71.

Herzig, Margaret E., Birch, Herbert G., Thomas, A., & Mendez, Olga A. Class and ethnic differences in the responsiveness of preschool children to cognitive demands. *Monographs of the Society for Research in Child Development*, Vol. 33, No. 1, 1968.

Hieronymus, A.N. Today's testing: what do we know how to do? *Proceedings of the 1971 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1971, Pp. 57-68.

Krathwohl. David R., Bloom, Benjamin S. and Masia, Bertram B. (Eds.) *Taxonomy of educational objectives handbook II: affective domain*. New York: David McCay, 1964.

Mercer. Jane R. A policy statement on assessment procedures and the rights of children. *Harvard Educational Review*, 1974, 44, 125-141.

Messick. S., & Anderson, Scarvia B. Educational testing, individual development, and social responsibility. *The Counseling Psychologist*, 1970, 2, 80-88.

Morgan. Don L. Evaluation: a semantic dilemma. *Educational Technology*, 1971, Vol. 11, No. 12, 46-48.

Snow. Richard E. Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 1974, 44, 265-292.

Stufflebeam. Daniel L., Foley, Walter J., Gephart, William J., Guba. Egon G., Hammond, Robert L., Merriman. Howard O., & Provus, Malcolm M. *Education evaluation and decision making*. Itasca, Ill.: Peacock, 1971.

Tyler. Ralph W., & Wolf, Richard M. (Eds.) *Crucial issues in testing*. Berkeley. Calif.: McCutchan Publishing Corporation. 1974.

Tyler. Ralph W. The use of tests in measuring the effectiveness of educational programs, methods, and instructional materials. In Ralph W. Tyler and Richard M. Wolf (Eds.). *Crucial Issues in Testing*. Berkeley, Calif.: McCutchan Publishing Corporation. 1974.