DOCUMENT RESUME

ED 242 427 PS 014 302

AUTHOR Haney, Walter
TITLE Thinking About Test Development.
INSTITUTION National Inst. of Education (ED), Washington, DC.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Jan 81
NOTE 24p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Educational Innovation; Elementary Secondary Education; Learning; Norm Referenced Tests; *Program Evaluation; Research Problems; Selection; *Standardized Tests; *Test Construction; Test Reliability; *Test Use; Test Validity
IDENTIFIERS *Inference; Project Follow Through

ABSTRACT

The question of how standardized tests can be better developed to improve educational program evaluation is probed in this paper. After the first section's brief introduction, section 2 explores the thesis that tests developed in terms of selection and inference may not serve current social functions of educational testing. To clarify this thesis, section 3 recounts an example of instrument development from the history of Project Follow Through, suggesting that the value of an instrument may be overlooked because the instrument is judged by criteria inappropriate to the original motivations behind its development effort. Section 4 attempts to go beyond the statement of the problem to suggest how thinking of a test as a source of individual learning might guide test development in nontraditional ways. Section 5 sums up some of the possible connections between testing and various social functions, pointing to some alternate ways in which standardized testing may serve goals of evaluation. (RH)

PS

THINKING ABOUT TEST DEVELOPMENT

by

Walter Haney
January 1981

The Huron Institute
123 Mount Auburn Street
Cambridge, MA 02138

Paper commissioned by the National Institute of Education for Project Follow Through Planning

I. INTRODUCTION

How can standardized tests be better developed to improve educational program evaluation? This question is the subject of this paper. I should hasten to make clear that I have no ready-made answers to this question. Rather I have some suggestions about ways of approaching the question -- approaches which lead, I think, toward some rather uncommon formulations of the possible relationships between standardized testing and educational evaluation.

By way of introduction I should explain what I mean by standardized test. I use the phrase standardized test in a fairly general sense to mean a systematic device for eliciting and recording a sampling of skills, knowledge, or attitudes. In this definition, I include such commonly recognized tests as aptitude and achievement tests, and norm- and criterion-referenced tests, and techniques such as systematic observation, and rating instruments, but exclude, at least for the sake of this discussion, teacher-made or classroom tests.

Specifically, my initial thesis in this paper is that educational tests are typically developed in terms of two functions traditionally assumed of educational tests -- namely, selection and formal inference -- but that these functions may not fit very well with some of the current social functions of educational testing. Educational tests, for example, seem to be serving more and more nowadays as a medium of communication, for discussion and debate over the goals and priorities of schooling. Indeed the minimum competency testing movement of recent years could be viewed as a social conversation on what should be the main aims of elementary and secondary schooling. For instance, when legislatures and

special committees debate whether high school graduation tests should cover "school skills" or "life skills," they are implicitly debating alternative aims of schooling.

A second example is that tests sometimes serve as social standards. Indeed, tests are widely perceived as devices for upholding educational standards, as for instance when they are viewed as antidotes to grade inflation or instruments for adding meaning to the high school diploma. To some extent, of course standardized tests already do serve as social standards. Indeed the notion is implicit in the phrase standardized tests. But note that if one set out to develop a test as a social or educational standard, it might not be necessary to employ the traditional techniques of test development.

A third example is that children learn directly from tests. Students may, of course learn indirectly as a result of tests in any number of ways -- because of college admissions decisions based on test results, or through teaching based on test results. But what I would like to explore is how individuals might learn directly from tests and test results, and how tests might be developed differently if this were one's aim.

To explore these issues, this paper is organized as follows. Section II describes the disjuncture to which I alluded above, namely that tests developed in light of the function of selection and inference may not well serve other functions. To make this thesis clearer, section III will recount an example of instrument development from the history of Project Follow Through to suggest that the value of an instrument may have been overlooked because it was judged by criteria inappropriate to the original motivations behind the instrument development

effort. Section IV attempts to go beyond the problem outlined to suggest how thinking of a test for a particular function, namely as a source of individual learning, might guide test development in ways somewhat different than those suggested by traditional standards of test development. The closing section, V, sums up some of the possible connections between testing and different social functions, and points to some alternative ways in which standardized testing may serve goals of evaluation.

II. THE PROBLEM.

The thesis outlined above was that traditional methods of constructing standardized tests are relevant to only some of the social functions which tests serve. To make this point clearer let me briefly describe some of the considerations which typically guide the construction of standardized tests.

Norm-referenced tests of achievement, aptitude and ability constitute the thickest branch in the family tree of standardized testing. The history of norm-referenced tests clearly suggests the success of such tests in informing selection processes. The original Binet test was designed, of course, to select French school children for special instruction because they could not profit from regular instruction. In the tremendous proliferation of testing in the first World War, the Army Alpha and Beta tests were used for military personnel selection. And the Scholastic Aptitude Test, introduced originally in 1926 and adapted into essentially its current form in the 1930s, is probably the pre-eminent example of a norm-referenced standardized test serving selection functions.

The tie between norm-referenced testing and the function of selection is apparent not just in historical perspective, but also in the techniques used to construct norm-referenced tests (NRT). Item difficulty and item-test correlations, for example, are two of the most widely used criteria in terms of which candidate items are selected for inclusion in norm-referenced tests. Also, of course, constructors of NRTs must pay heed to item content specifications, but as the technical report on the SAT notes, content specifications are "necessarily less rigorous" than difficulty and item-test correlations (Angoff, 1971, p.9). Now in terms of unitary selection decisions, these criteria contribute to important overall test characteristics. Difficulty contributes to the test's power to discriminate among test takers -- an important characteristic of a selection test, since practical selection decisions are almost always constrained in that some candidates must be selected, but not all can be. Similarly, item-test correlations contribute to the construct coherence of the selection instrument. If one is faced with a binary selection decision -- that is to select or not -- such an attribute surely can make matters simpler than if a selection instrument tapped several different constructs.

Nevertheless, desirable though these test characteristics may be from a selection perspective, critics of NRT have noted in recent years that these characteristics may not be desirable, or may even be undesirable, in light of other functions that tests may serve. Indeed, it is thinking along this line which has powered much interest in criterion-referenced tests in the last decade or so.

Several observers, for example, have directly criticized the widespread use of norm-referenced standardized tests in program evaluation (among others, Glaser, 1963; Carver, 1974; Popham, 1978; Madaus et al., 1979). The argument, in abbreviated form, goes roughly as follows. Since norm-referenced tests were designed to serve selection purposes and hence to discriminate efficiently among individual test takers, they have been constructed to be insensitive to effects of instruction in local school systems, which may have different curricula. Now tests are increasingly being used to evaluate educational programs and to guide instruction. However, precisely because of the way they are constructed, norm-referenced tests tend to be insensitive to the instructional effects of particular educational programs. Hence new types of tests are required for the purposes of program evaluation.

More extreme critics of norm-referenced tests have extended this argument; they predict that the weaknesses of norm-referenced tests will usher in a new period of educational assessment -- "the criterion-referenced measurement era" (Popham, 1978,p.2, emphasis in original). More moderate observers have suggested merely that curriculum-sensitive tests can play an important role in program evaluation, even though norm-referenced tests may continue to be valuable comparisons of the educational outcomes of programs that emphasize different aspects of instruction (Madaus et al., 1979).

If we are to judge from the continued popularity of norm-referenced tests, it seems doubtful that the criterion-referenced era is yet upon us. Nevertheless, there surely is much interest in criterion-referenced tests (CRT). According to one recent review of the state of the art of criterion-referenced measurement, so much has been written on this

topic that we now have available "more than fifty descriptions of a criterion-referenced test" (Berk, 1980, p.5). The most widely cited definition appears to be that of Popham, namely that a CRT "is used to ascertain an individual's status with respect to a well-defined behavioral domain" (Popham, 1978, p.93). Given this definition, it is not surprising to find it written that the most important step in the development of a CRT is "to define operationally the domain of content or behaviors the test is to measure" (Berk, 1980, p.13).

Yet when one examines recent literature on criterion-referenced measurement, a curious pattern is apparent. Far more has been written on technical issues of validity and reliability than on the "most important" step of defining what it is that a CRT is to measure. In Berk's (1980) book on the state of the art of criterion-referenced measurement, for example, the two brief chapters on domain specification/item generation contain a scant 34 references whereas the bulkier four chapters on validity and reliability contain over 180 references. In other words, work on criterion-referenced measurement seems to be progressing far faster on technical issues such as methods of item analysis, setting cut-off scores, assessing decision consistency, and applying generalizability theory to analyze variance in test results, than on the more fundamental issue of defining directly what it is that a criterion-referenced test is designed to measure.

Another means of illustrating this contrast is to cite an observation by Popham in the introductory chapter in the Berk (1980) volume. After recounting a variety of domain specification strategies that he has tried, Popham observes in closing:

> Once upon a time when I was younger and foolisher, I thought we could create test specifications so constraining that the test items produced . . . would be functionally homogeneous, that is, essentially interchangeable. But if we use the difficulty of an item as at least one index of the item's nature, then it becomes quite obvious that even in such teensy behavior domains as measuring the student's ability to multiply pairs of two digit numbers, the task of 11 x 11 = ? is lots easier than 99 x 99 = ?
> (Popham,1980,p.26)

Popham's observation nicely illustrates one of the essential problems of criterion-referenced measurement. It is that common constructs in terms of which we communicate about the substance and skills of learning often seem to have little coherence in terms of the common coin of educational measurement: right or wrong answers, item difficulties and test scores.

There may, of course, be strategies for surmounting this apparent problem, for example through longer tests, multiple measures, or statistical equating of various sorts. But my point in this paper is not on such theoretical problems. Rather I mean to suggest simply that many of the important social functions of educational tests may not depend on issues of formal inference, and that judging test instruments only or largely in terms of standards of formal inference may limit other social functions of tests. To illustrate this point I will bo on to suggest that if we view standardized tests not simply as measurement instruments but as sources of direct learning, then perhaps we might develop them in different ways.

III. AN EXAMPLE FROM THE HISTORY OF THE NATIONAL FOLLOW THROUGH EVALUATION.

To illustrate my thesis that judging test instruments in terms of techniques relevant to selection and formal inference may hinder their

application for alternative functions, in this section I recount one small portion of the history of the national evaluation of Project Follow Through. When Follow Through evaluation results were released in 1977, there ensued much debate about the narrowness of the outcome measures used, and the limited scope of the evaluation (House, et al; 1978). What was widely overlooked in the controversy over the FT evaluation results, however, was that a huge amount of effort was actually invested in assessing a wide range of the broad goals of FT. Indeed, through 1977 it was estimated that around $50 million or roughly 10 percent of total FT program costs were invested in the national evaluation (Haney, 1977,p.2). As far as I know this amount far surpasses typical program investment in evaluation. So if the FT evaluation was overly narrow it was surely not for want of resource investment in the task.

Now much of what was tried in the FT evaluation died or disappeared before it ever reached fruition. As I observed in writing a history of FT, the FT evaluation over time underwent "a sort of funnel vision," with dozens of questions asked of the evaluation at one time or another falling by the wayside (Haney, 1977, p.295). There were several reasons for the sloughing off of questions in the course of the FT evaluation. I will not even try to mention most of them here. Nevertheless, one cause relevant to the present topic, was the way in which evaluators went about developing and judging the quality of evaluation instruments.

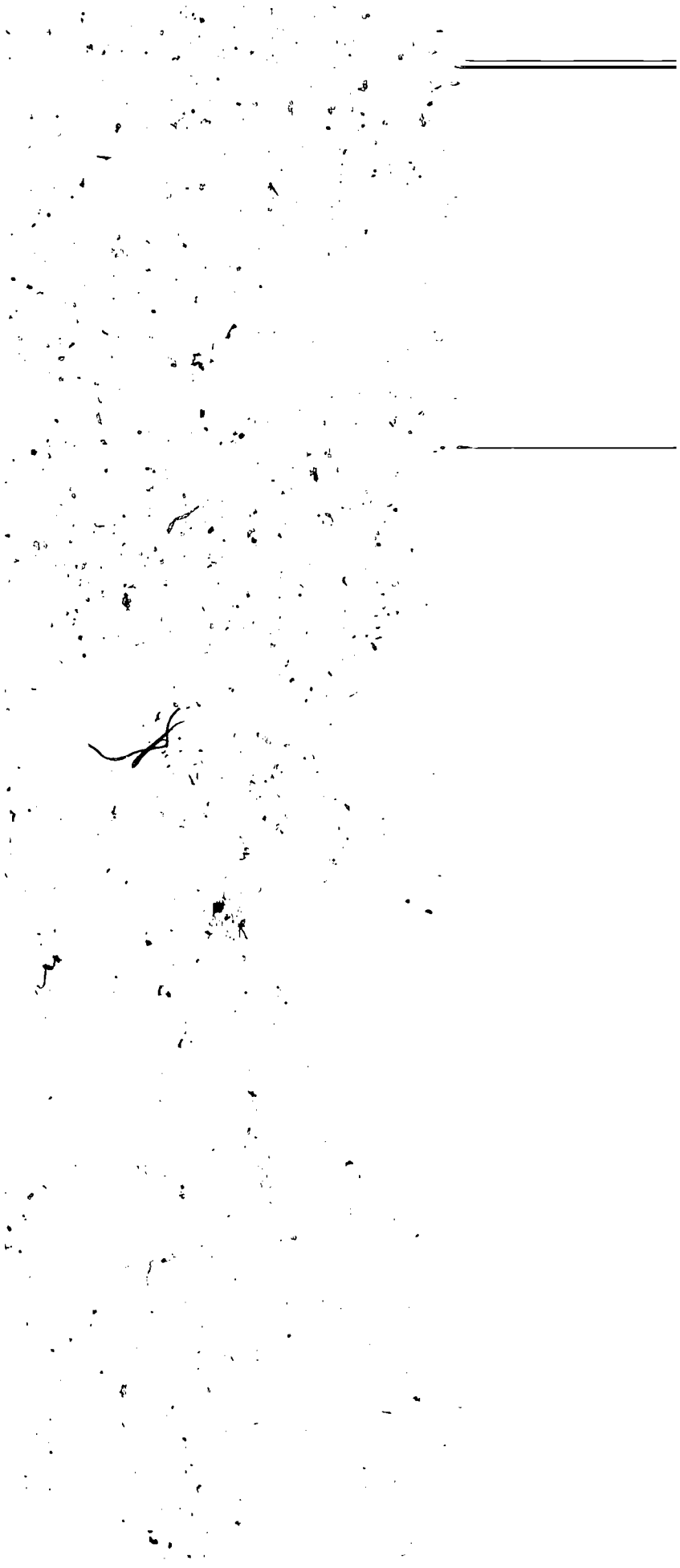
To illustrate how this worked, let me briefly recount the history of parent interview data in the FT evaluation (summarized from Haney, 1977, pp. 95, 258-269). From the very inception of FT, official program documents stressed the importance of involving parents in the program. Indeed, when official rules and regulations for the program were finally promulgated in 1977, one of seven explicitly stated evaluation criteria for FT

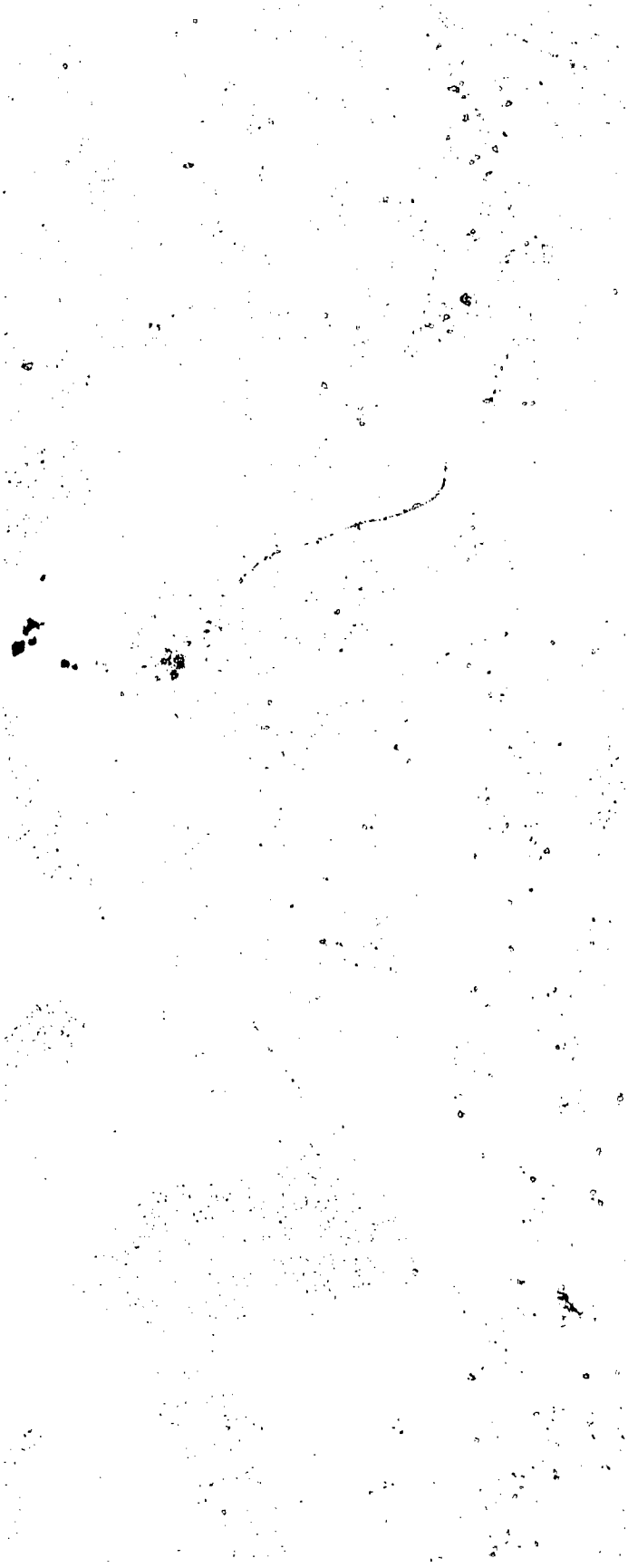
was the "extent of parent involvement." Given this emphasis, it is not surprising that considerable attention was given, as early as 1968, to interviewing parents of FT children, in part to obtain data on their involvement in the FT program. Between 1968 and 1975, over 60,000 parents were interviewed by the National Opinion Research Center to gather data for the national evaluation. Yet by the time of the final Abt evaluation report of FT, these data, gathered at tremendous expense, had almost completely disappeared from view. They were not even mentioned in the final "patterns of effects" chapter in the main volume of the final Abt report, nor in the Abt digest of evaluation findings.

There were several reasons for the virtual disappearance of the parent data, including ambiguity of purpose behind their gathering, organizational discontinuities, and simply too many demands on evaluators and too little time and resources to respond fully to all that different parties wanted done. But beyond such practical problems lay another cause, namely how evaluators went about analyzing the parent interview data, and assessing what they measured. Over the four years of the Abt evaluation effort, a variety of factor and cluster analyses were performed on the parent interview data. Now those techniques are widely recognized means of developing tests and understanding the meaning of test data, by identifying the constructs measured by data. But the problem which arose in applying these techniques to FT parent interview data was that from one year to the next, the results never turned out quite the same. Successively the Abt evaluators derived six clusters one year, eight clusters the second year, ten factors the third, and thirteen factors in the final, fourth year of analysis. Although some clusters and factors from the different

years of analysis contained the same parent interview questions, more often than not corresponding clusters and factors also included different interview questions. Such discontinuity across years of analysis quite effectively prevented any comparisons of results across years. While there are several alternative explanations for the virtual disapperance of the parent interview data in the national FT evaluation effort, one is this. The parent interview data-gathering was instituted to gather information on important aspects of the FT program, but data analysis designed to ascertain what constructs were represented in the parent interview data, revealed that they tapped no clearly consistent constructs across different years of data gathering. It is in a way, Popham's point writ large. Though the parent interview data had some coherence in terms of what was asked, the results of interview questions turned out to have little construct coherence in terms of interview responses.

This episode illustrates the disjuncture to which I alluded in the introduction -- namely a measurement procedure instituted for one set of reasons being judged in terms of techniques which imply another purpose. Specifically the parent interview data-gathering was instituted as a means of responding to one important aspect of FT, but the results came to be judged in terms of techniques -- namely factor and cluster analysis -- aimed essentially at identifying the construct coherence of parent interview responses. Such coherence is not, however, necessarily relevant to the original goals motivating the endeavor. Indeed when the parent interview data were reanalyzed, using simple cross tabulations, and on a judgmental basis grouping together items relevant to





specific FT goals, it was found that patterns of parental responses correspond in many cases with precisely what could be expected in terms of the goals of different FT models (Haney and Pennington, 1978, pp.103-104). Such correspondence could not of course be inferred with any great degree of confidence to be effects of FT model programs; but my point is that such simpler techniques, oriented more toward description than to inference, may have been more congruent with the original motivation behind introduction of parent interviews into the FT evaluation effort.

IV. DEVELOPING TESTS AS INSTRUMENTS FOR LEARNING.

If one accepts the proposition that commonly recognized techniques of test development, including both well-established techniques of NRT construction, and newer prescriptions on CRT construction, may be counterproductive with respect to functions of tests other than selection and formal inference, natural next questions are: 1) What other important social functions do tests serve, and 2) How could tests be developed so as to enhance those functions? In the introduction I suggested several different social functions which tests seem to be serving, namely as media for educational communication, as educational standards and as sources of learning. I will not try to speculate here on how tests might be developed differently if aimed at each of these, or other particular functions. Rather simply as a way of illustrating my more general point, I will attempt to suggest what considerations might go into developing tests as learning instruments.

A reasonable place from which to begin this exploration is simply to ask what makes for effective learning. Obviously different people have different answers to the question, but as a means of illustrating this approach to thinking about test development, let me work with one particular

set of theories of learning, namely Benjamin Bloom's writing on Human Characteristics and School Learning (1976), and his theory of mastery learning.

Bloom's theory encompasses the full range of the learning process including student characteristics, instruction, and learning outcomes. His observations on each of these areas have implications, I think, for how one might think about test development. Nevertheless, let me focus here on instruction, and specifically Bloom's observations on critical aspects of quality instruction. Bloom suggests that four characteristics seem to be important: cues, participation, reinforcement, and feedback. Before elaborating on what Bloom means by these terms let me note simply that one need not accept Bloom's theory lock, stock and barrel to be interested in these characteristics. As Bloom himself suggests, these aspects of quality instruction can be identified in other theroies of learning. Indeed, with respect to the first three, Bloom maintains that "although the terms may differ, they can be found in some respect in almost every theory of learning as summarized by Hilgard and Bower (1966)" (Bloom, 1976, p.172).

So what are these four features of quality instruction? Bloom describes them mainly in terms of tutor-student learning arrangements, but since I wish to suggest their broader applicability, I recount Bloom's description in paraphrase. Having done so, I will proceed to suggest what they imply for test development if we view tests as learning instruments.

Cues. It is made clear what is to be learned, what the student is to do, and how he is to do it. Cues can be altered or adapted to present those which work best for particular learners. For some students the cues

can be derived from written materials; for others it may be oral explanations; and for still others it may be combinations of demonstrations or models with explanations, and so forth.

Participation. The learner actively participates or practices the responses to be learned. While some of this participation may be overt and observable, it is also likely that covert participation may be as effective in some situations as the more overt or observable participation. There may be individual differences in the amount of practice or participation needed.

Reinforcement. Positive or negative reinforcement is used at various stages of the learning process. Reinforcers are adapted to the learner since what is an excellent reward for one student may not operate in the same way for another. A variety of reinforcers (both extrinsic and intrinsic) are used.

Feedback. Individual students receive evidence on the effectiveness of the learning process. Relatively rapid corrective feedback is provided when and where needed. "Furthermore, through the use of a variety of instructional materials, students helping each other, or tutors or aides, mastery learning procedures have made it possible to quickly apply correctives with regard to cues participation and reinforcement where the learners have specific difficulties in the learning process" (pp. 172-173).

Now suppose we accept Bloom's formulation of these aspects as critical components of an effective learning system. Suppose further that we view tests not just as measurement devices from which teachers or tutors derive information to use in applying Bloom's ideas to instruction, but also as learning instruments from which test-takers might learn directly.

From this perspective and in light of Bloom's critical features of a learning system, how might tests be developed differently than they typically are at present? Bloom's advice regarding cues suggests that tests might be more clearly labelled, not in terms of psychological constructs or abstract learning domains, but instead in terms more familiar to student test-takers. The idea of adaptable modes of presenting cues also might imply alternative means of test presentation; for example, oral, written and demonstration. When tests are viewed strictly as measurement devices, such alternative modes might be viewed as a problem, namely as extraneous sources of error variance. But from the learning perspective, alternative modes might be viewed more positively as differentially appropriate for students with different learning styles.

Bloom's notions of participation seem to imply several alterations from traditional test development procedures. At a minimum they suggest less emphasis on external control over administrative conditions, and scoring of results. Test items that are either self-scoring or scoreable by the student him- or herself would, for example, seem to have considerable potential for enhancing active learner participation in assessment. Likewise, the notion that different learners may need different amounts of participation and practice would suggest that tests would not necessarily need to be of uniform length for all test-takers.

Bloom's third aspect of quality instruction is reinforcement, either positive or negative, at various stages in the learning process. He notes further that what is excellent reinforcement for one student may not operate

in the same way for another student. This suggests that reinforcement which students derive from tests might best take different forms. For example, instead of all students receiving overall percentage correct scores -- or some norm-referenced or criterion-referenced score derived from percentage correct -- perhaps instrument scoring procedures could be adapted so that test-takers could receive results in the form of item types or sets in which they scored highest (positive reinforcement) or lowest (negative reinforcement).

Bloom's recommendations regarding rapid feedback suggest that tests might be constructed, not only so that they are self-scoring or scoreable by the test-taker him- or herself, but also so that results convey specific information or cues on types of errors or sources of information on corrective instruction. With regard to self-scoring, for example, might it not be possible for tests to employ materials and techniques already used in instant lottery tickets, so that test-takers could gain immediate feedback on whether their answers were right or wrong. Such self-scoring answer sheets have been used as far back as 1935 in the Henmon-Nelson Test of Mental Ability (which used the Clapp-Young self-marking device patented in 1929) as an aid to test administrators, but as far as I know such techniques have not been widely viewed as a potential source of enhancing test-taker participation in the assessment process.[1]

---

1. I know of little research bearing directly on the issue of immediate feedback of test-results. One relevant study, of computerized adaptive testing, concluded that "testees reacted very favorably to the provision of knowledge of results" and that this knowledge of results "increased average testee motivation." (Prestwood, 1978, p.105)

The only instrument of which I know that has employed such techniques in this way is the TORQUE developed at the Education Development Center, but unfortunately this unusual test development effort seems to have come to a halt before any large-scale try-out and evaluation could be accomplished.

In short, this brief review of how tests might be developed as instructional devices, specifically as direct aids to individual learning suggests that tests developed with this aim in mind might have several features which are not now found in most standardized tests. Specifically, they might

- be available in alternative modes of presentation
- be labelled in terms familiar to test-takers rather than in terms of psychological constructs on behavioral domains
- not require standardized administration
- be self-scoring or scoreable by individual test-takers
- be of variable length
- provide results not only on whether answers are right or wrong but on the nature of errors or sources of corrective instruction.

The process of developing tests with such characteristics obviously would entail less attention to the artifacts of tests -- namely the score results in terms of which the qualities of standardized tests typically are judged -- and more attention to the content of test questions and the way in which individual test-takers interpret and react to them. It would, for instance, require something akin to what curriculum developers call learner verification, and less attention to tests and test items as strictly measurement devices, to their discriminatory power, and to their empirical construct coherence.

V. LEARNING, MEASUREMENT AND EVALUATION.

These ideas obviously raise the question of whether tests with the characteristics I have described would really be tests, as this term is commonly understood. After all, standardized tests are more commonly thought of as instruments of educational measurement, than as instruments of learning, or educational standards, or media of communication. My answer is yes, for what I have been suggesting is exactly that standardized tests in the various roles they serve already are not and need not be viewed simply as measurement instruments.

Why? Because the limits of measurement are quite severe. In arguing this point, I discount the broader definitions of measurement -- for example, S. S. Stevens' view that measurement is simply "the assignment of numerals to things so as to represent facts and conventions about them" (Stevens, 1960, p.148), and Ernest Nagel's sweeping definition that "measurement can be regarded as the definition of and fixation of our ideas of things so that the determination of what it is to be a man or to be a circle is a case of measurement" (1960,p.121). Instead, I refer more narrowly to Lyle Jones definition that "measurement . . . is a determination of the magnitude of a specified attribute of the object, organism . . . or event in terms of a unit of measurement" (1971). Given this definition, and as long as we discount tautologies of the sort advanced with respect to intelligence tests--namely that intelligence is what intelligence tests measure--my point is simply that there is much in education and many sorts of learning which cannot be measured, whose magnitudes cannot be determined.

More generally, it seems quite clear that many social functions of standardized tests are not dependent on their qualities as measurement devices. This point can be illustrated by referring to the Eighth Measurements Yearbook (8MMY, Buros, 1978). As the introduction to this massive two-volume publication points out, the two most widely cited test instruments are the Minnesota Multiphasic Personality Inventory and the Rorschach -- each with around 5000 cumulative total references in the Buros' series of publications, while the average number of references for instruments listed in 8MMY is only 25 or so (Buros, 1978, p.xxxix). Why should these tests be so widely used? Surely it is not because of their proven validity and reliability as measurement instruments. As one reviewer of the Rorschach suggests,

> Certainly the validity research on the Rorschach does not warrant its popularity. Rather it seems it is the role the Rorschach has played within the psychodynamic oriented approach to psychopathology that has resulted in its popularity. Few instruments provide data so rich with hypothetical dynamic associations as does the Rorschach. When the goal of assessment is to formulate complex personality structures and complex dynamic interactions as the cause of the observed behavior, the Rorschach elicits responses which can be multi-interpreted and combined in an endless set of associations to produce speculative complex hypotheses and interpretations.
>
> (Peterson, in Buros, 1978, p.1042)

If I may offer a uni-interpretation of that passage, it seems as if this fellow is saying that the Rorschach is popular not because it helps answer questions, but because it multiplies them. This suggests that standardized tests, for at least some purposes, are valued not as valid and reliable measurement instruments per se but because they yield information which can be interpreted in numerous different ways.

It is an unusual perspective on the value of test information, but oddly enough it seems not too different from some recent thinking about program evaluation. Recall that not too many years ago, educational program evaluation was viewed mainly as applied social science research in the service of decision-making. Emphasis was on estimating effects of educational programs, most often by using standardized tests. But research on the utility of evaluation research has shown that evaluation findings rarely seem to have contributed directly to decision-making in the way that was expected (Cohen & Garet, 1975; Weiss, 1977). Instead, it seems often to be used in a more general way, indirectly influencing the way in which people think about education and educational programs. At least partly as a result, many seem now to think of program evaluation less as applied science and more as a descriptive enterprise, with more attention given to program implementation and depiction of how programs operate, even if their effects cannot be confidently measured. Evaluation as effects measurement is, of course still alive and well in some quarters, but we also now have evaluation as investigative reporting, evaluation as story-telling, and evaluation as art. From this angle a more general way of making the point of this paper, is simply to say that to the extent that program evaluation has shifted away from the goal of formal inference of program effects, perhaps also testing as part of the evaluative enterprise should also be aimed less at formal inference and selection and more at description. Test instruments as vehicles for communication and sources of direct learning may not, I realize, seem terribly

relevant to conceptions of evaluation as applied research.[1] But such roles may nevertheless serve the larger meaning of evaluation and its ultimate goal. For if we take the meaning of evaluation to be ascertaining values of programs, it is clear that this can never be reduced strictly to a technical or scientific affair. And if the goal of educational evaluation is improvement of education we need not restrict ourselves to a paradigm by which evaluators produce knowledge to give to educators for purposes of educational improvement. Perhaps instead we might view the role of evaluators as providing tools to educators and society generally with which to communicate about education goals and values, and as providing instruments to learners to improve learning.

---

[1] This point should not, however, be overstated. For one of the significant features of thinking on social science research in recent years is that it need not, and perhaps should not strive at building all powerful theories and parsimonious generalizations, but instead should attend to fuller and more thorough descriptions. For example, Cronbach recently argued:

> Social scientists generally, and psychologists in particular, have modeled their work on physical science, aspiring to amass empirical generalizations, to restructure them into more general laws, and to weld scattered laws into coherent theory. That lofty aspiration is far from realization. . . . Social scientist are rightly proud of the discipline we draw from the natural-science side of our ancestry. Scientific discipline is what we uniquely add to the time-honored ways of studying man. Too narrow an identification with science, however, has fixed our eyes upon an inappropriate goal. The goal of our work, I have argued here, is not to amass generalizations atop which a theoretical tower can someday be erected (cf. Scriven, 1959b, p.471). The special task of the social scientist in each generation is to pin down the contempory facts.
>
> (Cronbach, 1975)

REFERENCES

Angoff, W. (Ed.) College Board Admissions Testing Program Technical Report. New York: College Entrance Examination Board, 1971.

Berk, R: (Ed.) Criterion referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins, 1980.

Bloom, B. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Buros, O. The eighth mental measurement yearbook. Highland Park, NJ: Gryphon Press, 1978.

Cohen, D. & Garet, M. Reforming educational policy with applied research. Harvard Educational Review 1975, 45, 17-43.

Carver, R. The two dimensions of tests psychometric and edumetric. American Psychologist. 1974, 29, 512-518.

Cronbach, L. Beyond the two disciplines of scientific psychology. American Psychologist. Feb. 1975, pp. 116-127.

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

Haney, W. The Follow Through planned variation experiment. Volume V. A technical history of the national Follow Through evaluation. Cambridge, MA: The Huron Institute, 1977.

Haney, W. & Pennington, N. Reanalysis of Follow Through parent and teacher data from spring 1975. Cambridge, MA: The Huron Institute, October 1978.

House, et al. No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review 1978, 48, 128-160.

Jones, L. The nature of measurement. In R. L. Thorndike, Educational Measurement 2nd Edition. Washington, DC: American Council of Education, 1971.

Madaus, G. et al. The sensitivity of measures of school effectiveness. Harvard Educational Review. 1979, 49, 207-230.

Nagel, E. "Measurement" in A. Danto & S. Morgebesser (Eds.) Philosophy of Measurement. New York: Meridian Books, 1960.

Popham, J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice Hall, 1978.

Popham, J. Domain specification strategies. In Berk, 1980, pp.15-31.

Prestwood, J. S. Effects of knowledge of results and varying proportions correct on ability test performance and psychological variables. In Weiss, D. (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: Univ. of Minnesota, July 1978, pp.105-115.

Stevens, S. On the theory of scales of measurement in A. Danto & S. Morgebesser (Eds.) Philosophy of Science. New York: Meriden Books, 1960.

Weiss, C. Using social research in public policy making. Lexington, MA: Heath, 1977.