DOCUMENT RESUME

ED 081 842 TM 003 203

AUTHOR Wilson, H. A.
TITLE A Humanistic Approach to Criterion Referenced Testing.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
PUB DATE 18 Sep 72
NOTE 21p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; *Criterion Referenced Tests; Educational Objectives; *Educational Philosophy; Humanism; Item Sampling; Literature Reviews; Standardized Tests; *Test Construction; Test Validity

ABSTRACT

Test construction is not the strictly logical process that we might wish it to be. This is particularly true in a large on-going project such as the National Assessment of Educational Progress (NAEP). Most of the really deep questions can only be answered by the exercise of well-informed human judgment. Criterion-referenced testing is still a term in search of definition. It has been suggested that NAEL's exercises might be more properly called "objective referenced" tests. That is a reasonable title for our efforts since we are attempting to assess the degree of achievement of stated goals without reference to a predetermined level or criterion. Whatever the appropriate title may be, we share the concerns of all workers in the field with the same basic questions. But until satisfactory scientific solutions have been found; we, like the rest of education, must rely on the best human judgment available. (Author)

A HUMANISTIC APPROACH TO CRITERION REFERENCED TESTING

H. A. Wilson
Director/Exercise Development
National Assessment of Educational Progress

September 18, 1972

A HUMANISTIC APPROACH TO CRITERION REFERENCED TESTING

An adequate theory in science serves as a foundation for practical applications as well as a framework for experimentation. Intellectual activities that lack such a solid theoretical basis might be realistically considered arts rather than sciences. Much of education is art in that activities are based on the intuition and judgment of practitioners rather than logical extensions of quantifiable theory. The field of educational measurement is an exception since an extensive body of theory has been developed that guides the activities of norm-referenced testing. However, a much older tradition exists in educational measurement that attempts to determine the aboslute achievement of the individual or population without regard to interpersonal comparisons. That tradition, which is currently called criterion referenced testing, can call on much of the statistical technic that is used in other fields of measurement. It is faced, however, with important problems in basic theory that norm-referenced testing can, by definition, safely ignore.

Activity in criterion referenced testing, like the rest of education, cannot be delayed until basic applicable theory is developed. Schools cannot close their doors until a comprehensive theory of learning is found. Neither can assessment activities be halted. We must, instead, rely on human judgment to solve practical problems while we work on basic theory. This is the situation currently faced by the National Assessment of Educational Progress (NAEP).

A brief overview of the history and purposes on the National Assessment might be useful as background for a discussion of NAEP's responses to important theoretical questions in the area of criterion referenced testing.

> By the early 1960's many billions of dollars were being invested annually in the formal education of our young people. The only available measures of educational quality resulting from this investment had been based upon inputs into the educational system such as teacher-student ratios, number of classrooms, and number of dollars spent per student. The tenuous assumption had been that the quality of educational outcomes. --what students actually learn--was directly related to the quality of the inputs into the educational system. No significant direct assessment of educational outcomes had been made. The typical state-administered or school-administered achievement tests, which provided scores whereby one student could be compared with others, were useful for categorizing students; but they provided very little information about what students were actually learning.
>
> This insufficiency of information became the concern of Francis Keppel, United States Commissioner of Education (1962-1965), who initiated a series of conferences to find ways in which it might be overcome. In 1964, as a result of these conferences, John W. Gardner, president of the Carnegie Corporation, asked a distinguished group of educators and lay persons to form the Exploratory Committee on Assessing the Progress of Education (ECAPE). This committee, chaired by Dr. Ralph W. Tyler, was to examine the possibility of conducting an assessment of educational attainments on a national basis.
>
> After much study, ECAPE deemed that it was feasible to assess the knowledges, understandings, skills, and attitudes in 10 subject areas [1] at four age levels (9, 13, 17, and adult--ages 26-35). The project began its first assessment of the subject areas Science, Citizenship, and Writing in the Spring of 1969. Later that same year, the project came under the auspices of the Education Commission of the States and was named the National Assessment of Educational Progress (NAEP).

---

[1] Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing.

> For the first time, there would be a direct measure of educational outcomes which could be utilized by school systems to improve the educational process. Since NAEP is to be an ongoing project, it will eventually be able to assess changes in these knowledges, understandings, skills, and attitudes to determine any changes in educational outcomes.[2]

Many people, prominent in education and measurement have contributed heavily to the purposes and processes of NAEP. A brief and very incomplete roster would include, besides Tyler, Keppel and Gardner, Jack Merwin, Frank Womer, Stanley Ahmann, John Tukey, Frederick Mosteller, and Lee Cronbach.

Two subject areas are currently being assessed each year with a five year cycle for reassessment within a given subject area. The five year assessment-reassessment cycle and the 210 minutes allotted to each subject area at each age level in an assessment year place very practical constrints on the design and production of exercises (test items). The five year cycle requires continuous exercise development effort and limits experimental and validation activities. The time allotment limits the number of exercises administered and hence, the depth of coverage for each objective.

Universe Definition

Some of the most intriguing questions in the field of criterion referenced measurement have to do with the rigorous definition of a

---

[2] This section was adopted from What is National Assessment by Dr. Frank Womer and The National Assessment Approach to Exercise Development by Drs. Carmen J. Finley and Frances S. Berdie and may be obtained from: National Assessment of Educational Progress, Public Information Department, 300 Lincoln Tower, 1860 Lincoln Street, Denver, Colorado 80203.

domain of reference (subject matter) and of a universe of behaviors within that domain. This paper will briefly summarize some of those questions and indicate the general thrust of NAEP's responses. The responses discussed in this paper are to be viewed only as current positions of NAEP regarding the basic problems. They are in no sense offered as definitive solutions.

Two questions must be asked:

1. What constitutes a definition of a domain of reference or a universe of behaviors?

2. When can we be sure that a complete definition is achieved?

Since the problems of defining a domain of reference and a universe of behavior are parallel, discussion of a domain of reference can serve as a model for discussion of a universe of behaviors.

It is clear that a complete definition of a domain of reference must include all knowledge, skills and attitudes directly related to the subject area and exclude all those that are not related. A similar statement could be made for defining a universe of behaviors by substituting "behaviors" for "knowledge, skills and attitudes." Such a definition need not be an enumeration. Indeed, such an enumeration would be useless because of its extensive, if not infinite, length.

What is needed then is a method of statement generation that will produce relevant and only relevant statements. We can be sure that a complete definition is achieved only when it can be logically shown that any statement or question that can be made by our statement

generation mechanism is or is not a member of the set of questions and statements contained in that domain or universe.

Lacking a logically complete knowledge generator, it is not possible to make statistically defensable and generalizable statements relating individual or group performance to a subject area by means of a restricted set of items. Without a complete definition of the domain of reference and a universe of behaviors, all statements about the results of a criterion referenced test must be confined to the items in that test without further generalizations. Clearly, this is not the purpose of any test maker.

Several approaches to the problem of generalizability can be found in the literature. One approach is to ignore the problem altogether. Another is to indicate how certain domains and universes can be defined and systematically sampled. Unfortunately, those domains and universes that have been discussed are typically narrowly restrictive or trivial or both. For example, tests of knowledge of word meanings can be constructed by defining the domain of reference as the Merriam-Webster Collegiate Dictionary, 7th Edition. All statements about words contained in that dictionary are relevant and all statements not contained in that dictionary are not relevant. One can then define the universe of behaviors as responses to a cloze test on the definitional entry for each word. Many schemes can then be devised for systematically sampling both the domain of reference and the universe of behaviors. Item generation rules can be devised which will produce any number of equivalent tests and the results of those tests can indeed be generalized

to knowledge of word meanings as defined in the domain of reference. Such schemes are of little value, however, in constructing tests to assess knowledge, skills and attitudes in broader areas such as social studies, literature, music or art.

Objectives

It is clearly beyond the current state of the art to define the universe of discourse for a complex area in the strict sense discussed above. Yet it is equally clear that a set of exercises (test items) which form a coherent assessment of a subject area cannot be constructed without some definition of the domain to be tested. Faced with this conundrum, NAEP has taken a humanistic rather than a statistical approach to universe definition.

The term "humanistic" is used to indicate reliance on human judgment rather than logical or statistical proof. We define our universe by producing a set of objectives that represent a consensus of opinion covering many segments of our society regarding the important goals and outcomes of our educational processes in respect to a given subject area.

The question might well be raised, "Why add yet another formulation of educational goals and objectives to the already existing plethora of such documents?" It is certainly a reasonable question and yet one that is easily answered in terms of NAEP's mission. NAEP, as its name states, is a national assessment and as such is compelled

to attend to those aspects of education whose definition and evaluation can be agreed upon for the society as a whole. Most of the myriad statements of objectives are produced by and for the use of schools at the local and state level. NAEP must go beyond that restricted viewpoint to identify goals that are accepted nationally.

Since NAEP is also an assessment of change in educational outcomes over time we have the further responsibility to examine and revise our codifications of objectives on a systematic cyclical basis. These twin requirements of demonstrable national significance and continuous revision justify the effort to produce statements of goals and objectives that are unique to our own needs and purposes.

NAEP defines the domain of reference in a subject area by arriving at a national consensus statement of goals in that area. Goals are stated in the form of overall objectives with attendant levels of sub-objectives. The form and structure of the objectives varies from one subject area to another and between assessment cycles within a single subject area. For example, a major objective and its sub-objectives for cycle 1 of Music were stated as follows:

III. LISTEN TO MUSIC WITH UNDERSTANDING.

A. Perceive the various elements of music, such as timbre, rhythm, melody and harmony, and texture.

1. Identify timbres.

Age 9 Identify by categories the manner in which the instrument is played (e.g., struck, bowed).

Identify individual instrumental timbres--unaccompanied.
Identify individual instrumental timbres--with accompaniment.

Age 13 (In addition to Age 9)

Identify individual vocal timbres--with accompaniment.
Identify ensemble timbres, instrumental and vocal.

Age 17 Adults
Identify by categories families of related timbres(e.g. woodwinds, plucked strings).
Identify individual instrumental timbres--unaccompanied.
Identify individual instrumental and vocal timbres--with accompaniment.
Identify ensemble timbres, instrumental and vocal.

A much more loosely defined objectives structure was produced for the first cycle of Literature assessment as shown by the following example:

III. DEVELOP A CONTINUING INTEREST AND PARTICIPATION IN LITERATURE AND THE LITERARY EXPERIENCE

This goal is directed at assessing the interests and attitudes; for the most part the goal is relevant to Age 17 and Adult.

A. Be intellectually oriented to literature.
This goal asks of the individual a recognition of the importance of literature to the individual and society, and a recognition that literary expression requires a number of forms to enable it to become an art.

All ages
Recognize the importance of literature to an understanding of cultures distant in time or distinct in history.
Recognize the importance of literature to a comprehension of the diversity and homogeneity of man.
Recognize that participating in the literary experience is a prime form of enjoyment.

Age 17 Adults	Recognize the necessity of a free literature in a free society. Recognize that the art of literature involves a close connection between form and content.

The process of identifying and explicating objectives or revising those used in the previous cycle of assessment of a subject area is somewhat complex and occupies a time span of approximately nine months. A search of recent literature is made to identify new trends in the subject area. The literature search is coupled with an examination of existing sets of written objectives such as those brought together by The Instructional Objectives Exchange. This material forms a background for a number of working and review panels that produce and refine the objectives to be used as the basis for exercise development and for reporting of assessment results.

In the early years of NAEP, objectives development was done by sub-contractors (AIR, ETS, SRA, etc.). They studied the literature, examined existing objectives and produced a document that was critiqued by a variety of consultants and then revised. This plan was followed for the objectives development of most of the first cycle assessments. Leaving objectives development in the hands of the contractors who then wrote the exercises not only produced objectives of uneven quality but also was fraught with the danger of producing only those objectives that were most easily measured and neglecting those that might be at least as important to the education community but are difficult to measure.

With these considerations in mind, the task of producing objectives was removed from the purvue of sub-contractors and made part of the direct responsibility of the Exercise Development department of NAEP. A standardized procedure for developing objectives is now followed that begins with a mail review by subject matter experts of the objectives from the previous cycle. This mail review is followed by a conference in which consultants determine the broad outlines of the desired revision. A sub-set of consultants from the first review conference produce a first draft of the revised objectives within the guidelines from the conference. This draft is reviewed by mail by members of the first conference and a second draft is produced based on the resulting comments. The second draft is then reviewed by a second conference of consultants some of whom were present at the first revision conference. Consensus is reached on the remaining points at issue among the consultants and the document is adjusted accordingly and given a final editorial polish.

The working and review panels are composed of consultants drawn from three major groups: scholars and educators within the subject area and qualified and interested laymen. Between 35 and 50 consultants are involved at one time and another in the development of objectives. Consultants are chosen with serious attention to representation by region (northeast, southeast, central and west), type of institution (university, four year college, junior college, secondary and elementary schools and private schools), race and sex. Wherever clearly defined schools of thought hold differing positions in a subject area, care is given to assure

representation of each of the conflicting points of view. The above describes the selection of consultants who serve actively on panels. In the case of mail reviews, a much larger number of people are involved.

The method just described, while far from ideal, does produce a set of objectives that represent as nearly as possible, within the constraints of time and money, a national consenses on educational goals and objectives that are currently valued by our society. Great emphasis is placed on producing objectives that are important without regard to their measurability. NAEP views the objectives as defining the broad domain within which exercises are to be written and as a mandate from our society to produce data on related educational outcomes.

There are a number of important questions still unresolved in the area of objective development. The question with the deepest theoretical implications is , "To what depth of sub-objective level and of age specific behavior should objectives be taken?" The major objectives are generally few in number and are of such a general nature that they provide only an ambiguous guide to exercise development. At each level of sub-objective, the domain of reference is more clearly defined but how clear that definition can be or should be is still an open question. There is currently a large variation in this matter from subject area to subject area and between assessment cycles within any given subject area. The use of age specific behaviors in the objectives furnishes the clearest definition and guide for exercise development. However, it is again a

question of viewing age specific behaviors as an exhaustive list of all possible behaviors (an obviously impossible task) or simply as guidelines and illustrations for the exercise developers.

A second and related question has to do with the feasibility of developing some sort of hierarchial scheme of cognitive and affective objectives. Many such schemes have been devised but is it possible or even advisible to choose one plan to the exclusion of all others?

A final question has to do with standardizing the format of objectives. It has been suggested that from a quality control standpoint, a standardized format and framework of objectives should be developed and applied to all subject areas. There is no solid agreement, however, that this plan, if it could be implemented, would be desirable. The discussion on this point revolves around the issue of the amount of freedom allowed to the developers of the objectives to express in their own way those aspects of the subject area that they feel to be most important in our educational scheme.

Item Generation Rules

In criterion referenced testing it would be desirable to identify a generally acceptable method for item construction. In the strict sense, such a method should provide a systematic sampling of a previously defined universe of behaviors. Further, it should be a set of rules which, if followed by more than one person or group of item writers with equivalent knowledge, would produce equivalent tests. We have

already discussed the difficulties involved in domain and universe definition in the complex areas of interest to NAEP. Since the universe of behaviors has not been well defined, a systematic sampling scheme is difficult to devise. When the notion that a set of rules may be clearly enough stated that equivalent tests may be generated from them is examined closely, it is easily seen that such rules, while useful in narrowly specialized areas, are not definable in other more complex areas. Tests of arithmetic computational skills, tests of word meanings and spelling tests have been constructed using such rule sets. Indeed, on occasion, rules have been embodied in computer programs which will generate equivalent tests ad infinitum. Unfortunately, such tests, while complete in themselves, fall far short of being comprehensive tests of mathematics, reading or writing.

Assuming for a moment that solutions were at hand for problems of defining the universe of behaviors and of stipulating an adequate set of rules for generating items, we are still faced with a question of serious theoretical consequences. The question might be phrased as,"How much is enough?" How many items are necessary for an acceptable test of an objective? If the objectives are complete through the identification of one or more levels of sub-objectives under each major objective, and if each sub-objective is adequately tested, then we can certainly claim that we have an adequate test of a major objective. However, such a plan simply puts off the problem to another level of detail. We are still faced with the central question of how many items are necessary to test the lowest level sub-objective or any given age specific behavior.

NAEP Exercise Development

In light of the problems outlined above, we may move to a brief discussion of the methods used by NAEP in generating exercises (test items). None of the activities to be described below are presented as final solutions but it will be seen that many of our item generating activities, while perhaps tangential to the central problems as stated above, do stem from our abiding concern for such problems. Again, as in the definition of domains of reference and universes of behavior, it will also be seen that we continue to use a humanistic approach in the sense of relying primarily on the judgment of experts in the subject matter area.

Following the development of objectives, contracts are awarded through competitive bidding for the generation of exercises to assess those objectives. The amount of exercise material to be developed for each sub-objective is based on a "weighting" scheme. Weights are assigned by subject matter experts who are experienced with students at the four age levels. For example, the major objectives are weighted for their relative importance for nine-year-olds by teachers who have experience with that age group. Each sub-objective is then weighted for its relative importance within the major objective. This scheme is continued to the lowest level of sub-objective. The weights for an objective may differ widely over age groups reflecting the importance of that objective at one age as opposed to another.

The use of weights is in some sense a response to the problem of providing adequate coverage for each sub-objective. Since the weight of the sub-objective is an index of its importance in relation to other objectives at a given age level, such weights can easily be translated into percentages of the total assessment time that it would be reasonable to spend in assessing that particular sub-objective. This method of specifying coverage of course accounts only for amount of material related to its importance and does not speak to the issue of relating coverage to the complexity of the various objectives and sub-objectives.

Much attention has been paid by NAEP to the problem of giving contractors an adequate framework for preparing the kinds of exericses that will achieve coverage through a variety of approaches. We have arrived at a general notion of exercise prototypes which are not rules for exercise generation nor are they examples of specific exercises, but rather attend to those aspects and variables of exercise generation that can be discussed. NAEP exercise prototypes are actually a tree structure showing mutually exclusive categories for four variables: Administration mode, stimulus mode, response mode and response category. The administration mode is dichotomous: an exercise can be administered either individually or to a group. Branching from administration mode we define the stimulus mode as audio, visual, other senses (tactual, olfactory, etc.) or some combination of the three. From each stimulus mode we show a dichotomy of response alternatives or response mode: objective (multiple choice) and free response. Finally,

branching from each response mode we define response categories as written, verbal, role playing, group interaction, and other physical action.

Such a tree structure results in 80 (2 x 4 x 2 x 5) possible prototypes. It is clear that not all possible prototypes are applicable to any given subject area. A panel of subject matter experts selects those prototypes that are most reasonable for assessing a subject area. Their input, in conjunction with practical considerations of cost of administration and scoring, provides the specification of percentage ranges (minimum and maximum) in terms of minutes of material as guidelines for the contractor. The subject matter experts also produce exemplary exercises within the subject area for each prototype specified. The use of prototypes as a control for coverage through a variety of approaches is frankly experimental. Its first use will be in the current redevelopment of literature assessment, but it is expected to provide a more balanced body of exercises.

Working within the weighted objectives, prototypes, and exemplary exercises, contractors produce the specified minutes of exercise material for assessment of a subject area. Each exercise produced by the contractor must be accompanied by a rationale relating that exercise to the sub-objective that it is proporting to measure. It must also be accompanied by a rationale relating that exercise to other exercises within the body of material to be used in the assessment.

The exercises received from the contractor are subjected to at least four reviews by each of three groups: the NAEP staff, subject matter experts (scholars and educators) and qualified laymen. In addition to reviewing the exercise itself, the rationale relating that exercise to a sub-objective and to other exercises in the body of material is also brought under scrutiny. Some exercises survive each review session; others are sent back to the contractor for suggested revisions and others, hopefully a small percentage, are rejected as being without merit and are no longer considered for use in the assessment.

Those exercises that have survived the reviews, either in their original or revised state, are then given a full field trial. Each exercise has been tried out during its developmental stages by the contractor and is submitted to NAEP accompanied by data from three sub-units of the population: extreme inner city, extreme rural, and affluent suburb. Data from the developmental tryouts consists of timing information, overall percentage correct responses, percentages of responses for each foil in a multiple choice exercise, and the beginnings of a scoring guide or response categorization in the case of free response exercises. While these data are gathered from three sub-units of the population, the number of subjects contributing from each population is necessarily small. For increased reliability of this sort of data, we run extensive field trials on a national sample. The field trials, while far less extensive than the actual assessment, are large enough to yield

reliable data and also point up regional biases and administrative problems that might otherwise be missed.

Following the field trials, the pool of exercises is reviewed by the United States Office of Education for possible offensiveness in sensitive areas. Exercises surviving this last review by USOE are then examined by successive panels of subject matter experts in a selection conference.

Since the attrition rate through all the reviews is unpredictable in any precise way, we order from the contractor a considerable overage of material. This overage is on the order of 100% plus an additional 20% that allows for contractor creativity outside of the specifications and guidelines furnished by NAEP.

Since we are constrained to a total of 210 minutes of assessment for each subject area, a selection conference is necessary to choose the best among surviving exercises. Consultants at the selection conference are required to pay close attention to maintaining the balance over objectives and sub-objectives that was specified in the original contract and to the relationships between exercises that forms a coherent assessment.

Validity

Two main concerns of NAEP for the assessment exercises is for their content validity and their importance. Two questions are continually asked at every exercise review conference: "Is this

... a valid measure of the objective for which it was written?" ... If it is valid, is it an important or a trivial measure of the objective?" Importance can only be established by following the judgment of subject matter experts. Human judgment is also the primary check on validity.

However, another measure of validity is available for some of the exercises by examining the assessment response data. If an item is administered to two groups, one of which has had no training or experience in the area and the other has had extensive training, the results can be viewed as one measure of the item's validity. In the ideal case, a valid item would yield a score near zero for the untrained group and approaching 100% correct for the highly trained group. Such a test is approximated for those NAEP exercises that overlap age groups. It may be assumed that seventeen-year-olds have had more training in a given subject area than thirteen-year-olds when training in that area is a continuous process. The same assumption may be made for comparison of thirteen-year-olds and nine-year-olds. If the same exercise is administered to the three age levels, an increasing percentage of correct responses from nine- to seventeen-year-olds can be accepted as some assurance of the item's validity. In general, such has been the case with NAEP data. If in the field trials a contrary instance is found, that item is examined closely. If an adequate explanation is not evident, the item is dropped from the assessment.

Summary

Test construction is not the strictly logical process that we might wish it to be. This is particularly true in a large on-going project such as NAEP. Most of the really deep questions can only be answered by the exercise of well informed human judgment. Criterion referenced testing is still a term in search of a definition. It has been suggested that NAEP's exercises might be more properly called "objective referenced" tests. That is a reasonable title for our efforts since we are attempting to assess the degree of achievement of stated goals without reference to a predetermined level or criterion. Whatever the appropriate title may be, we share the concerns of all workers in the field with the same basic questions. But until satisfactory scientific solutions have been found; we, like the rest of education, must rely on the best human judgment available.