DOCUMENT RESUME

ED 141 394 TM 006 343

AUTHOR Rubinstein, Sherry Ann; Nassif-Royer, Paula
TITLE The Outcomes of Statewide Assessment: Implications for Curriculum Evaluation.
PUB DATE [Apr 77]
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Content Analysis; *Criterion Referenced Tests; Decision Making; *Educational Assessment; Evaluation Methods; Item Analysis; Needs Assessment; Norm Referenced Tests; *State Programs; *Test Construction; *Test Validity

ABSTRACT

State Departments of Education are turning to the use of criterion referenced, as opposed to norm referenced, models for statewide assessment. The underlying assumption in this turn of events is that results generated by criterion referenced tests within the statewide assessment context permit the drawing of value inferences about the effectiveness of the educational curricula under study. The tenability of this assumption is examined in light of rigorous requirements for test construction and validation. The extent to which the test construction steps can be followed closely to yield a content valid test determines the extent to which the tests can be justifiably used to evaluate the curricula or programs under study. In summary, it is to be concluded that the content validity of measuring instruments must be carefully established in order to ensure meaningful and defensible decision making. The risk involved in using an invalid test must be judged in terms of the costs (psychological, financial, etc.) attendant on making erroneous decisions in a given situation. (MV)

THE OUTCOMES OF STATEWIDE ASSESSMENT:

IMPLICATIONS FOR CURRICULUM EVALUATION

Sherry Ann Rubinstein, Ph.D.
Paula Nassif-Royer, Ed.D.

National Evaluation Systems, Inc.
Post Office Box 226
Amherst, Massachusetts 01002

Paper presented at the Annual Meeting
of the American Educational Research Association
New York 1977

The Outcomes of Statewide Assessment: Implications for Curriculum Evaluation (EVA, VAR)

SHERRY ANN RUBINSTEIN, PH.D.

PAULA NASSIF-ROYER, ED.D.

National Evaluation Systems, Inc.

The paper analyzes the methodologies and results of selected statewide assessment programs in terms of their implications for evaluating curricula in related subject-area domains. The purpose is to determine the scope and limitations in usefulness of statewide data, given different characteristics of the tests designed to measure the objectives within a domain. Critical issues related to the selection of objectives and items for criterion-referenced instruments and the resulting content validity of the tests are discussed.

Purpose

In response to increasing claims about the merits of criterion-referenced testing (for example, Popham & Husek, 1969; Hambleton & Novick, 1973; and Popham, 1976) as well as to the impetus provided by the National Assessment of Educational Progress (NAEP) (Finley & Berdie, 1970; Womer, 1970), State Departments of Education are turning more readily to the use of criterion-referenced, as opposed to norm-referenced, models for statewide assessment. The underlying assumption in this turn of events is that results generated by criterion-referenced tests within the statewide assessment context permit the drawing of valid inferences about the effectiveness of the educational curricula under study.

This paper examines the tenability of this assumption in the light of rigorous requirements for test construction and validation. Some tentative suggestions for overcoming certain obstacles to this process are offered, with the ultimate goal of underscoring the degree to which the usefulness of statewide assessment data depends on the development of content valid tests.

A Note on Definitions

While numerous definitions of criterion-referenced tests have been offered, the common denominator appears to be that such tests are intentionally constructed so as to yield information on the competence of individuals relative to specified instructional performance tasks (for example, Hambleton and Novick, 1973; Harris, 1972 ; Glaser & Nitko, 1971).

The salient distinction raised by some authors (for example, Hambleton et al., 1975) is that measurements that are directly interpretable in terms of specified performance standards need not be analyzed solely to permit mastery decisions. Given the purpose of the testing program, one of two analytic approaches may be most suitable: 1) the determination of a "level of functioning" score for each individual or what Millman (1974) has termed the "estimation of domain scores," and 2) the assignment of individuals to mastery states (e.g., masters and non-masters) based on a criterion cut-off or threshold score.

Most typically, statewide assessment programs adopt the "level of functioning" analysis (generally applied in terms of the average percentage of students answering correctly the items referenced to each performance objective). The ascription of mastery states has been generally bypassed due to difficulties in gaining consensus on specific cut-off scores and due to the problem of adequate numbers of items per objective discussed later.

Since the word "criterion" refers to that "minimal acceptable level of functioning that an examinee must achieve in order to be assigned to a mastery state" (Hambleton et al., 1975), it would seem that "criterion-referenced" is not the best modifier for the majority of statewide assessment programs. In the cases where mastery ascriptions are not a focus of the program, "domain-referenced" (Hambleton et al., 1975) or "objective-referenced" (Schooley, 1976) would seem more suitable and less misleading terms. These terms imply more directly the statewide practice of designing high-priority learning objectives each of which specifies (or attempts to specify) a domain of items from which a sample of items should be

selected. However, because "criterion-referenced testing" is the more familiar term within the context, the terms are used interchangeably here.

A Note on the Purposes of Statewide Assessment

It should not be assumed that all statewide assessment programs embrace only one purpose, or for that matter, even the same set of purposes. At a general level, all can be said to have at least one underlying goal the provision of appropriate information for decision making—which is the essence of any "good evaluation system" (Schooley et al., 1976).

Beyond this generic goal, however, the purposes are diverse. Reinstein (1976) has provided an excellent review of these purposes some of which are reproduced here: 1) developing state planning statements and priorities; 2) determining the extent to which students in a state have attained the skills, knowledge, and attitudes reflected in the educational goals of the state; 3) determining if students are acquiring "survival level skills" or "minimum competencies"; and 4) allocating state grants-in-aid to alleviate weaknesses in instructional programs.

Without de-emphasizing Reinstein's (1976) cautions regarding the implementation of these purposes, it may be seen that many of them correspond nicely to two of the major uses of criterion-referenced tests outlined by Millman (1974), including "needs assessment" and "program evaluation."

Millman's third use, "individualized instruction," only marginally applies; while some statewide programs include reporting results for individuals, the time frame usually permits only summative as opposed to formative evaluations. Millman's fourth use, "teacher improvement and personnel

evaluation," is most often avoided by statewide assessments primarily because of the political problems that obtain.

However, as other authors have pointed out (notably Hambleton et al., 1975), these different uses of criterion-referenced tests do not have differential implications for the construction of tests. Morgan et al. (1976) tend to disagree, and specify a set of evaluative criteria for determining the applicability to a given purpose of a test constructed in a given way. Nevertheless, those criteria which seem most relevant in the present context (subject area coverage, testing time, curriculum match, and stability/number of items per objective) seem to be subsumed in the comments of Hambleton et al., (1975) and will be discussed in detail later. The point is that defensible construction and content validation procedures must be observed regardless of the intended use of the criterion-referenced measurements. If such is not the case, and the resulting instruments are not content valid, then decisions based on the data generated are likely to be unjustified at best and wrong at worst.

The Developmental Process

In their excellent monograph on criterion-referenced testing and measurement, Hambleton et al., (1975) in close agreement with Fremer (1974), outline what they consider to be the major "domain-referenced test construction steps": 1) task analysis, 2) definition of content domain, 3) generation of referenced test items, 4) item analysis, 5) item selection, and 6) test reliability and content validity check.

The succeeding discussion focuses heavily on the first two steps since these activities are the most problematic within the statewide context

and treats more briefly steps three through five. The discussion is directed at the issue of content validity—the production of criterion-referenced tests that permit valid inferences about the curriculum under study. The determining factor is the degree to which the content of the tests may be justified on the basis of a well-defined content domain and a representative sample of items which permit generalizations to the domain. It is encouraging to note that it is the position of Hambleton et al. (1975) that, if the test development steps are carefully followed, the content validity of the tests should be guaranteed. This is critically important since the empirical validation technique suggested by Cronbach (1971) (computing a correlation coefficient for two *parallel* criterion-referenced tests constructed by separate teams on the basis of the same domain specifications) is financially beyond the means of most statewide assessment programs.

Step 1: Task Analysis

While this term is not commonly used in connection with statewide assessment, it legitimately refers to the process of defining the purpose and parameters of the test in terms of the subject area and domain to be assessed. In general, this process is implemented by an advisory committee representing a cross-section of the state's educators, administrators, and consumers of education (perhaps parents, students, or business people). The subject area to be assessed is generally mandated at the outset, and it is the task of the committee to specify in more detail the domain(s) which will define the scope and depth of the assessment instrument. In the experience of this author, this activity represents no small task.

The context problem. The major obstacle to the development of content-valid statewide assessment tests resides in the attempt to apply the criterion-referenced approach in a context for which it was not initially developed. The criterion-referenced model was pioneered for use in classroom management; that is, in the evaluation of learning outcomes relative to objectives for a specific curriculum with identifiable characteristics. Accordingly, it has been noted that criterion-referenced tests are generally administered before or after small units of instruction (Hambleton & Novick, 1973) and are most useful when used in a pretest—teach—posttest mode (Schooley et al., 1976). Difficulties arise when the model is applied, within the statewide context, to a diversity of curricula considered as one comprehensive program purely on the basis of the geographical boundaries of the state. Developing objectives, or domains, in this instance is not as straightforward a task as the one undertaken by a classroom teacher in the articulation of prescribed learning outcomes for a particular semester or year-long course.

Thus, the context problem stems from the need to treat all local-district programs within a given subject area as comprising one common curriculum and, therefore, to reflect in the tests being developed the diverse content of these programs. While, as Reinstein (1976) points out, the lack of congruence among local programs is "probably within acceptable limits for convention-based studies" such as reading and mathematics, incongruence is a major concern in other areas like science and social studies. In the latter case, a high degree of latitude is evident in the philosophies brought to bear on the task analysis and this causes consternation in the process of attempting to gain committee consensus. This author has observed

the penetration of varying philosophies into the more convention-based curricular areas (such as mathematics) as well, and concludes (as does Reinstein, 1976) that it is something of a problem in almost every subject area.

These difficulties are observed in the verbal attempts by committee members to find some common framework within which to evaluate differing philosophies in terms of their appropriateness to statewide assessment efforts. In essence, the committee seeks "guidelines" for formulating the content of the test.

One attempt to provide guidelines for the task analysis comes in the form of instructions to develop a test that reflects "survival skills" or "minimum competencies" within the subject area. One should be alerted to the problems to be encountered in attempting to define such concepts. Some practitioners have a tendency to simplify these concepts to the point of questionable usefulness by specifying a purely empirical definition (for example, a minimal competence is a performance which 90% of the current student population within a grade level are expected to have mastered). This strictly empirical approach is beset by theoretical difficulties in that the definition of "minimal competency" is subject to change from year to year based on the competencies existing in the population at a given time. While it is certainly possible and acceptable that a set of minimal competencies will change in a world that is characterized by changing demands on individuals, it is counterintuitive that the changes in definition should be based solely on the changing skill levels of individuals.

A more useful approach is suggested here: charge the advisory committee to identify those domains that are reflective of curriculum content

to which, in the committee's view, all students in the grade at the time of testing have been exposed. This does not imply, as noted earlier, that some large proportion has *mastered* the objective, only that the content has been widely taught. An immediate reaction may be anticipated— the complaint that, given varying curricular programs across schools, the approach is difficult, if not impossible, to implement. This author contends that, if the approach is used to provide *focus* to the committee, rather than to rigidly restrict the test development effort, it can provide a useful means for a "first pass" delimitation of the domain to be assessed.

Once this "common-ground" approach has been used to delineate domains, the committee may then apply additional guidelines to expand the coverage. There may be, for example, an interest in identifying additional domains that represent "ideal" outcomes. This interest is often a function of the transitional phases within the subject area (as, for example, in the "metric movement" in mathematics instruction of the intercurricular-concepts movement in social studies and career education). In these cases the committee is specifying a domain which the committee does not fully expect all students to have encountered, but which represents an ideal learning outcome. The inclusion of these "ideal" domains in the set may serve to generate 'baseline' data and/or to set a policy direction for high-priority curricular or program development within the state.

It should be pointed out here that a task analysis based on the "common ground" and/or "ideal" approach does not necessarily ensure that the complete set of resulting domains will match exactly the curriculum in every school in the state. In this regard, the approach may be open to charges of irrelevancy to local needs and goals similar to those originally levied

against norm-referenced tests which were discarded in favor of criterion-referenced tests. However, where local districts are using the results for their own evaluational purposes and where such "mismatch" occurs, data on irrelevant domains may simply be ignored by district personnel. Further, if state agencies are using the results to monitor local performance, emphasis should be placed on domains relevant to the local situation.

Committee members have been observed to achieve a high degree of consensus on the task analysis using the above approach (see, for example, the Connecticut Assessment of Educational Progress in Mathematics, 1976) in spite of the context within which they must operate. The outcome of this process generally takes the form of a topical outline. It is this outline, a list of domain descriptors (e.g., "Addition") or general behavioral objectives (e.g., "Possesses numerical skills useful in the world of work"), which forms the basis of the detailing process in Step 2.

Step 2: Definition of the Content Domain

Defining specifically the content domain of statewide assessment tests is equivalent to writing (or selecting) behavioral objectives. Given the time commonly available, it is beyond the means of the committee to specify content either via item generation rules (Bormuth, 1970; Hively et al., 1973) or via "amplified" objectives (Popham, 1974). The goal, then, is to produce a set of objectives, each of which is explicit enough to define the domain of items which may be legitimately referenced to it. This is important primarily to the content validity of the test, and secondarily to the need to specify clearly to local consumers of the test results the domains assessed.

The first problem is to determine the number of objectives to be identified or, as Popham (1976) puts it, "How large a chunk of learner behavior should be assessed by the test?" It is understood that an explicit objective, by definition, must display a certain degree of specificity. Given the level of specificity adopted, one or more (perhaps numerous) objectives may be identified for each of the domains in the topical outline. And certainly, for each objective identified, there must be "room" on the test for a "sufficient" number of items to permit generalization from performance on the item set to performance on the objective

Since there are rather severe time limitations on statewide assessment tests, the committee must deal with the trade-off between the number of objectives that can be assessed and the number of items per objective that can be included. Given the task of assessing a subject area, committee members tend to be highly concerned with subject *coverage*, and in spite of time constraints, tend to resist limiting the knowledge or skills covered by the test. Unfortunately, they tend, therefore, to reduce the specificity of objectives in order to widen the domain of (types of) items which may be matched to them. This practice violates Popham's (1976) rule that the magnitude of behavior(s) assessed should not sacrifice the test's descriptive clarity.

It is strongly recommended that committee be apprised of this problem and urged to limit the test to a set of high-priority objectives characterized by sufficient explicitness and specificity. This does not imply that the objectives finally selected must be so explicit as to be trivial (see Ebel, 1971), but rather, that they must be narrow enough to focus on

a restricted range of item types. Neither does it imply that the committee does not recognize excluded domains to be important in the general sense; rather, it suggests that the domain has been limited to maximize the usefulness of the test for decision-making purposes.

To guide the committee through this content definition activity, it is recommended that the parameters of the test be considered at the outset. That is, the committee must consider the time allocated for testing and the total number of items which can be administered within that time. (Given conventional multiple-choice items, one minute per item is a useful rule of thumb; where items are unusually long, as in reading comprehension, or where open-ended items are involved, time per item must be adjusted accordingly.)

The next step is to set a minimum number of items per objective. Unfortunately this number usually cannot be one that is ideal in terms of ensuring test reliability. Hambleton et al. (1975) indicate that some number less than 25 items per domain is recommended, while Popham (1976) suggests that, in order to reliably assess a domain, the number of items should "more than likely be between 10 and 20 than between one and five." These guidelines would restrict a one-hour conventional multiple-choice test to the measurement of between three and six objectives—a situation that most statewide committees would find difficult to live with. A common practice is to adopt a minimum of four items per objective (which meets the minimum for stable reliability estimates set by Schooley et al., 1976.

If a minimum of four items per objective is established, the test described above could contain up to 25 objectives, but more likely (given

varying item lengths) would contain something closer to 15 objectives. Recall that the issue is to delimit the number of objectives which the committee may select for the test. Once this limit is determined, given the test parameters, the committee can be guided to write or select objectives that are specific enough to be measured by a sample of only four items. Clearly, this gauging of the appropriate level of specificity is, at present, an intuitive process. In practice, committee members exercise their individual intuitions, achieve a surprising degree of agreement on level of specificity when the issue is clearly understood. If they are urged to consider the range in type and number of items which would be subsumed by a given objective, they are able to identify those objectives which are too general or broad to be of use.

These suggestions serve only as a practical guide for completing the most difficult step in the test production process. The author admits that the suggestions perhaps may be more practically-sound than theoretically sound. However, given the current state of the art of developing criterion-referenced statewide assessment tests, they may serve to bring us one step closer to the production of fully valid and reliable tests.

Step 3: Generation of Referenced Items

Many statewide assessment programs involve the generation of an item pool for each specified objective through a search for existing materials. One problem that often arises with this approach is that a sufficient number of items appropriately matching each objective cannot be located or obtained. The question remains whether the perceived dearth of materials is a result of time constraints which make unfeasible a

comprehensive search, or whether extensive materials are not yet available in the field. This problem is compounded when a committee has elected to write original and rather uniquely-defined objectives. Often, after reviewing the available items, the committee is forced to reevaluate their objectives, and perhaps, to rewrite them.

This activity may, in fact, be valuable since it can result in the refinement of the objectives and focuses attention on the need for objective/item congruence. What should be avoided is the tendency to "cling to" the phraseology of the objective and to permit "slight deviations" in the types of items included in the matching pool. This tendency to create an item pool where truly none exists defeats the purpose of criterion-referenced assessment and results in tests that are limited in content validity and, therefore, usefulness. Where a sufficient item pool is unavailable, objectives should be redefined or discarded.

Where time constraints are not an issue (as, for example, in those programs which include at least a full year developmental phase prior to actual testing), generation of original items tends to follow conventional guidelines. It is encouraging to note that, in these cases, item writers are sometimes provided with either amplified objectives or item prototypes (see, for example, the Ohio Statewide Student Needs Assessment, 1976-77). Highest productivity is achieved when the item-writing team has the opportunity to interact with the objectives-writing team since objective/item congruence is then maximized. Schooley et al. (1976) suggest that the item and objective teams should be one and the same. This is sometimes the case in statewide assessments (see, for example, the Missouri Statewide Assessment, 1976), but is not a frequent occurrence.

Any of the above approaches (separate teams, interacting teams, or the combined team approach) is fully workable, given the production of objectives whose substance can be clearly agreed upon.

One additional approach, however, is not recommended: the generation of item pools without a corresponding predetermined set of objectives. Some statewide committees who use the "available materials" method have adopted this approach due to perceived difficulty in gaining consensus on objectives at the outset. Rather, the committee reviews all of existing items which can be located, and determines for each one whether or not it is appropriate for statewide assessment. Here, the committee members are using an internalized, but unarticulated, set of standards to identify the item pool. For some reason, they find it easier to achieve consensus on individual items than on objectives. Once the item pool is generated, they then return to the bypassed step of domain specification and write objectives based on the items identified. This approach is not recommended because it generally results in a pool of items that cannot be well justified in terms of objective and curriculum coverage.

Step 4: Item Analysis

This step, the procedure of checking the quality of the items, applies primarily in cases where original items are produced for the statewide tests. Where existing items are used, they were previously used. Where new items are written for the assessment, this author has noted the use of many of Hambleton et al.'s (1975) procedures for determining the extent to which items reflect their respective content domains. These include content specialist ratings, item difficulty and item discrimination indices.

The only method not observed is that involving "item change statistics," since statewide assessments do not involve testing before *and* after instruction.

It is encouraging that item analysis techniques are commonly being used in connection with statewide assessment tests. What is less encouraging is the fact that these techniques are being implemented without reference to available statistical procedures. Ratings by content specialists (for example, on a four-point relevancy scale for each item), item difficulty indices (for example, the percentage of students scoring correctly on an item in a field test situation), and item discrimination indices (for example, the proportions of students in "high" and "low achieving" groups correctly answering each item) tend to be evaluated on a visual scanning basis. That is, for example, if the range in difficulty level across the set of items referenced to an objective *looks* too wide, "deviant" items are deleted from the pool. There is little evident use of statistical procedures suggested by Hambleton et al. (1965): 1) Cohen's (1960) coefficient kappa to measure the agreement between ratings of items made by different content specialists, or 2) Cochran's Q test to determine whether item difficulties are equal.

The reliance on visual scanning methods as opposed to statistical techniques suggests that practitioners have not yet taken advantage of recent developments in data analytic procedures for criterion-referenced tests. This may reflect the traditional gap between theory and practice, but nevertheless, should be corrected if statewide assessment tests are to approach an optimal level of validity.

Step 5: Item Selection

It has been contended that "strong" criterion-referenced interpretations of test scores are made possible only by a *random selection* of items from the domain (Hambleton et al., 1975; Millman, 1974). It is here that statewide assessments encounter the most difficulty in producing tests that are useful to the decision-making purpose. In assessments that focus on identifying *existing* test items, it is frequently difficult to locate enough matching items for each objective from which to randomly select. If randomly selecting four out of five existing items truly qualifies as random selection from the domain, then the problem is resolved. However, it is unclear whether such limited randomness permits valid generalizations from the items to the domain. Where larger numbers of matching valid items are available, it is a straightforward matter to randomly select from the pool.

In assessments that involve production of original items, the random selection requirement implies that a greater number of items than will actually be used are to be written. While in practice, this is often the case, the number of items produced rarely exceeds the number required for the test by more than four or five due to the time and expense involved in item writing. This "over production" of items is generally not intended to permit later random selection, but rather to allow for the deletion of items that, on the basis of item analysis, do not prove to be valid indicators of the domain. If useful tests depend on the random selection process, then increased funding must be made available to permit the generation of larger numbers of items.

It is important that committee members understand the necessity of random selection in order that they do not insist on selecting the items they "like best." The most useful and practical approach is to instruct committee members to review all available valid items and to identify those that are "acceptable" for statewide assessment purposes. In theory, all valid items should be acceptable; however, certain eccentricities prevent the translation from theory to practice. The review for acceptability will result in a restriction, usually minor, of the item pool, but should allow the random selection process to be implemented without complaint from the committee.

Number of items per objective. Given that the number of items for each objective must meet the minimum required for reliability, the actual number selected may be constant across all objectives or, alternatively, may vary across objectives. Regardless of which alternative is adopted, the choice should reflect some theoretical rationale, as opposed to merely the number of items is constant across objectives. This implies that in the committee's estimation, the objectives are of equal importance. If the number of items vary across objectives, they should vary in terms of the relative importance ascribed to each objective.

Since results are almost uniformly reported in terms of the proportion of items for each objective answered correctly, a constant number of items across objectives tends to increase the ease with which results can be meaningfully interpreted. In a sense, this implies to users of the results that the proportional results can be given equal weight. Popham's (1976) comment on "behavioral homogeneity" may be useful here. If all objectives can be designed to reflect approximately equal amounts of

"required instructional time," then selecting equal numbers of items across objectives becomes particularly defensible. If this is not possible, then the differing number of items per objective should reflect some other criterion the committee is using regarding the relative importance of the objectives. Further, the number of items per objective should be reported along with the average scores to increase data meaningfulness.

Step 6: Test Reliability and Validity

The issues involved in determining test reliability are not treated here since the focus is primarily on the content validity of criterion-referenced tests. It should be sufficient to note here the position of Hambleton et al. (1975) that if the foregoing steps in test construction are followed closely, then the content validity of the tests should be ensured. The previous discussion has highlighted problem areas in each step that are encountered in the context of statewide assessment. Only to the extent that these problems are overcome by appealing to the guidelines suggested will the resulting tests be content valid and useful for decision making. It would, of course, be desirable to check the content validity of the tests through the use of Cronbach's (1971) techniques for test construction described earlier. However, this procedure continues to seem beyond the means of most statewide assessments. In the absence of such validation procedures, following closely the guidelines for valid test construction becomes all the more important.

Summary

Underlying the criterion-referenced test construction procedures outlined in this paper is the need for allocating sufficient time and resources to the developmental process. This need, raised by Reinstein (1976) in connection with criterion-referenced test development at the local level, is magnified in the context of statewide assessment. In the press to shift from norm-referenced to criterion-referenced testing at the state level, the need to make available increased and adequate time for test development is too often overlooked. The task of selecting and ordering an existing norm-referenced test is far less awesome than the task of developing from "scratch" a valid criterion-referenced instrument. Some states (e.g., Minnesota and Rhode Island) have found that a two-year timeframe is required to permit the implementation of valid construction methods, while other states have required that the process be completed within three to six months. Committees that are severely restrained in terms of time and resources available cannot be expected to produce something other than a hastily produced test that cannot be justified in terms of curriculum coverage or content validity.

The extent to which the test construction steps can be followed closely to yield a content valid test determines the extent to which the tests can be justifiably used to evaluate the curricula or programs under study. From a decision-theoretic point of view, the scores of students on the tests are used to make decisions about the performance status of individuals and/or the effectiveness of statewide curricular programs. If the content of the test does not adequately reflect the behaviors legiti-

mately inferrable from those delimited by the criteria (Popham & Husek, 1969), then decisions based on the results are likely to be erroneous.

Teachers have every right to expect that, if evaluations of the learning outcomes of their courses are to be made, the evaluations must be made on the basis of inferences that are well-founded. Teachers as well as other local consumers of statewide test results are very sensitive to the content validity of the tests. They frequently make accurate judgments as to the importance of the objectives selected and the "goodness-of-fit" of the items referenced to each objective. Where the tests are weak in these respects, results tend to be disregarded for local purposes and the evaluations or recommendations made by state agencies on the basis of test results tend to be ignored. While students have an equal right to be evaluated on the basis of sound instruments, they rarely have an opportunity to reject the conclusions drawn on the basis of their test scores.

In summary, it is to be concluded that the content validity of measuring instruments must be carefully established in order to ensure meaningful and defensible decision making. The risk involved in using an invalid test must be judged in terms of the costs (psychological, financial, etc.) attendant on making erroneous decisions in a given situation.

REFERENCES

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Finley, C. J., & Berdie, F. S. The National Assessment approach to exercise development. Ann Arbor, Mich.: National Assessment of Educational Progress, 1970.

Fremer, J. Handbook for conducting task analyses and developing criterion-referenced tests of language skills (ETS PR 74-12). Princeton, N.J.: Educational Testing Service, 1974.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Symposium presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Hively, E., Maxwell, G., Rabehl, G., Senison, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: University of California, Center for the Study of Evaluation, 1973.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, Calif.: McCutchan, 1974.

Morgan, P., Kosecoff, J., Walker, C., & Keesling, J. W. It's the metric that counts, or criterion-referenced schizophrenia. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Popham, W. J. Expanding the technical base of criterion-referenced test development. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Reinstein, B. J. Public school perspectives on the uses of large-scale testing programs. In Dissemination and utilization of large-scale test results. Symposium presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

Schooley, D. E., Schultz, D. W., Donovan, D. L., & Lehman, I. J. Quality control for evaluation systems based on objective-referenced tests. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

Womer, F. B. What is National Assessment? Ann Arbor, Mich.: National Assessment of Educational Progress, 1970.