DOCUMENT RESUME

ED 109 222 TM 004 683

AUTHOR Klein, Stephen P.; Kosecoff, Jacqueline P.
TITLE Procedures and Issues in the Validation of Criterion-Referenced Tests.
PUB DATE 75
NOTE 10p.; 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 Academic Achievement; Class Management; *Criterion Referenced Tests; Curriculum Development; Instruction; *Predictive Validity; Student Evaluation; *Test Validity

ABSTRACT

Four common uses for criterion-referenced tests (CRT) are outlined: describing student achievement, improving curriculum development, being sensitive indicators of the effects of instruction, and facilitating classroom management decisions. These uses parallel various forms of empirically establishing the content, concurrent, and predictive validity of CRT. It is found disconcerting that the developers of CRTs have generally not conducted such validity studies, or at least they have not reported on them in the technical manuals for the CRT systems. Proof of their utility is called for. (Author/DEP)

Procedures and Issues in the Validation of Criterion-Referenced Tests

Stephen P. Klein
Jacqueline P. Kosecoff

When one examines the literature on criterion-referenced tests, which we will refer to as CRTs, one quickly notices two phenomena. First, no one agrees as to what a CRT is, precisely, in relation to objectives-based tests or domain referenced tests, the latter being abbreviated by the term DRTs (which may have to do with what some people think of them). Second all of the relevant articles essentially deal with such issues as how the items should be constructed (Bormuth, 1970; David & Diamond, 1974; Hively, 1973; Popham, 1974); how many of them are needed for one to say that an examinee has "mastered" an objective (Millman, 1973; Novick & Lewis, 1974); and how one should determine their reliability (Harris, 1975). Very few researchers or publishers have seemed to concern themselves with determining whether CRTs have empirical validity.

One reason that most CRT developers ignored empirical validity is that they seem to assume that the CRT construction process itself will almost automatically lead to a content valid test. The typical construction process usually involves identifying a set of supposedly important objectives and then developing "good" items to measure these objectives. The "goodness" of an item is determined by two factors: (1) consistency of the item with the objective it is supposed to assess; i.e., does the item measure the intended objective or some other objective? and (2) technical quality of the item, i.e., is the item free of technical flaws and biases? If the items in a CRT are selected so as to assure that the specified domain or objective is adequately covered then we are supposed to believe that the CRT itself is "good" and useful for a wide range of educational decision making. We will

A presentation in a symposium at the Annual Meeting of the National Council on Measurement in Education, March 31 through April 3, 1975, Washington, D.C.

leave the issue of what constitutes "adequate coverage" of a domain for another time. Needless to say, what constitutes adequate coverage is not immediately apparent either from the literature (Klein & Kosecoff, 1973) on the subject or an in-depth inspection of available CRTs (Kosecoff & Klein, 1975).

To summarize, then, a CRT is generally considered to be valid if its items cover the specified domain or objective, if the items are judged to be appropriate members of the sample of items that might be constructed to measure that objective in terms of the kinds of skills and content assessed, and if the items are free of technical imperfections - a condition which is by no means guaranteed even by the most rigorous of the current item-writing rules (Skager, 1975). The foregoing criteria may lead to the development of a test that is content valid in that scores on it describe an examinee's skills or knowledge. This information, however, is generally not directly useful to most users of CRTs, because most CRTs are administered in order to gather information that will be used in a wide array of educational decision making (Klein, 1970). Such decisions might deal with classroom management problems of assigning students to groups for instruction or deciding whan a student or group of students is ready to progress to the next major unit of instruction. CRTs may also be used for evaluating the effectiveness of educational programs and determining the kinds of curriculum that should be provided to students. In other words, users of CRTs want to be sure that the tests really provide valid information for making these important decisions.

The remainder of this paper will consider four features of CRTs. These features represent some of the supposed major advantages that CRTs have in comparison to norm-referenced tests (Popham, 1971). For each of these features, we will outline empirical procedures that one might use to assess the

extent to which a given CRT actually contains these desirable characteristics.

Describing Student Achievement

One advantage of a CRT is that it is supposed to provide a clear description of what the student does or does not know or what the student can or cannot do. Student mastery of a given objective is supposed to be meaningful in and of itself. The CRT accomplishes this by being based on a very specified objective or set of objectives, and, all the items on the CRT, presumably, are indicators of the extent to which the student has or has not mastered that objective. In other words, the items are consistent or congruent with the objective.

One way of ensuring such consistency is to have expert judges independently evaluate each individual item to determine whether it actually belongs with a given objective or with some other objective (Dahl, 1971). This determination could be made by having the judges sort all the items for a variety of CRTs according to the list of objectives that was used in developing these measures. A better technique, however, would be to have the judges form their own clusters of items and then see whether these clusters correspond to the initial set of objectives. Alternately, one could have judges infer the objective from an item in terms of the kinds of skills and content knowledge that would be required to answer that item. This inference should closely correspond to the original objective on which the item was based. Finally, the construction process itself might be validated by having two teams of item writers develop items. Judges would be given the items written by both teams, in a completely scrambled fashion, and then be asked to perform the kinds of tasks noted above in order to

determine item-objective consistency. If the development procedure is appropriate, then the two sets of item writers should produce comparable items in the sense that the judges do not differentiate between them.

The foregoing techniques all require expert judges, but as many of us have learned, judges are sometimes not as expert as we believe. It may be necessary, therefore, to use actual student response data to insure that an item is indeed measuring the objective for which it was intended. This could be done by using a sample of students who vary in their levels of performance with respect to a variety of CRTs. Factor analyses of these data would indicate whether the items in a given CRT correlate more highly with each other than they do with items in other CRTs. If they do not, then one would have serious questions about the viability of the CRT as being a good measure of a well-defined objective. Before one believes all the propaganda about the value of CRTs for describing student achievement, then, one should be certain of the content validity of those CRTs as established by empirical data.

Curriculum Development

A second supposed advantage of CRTs is that they operationally define important en route or component objectives that must be mastered in order for students to achieve some desirable goal. Teachers and evaluators can use CRTs, therefore, as a means for monitoring student progress towards the achievement of this goal. The importance of such objectives, however, is generally established by theory and opinion rather than on the basis of empirical data.

To the authors' knowledge, there is only one study that has attempted to establish the importance of an objective as operationally defined by the

CRT that was used to measure it. In this study, McNeil (1975) divided a sample of students into two groups - those who could read a series of passages aloud essentially without error and those who could not perform this task. He then compared the performance of the two groups on a series of CRTs that presumably assessed the component skills needed for performing this criterion task of reading the passages. McNeil found that only a few of the CRTs were able to discriminate between the two groups. On the basis of these results, he concluded that it may not actually be necessary to teach certain objectives in order for a student to perform certain criterion tasks that are considered important in themselves.

It is apparent that McNeil used a concurrent validity model to determine the relative importance of certain objectives. In so doing, he also validated the relevance of the CRTs he used to measure those objectives which were deemed necessary for goal attainment. In other words, the fact that a given CRT was able to make the necessary discriminations between those who did versus those who did not master the goal indicates that performance on that CRT was relevant to that goal.

There are a few problems in the McNeil study that other researchers should be aware of before they try to replicate its approach. One problem is that a student who has mastered the goal may have forgotten how to perform some of the en route tasks that were required as part of the learning process. For example, of those adults who use good grammar, how many of them are still able to diagram sentences properly? Further, it is also possible that goal attainment could be achieved in a variety of patterns or that the criterion measure of the goal itself is faulty. While these problems are not easy to resolve, it would be well worth the effort especially

considering how much time is now spent on instructing students so that they can pass a group of CRTs whose importance is based on conjecture.

Sensitivity to Instruction

One of the most highly touted advantages of a CRT is that it is sensitive to the effects of instruction. Teachers are told by program evaluators that they no longer have to put up with test questions that are not germane to the particular instructional objectives they are trying to get their students to achieve. Thus, it is fair to use CRTs for assessing program outcomes.

There are two models for empirically establishing sensitivity to instruction. The first model focuses on whether specific items within a CRT differentiate between those who have versus those who have not mastered the objective, after they have had instruction in the area to be covered by the test. An item that is sensitive is one that students fail prior to instruction and pass after instruction (Kosecoff & Klein, 1973). The second model focuses on whether the CRT itself is sensitive to instruction in the sense that students who receive instruction perform better on the CRT than students of comparable ability who do not receive such instruction.

If a CRT fails to show the necessary sensitivity in one instance, one could argue that it was the fault of the instruction and not the test. But if this pattern occurs frequently, one should question the validity of the CRT itself in terms of its being sensitive enough to detect instructional outcomes for such purposes as program evaluation.

Classroom Management

The fourth major supposed advantage of CRTs is that they are useful for classroom management, especially where some form of individualization of in-

struction is in use. For example, the curriculum may be organized so that essentially all students proceed through the same sequence of objectives - such as steps in a given strand of mathematics - but they do so at their own rate. In this context, CRTs are presumably the ideal tool for checking on whether a student is "ready" to move on to the next step.

The assessment of a CRT's utility for making these kinds of progress decisions would involve examining its predictive validity. This could be done by measuring the extent to which students who passed or mastered the CRT actually performed better in a subsequent instructional unit than students who did not pass (Keesling, 1974). Such performance would be indicated by test scores in the subsequent unit and/or by the time it took the student to master its objective and/or by other relevant indices of competence. One important side benefit of this kind of validation study is that it provides an empirical basis for setting mastery levels on CRTs. In other words, "mastery" could be operationally defined as that performance level at which one has essentially eliminated such potentially costly classification errors as saying a student has mastered the objective when he has not.

A predictive validity model could also be used in situations where CRTs are employed for grouping students for instruction in the sense that all students may not receive the same set of objectives and/or at the same rate and/or in the same order. The issue under investigation would again be the ability of the CRT to make the relevant distinctions between student performance levels. For example, if CRTs really facilitate the forming of effective groups, then the subsequent overall performance of the classes in which grouping occurs should be better than in those classes that do not use CRTs for this purpose.

In short, if CRTs are truly useful for making classroom management decisions, then this advantage should be reflected in the performance of students. While this may not happen because of other extraneous factors, in every instance in which the CRT is used, there should at least be some indication of its utility when one examines its effectiveness across a variety of sites.

Summary

In this paper, we have outlined four common uses for CRTs: describing student achievement, improving curriculum development, being sensitive indicators of the effects of instruction, and facilitating classroom management decisions. These uses parallel various forms of empirically establishing the content, concurrent, and predictive validity of the CRTs. What is disconcerting, however, is that the developers of CRTs have generally not conducted such validity studies, or at least they have not reported on them in the technical manuals for their CRT systems (Kosecoff & Klein, 1975). It is time, therefore, for those of us who believe in the value of CRTs, to start providing proof of their utility for the tasks we claim they can perform, just as we have required such evidence from the developers of norm-referenced tests.

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

Dahl, T.A. The measurement of congruence between learning objectives and test items. Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Davis, F.B. & Diamond, J.J. The preparation of criterion-referenced tests. CSE Monograph Series in Evaluation, Volume 3. Center for the Study of Evaluation, University of California, Los Angeles, 1974. Pp 116-138.

Harris, C.W. Survey of techniques for analyzing reliability: Criterion-referenced measures. Paper presented at the American Educational Research Association Annual Meeting, Washington, D.C. March 31, 1975.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluations: A technical handbook and a case study from the MINNEMAST project. CSE Monograph Series in Evaluation, Volume 1. Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Keesling, J.W. Empirical validation of criterion-referenced measures. CSE Monograph Series in Evaluation, Volume 3. Center for the Study of Evaluation, University of California, Los Angeles, 1974. Pp 159-176.

Klein, S.P. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2(2), 1-6.

Klein, S.P. & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 24, 1973. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Kosecoff, J. & Klein, S.P. A review of currently available criterion-referenced assessment systems. Paper presented at the American Educational Research Association Annual Meeting, Washington, D.C. March 31, 1975.

McNeil, J.D. False prerequisites in the teaching of reading. Journal of Reading Behavior, 1975, 421-427.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-216.

Novick, M.R. & Lewis, C. Prescribing test length for criterion-referenced measurement. CSE Monograph Series in Evaluation, Volume 3. Center for the Study of Evaluation, University of California, Los Angeles, 1974. Pp 139-158.

Popham, W.J. (Ed.) Criterion-referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Popham, W.J. Selecting objectives and generating test items for objectives-based tests. CSE Monograph Series in Evaluation, Volume 3. Center for the Study of Evaluation, University of California, Los Angeles, 1974. Pp 13-25.

Skager, R.W. A classical scheme for assessment instruments: A theoretical conception. Paper presented at the American Educational Research Association Annual Meeting, Washington, D.C. March 31, 1975.