DOCUMENT RESUME

ED 117 145 TM 004 999

AUTHOR Mathis, William J.
TITLE Large-Scale Objective Referenced Testing: Some Practical Problems and Concerns.
PUB DATE Apr 75
NOTE 7p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Criterion Referenced Tests; Educational Assessment; *School Districts; *State Programs; Test Construction; *Testing Problems; *Testing Programs; Test Validity

ABSTRACT

This paper was presented with other papers in a forum dealing with statewide testing programs. The primary purpose of the paper is to address practical considerations and methods of resolution for large districts or states who are planning on conducting large scale testing or assessment programs with criterion or performance referenced measures. The first section lists the parameters and limits within which these programs generally operate. These limits are translated into practical problems and decision points. Methods of resolving the problems are then addressed with emphases being given to professional and community involvement. The paper closes with comments on test validity and how it is affected by these problems and concerns. (Author)

LARGE-SCALE OBJECTIVE REFERENCED TESTING: SOME PRACTICAL PROBLEMS AND CONCERNS

WILLIAM J. MATHIS, Ph.D.

Division of Research, Planning, and Evaluation
New Jersey Department of Education

Paper presented at the 1975 NCME convention in the symposium sponsored by the National Council on Educational Assessment, Washington, D.C., April, 1975.

LARGE-SCALE OBJECTIVE REFERENCED TESTING: SOME PRACTICAL PROBLEMS AND CONCERNS

I. Introduction

An increasing number of states and large districts are moving toward objective referenced testing. Mary Hall of the Oregon Department of Education (1975) noted that twenty-eight percent of the states now use objective referenced tests, twenty percent implement eclectic approaches, and fifty-two percent use norm referenced materials. With the implementation of these types of examinations, a new set of problems are faced and new sets of procedures are required.

In understanding these problems, we must first examine some of the parameters within which we operate, as the problems are closely akin to these limits. After noting these parameters and problems, this paper will be concerned with process requirements for solving the problems and close with some thoughts on validity of objective referenced testing for large populations.

II. Parameters of Operation

For states or large districts, limits are placed on the development of tests. Coupled with these limits are the requirements of objective referenced testing.

The first of these is concerned with program diagnostics. Any objective referenced test will have to provide results in a format which can readily be used to assess program strengths and weaknesses. This calls for

careful report design and more extensive and detailed information. I must whole-heartedly endorse Lorrie Sheppard's comments on designing the report as a first step (1975).

The second requirement is the provision of "semi-diagnostic" individual student profiles. The word "semi-diagnostic" is used because a totally diagnostic picture is not practical with a large scale program. With the need for scoring thousands of answer sheets, the form is virtually forced into multiple-choice, close-ended types of questions. The length of testing time, usually less than three or four hours, precludes the in-depth diagnostic picture which would be desired.

A third need is for fast turn-around. A test which is to be of any value must have results returned in a very short period of time. Results which are three or four months stale prevents the program components from being reactive to identified needs and prevents the proper use of individual diagnostics.

III. Practical Problems

The above noted limits pose problems within themselves in that they define a rather tight space of operation. These limits are compounded with other test construction problems.

Probably the first problem of test development will be the breadth of the subject matter to be tested. Regardless of the area selected, the very breadth of the field will preclude the use of behavioral objectives with three items per

objective. For example, the simple specification of adding two, two-digit numbers could be expanded to five or more behavioral objectives when considering addition with zero, addition with carrying, without carrying, vertical formats, and horizontal formats.

In addition to this concern, taxonomies or classification systems for arranging the subject matter are always open to question. For mathematics, the decisions are relatively simple. For reading, however, organization is more difficult because there are several differing theoretical views on reading which have various implications for test content as well as test organization. As areas such as history or civics are more complicated, the degree of basic agreement on content is not found as in the basic skills. The affective areas show even less central agreement, with a great variation in the attributes to be measured and the dimensions along which they should be assessed.

These concerns will rapidly be translated into breadth versus depth decisions. Either a particular area of the discipline may be tested in a thorough fashion or a broad brush must be applied with the loss of discrete information in particular sub-areas. Trade-offs are required which will, in turn, require close consultation with the users.

Closely tied to the breadth/depth problem is that of test balance. Decisions must be made on whether encoding is more important than comprehension. Should study skills be included? To what degree?

All of these practical problems must be addressed in terms of the potential test user. If the program is very broad based, it is likely that little agreement will be found on these test building concerns. However, the tests will be valid only to the degree that there is agreement with the user on content. Consequently, mechanisms must be found to aid the process of consensus in addressing these problems.

III. Resolving Problems

In many respects, the process of resolving the problems is more important than the product itself. If the users are not directly involved, it is doubtful that the tests will enjoy the acceptance or that the test makers might wish. Consequently, the establishment of good, working interaction and feedback channels is absolutely necessary if the program is going to be effective.

In addressing the problems of breadth versus depth and taxonomies, it is important to incorporate subject matter specialists in the subject matter areas to assist in determining operational definitions. Even more important is the use of teachers. All too often, central staff and test developers have rosy-eyed views of the reality of what goes on in the classroom. Certainly, we must concern ourselves with <u>should</u> questions but we must not allow confusion with what <u>is</u> reality in the classrooms.

Our concerns with test balance, breadth, and consensus can be largely resolved by close working relations with the people using the scores.

IV. Comments on Validity

With any broad-based testing program, validity problems will increase as the number of schools and districts being tested increases. Likewise, validity concerns will be a direct function of the degree of district and school latitude in establishing their instructional objectives and materials. With the strong tradition of local control of education in the United States, state assessment programs will always face validity questions. It is consequently necessary that continuous re-validation of testing objectives be undertaken. This constant re-validation necessitates a further swap-off in that a common core must be retained if we are to have a worthwhile longitudinal data base.

As noted above, the further away from areas of high agreement, the more general the test content becomes and the potentially less valid the results for local use. The lack of composite scores in a norm-referenced fashion provides distinct advantages in handling this validity concern. The districts or schools which do not find the particular objectives relevant must be given the option of declaring these objectives and test scores as non-relevant for them.

In addressing all of the practical problems and procedures, it cannot be over-stressed that the key to resolution is in the quality of the interactions with the user. Any beginning assessment program will have practical problems and will require revisions. If the interactive relationships are positive, clear and precise information on where program improvements are needed will be received as well as the time and latitude to make these improvements.