DOCUMENT RESUME

ED 110 496 TM 004 769

AUTHOR Harris, Chester W.
TITLE Techniques for Analyzing Test Response Data.
PUB DATE [Apr 75]
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; *Instructional Programs; Item Banks; *Item Sampling; *Response Style (Tests); Statistical Analysis; *Test Construction; Test Interpretation; Test Reliability

ABSTRACT

Achievement tests which are specifically linked to an instructional program and have been developed in relation to an objectives base and/or to an item generation rule are considered, as well as student response data. Three types of studies are outlined and the kind of procedures thought useful illustrated. As various methods for examining stability, equivalence, and sensitivity to instruction of both test items and test scores are examined, an effort is made to coordinate sampling procedure and experimental design, choice of statistic, and method of aggregating the statistic so as to provide generalizations for the pool of items or the universe of test scores. (BJG)

Techniques for Analyzing Test Response Data

Chester W. Harris

Center for the Study of Evaluation, UCLA
and
University of California, Santa Barbara

Paper presented at the AERA annual meeting, Washington, D.C., April 1975.

I wish to emphasize at the outset that I am considering only the domain of achievement tests, rather than other domains such as those of personality tests and of intelligence tests. Further, I am considering only achievement tests that are specifically linked to an instructional program and have been developed in relation to an objectives base and/or to an item generation rule. These may or may not be criterion-referenced tests, dependupon the definition employed for that term; however, they quite likely are tests for which there is an interpretation of a particular student's obtained score that does not depend on knowledge of the scores of any other student. I think of mastery tests as falling into this category, and I have discussed some technical characteristics of such tests in one of the Center's monograohs. (Harris, 1974)

Today I wish to report on an inquiry which is now underway but not completed. The inquiry is an attempt to examine the grounds and methods for studying student response data to the type of test I am considering. Such study of student response data is intended to throw light on the complex of instructional programs plus test development and interpretation. This differs from the typical practice of finding numbers to be used to choose items from an undefined or accidental pool of items on the grounds that such numbers mean that these items will work well in a particular sample of students whose instructional history is notknown or possibly not considered relevant.

Let us assume that, for the type of test that I am considering, a pool of items has been carefully conceptualized and constructed to represent the behaviors that the instructional program is designed to foster and that rules have been developed for sampling this pool of items in such a way as to yield aggregates of items for which one or more instructionally relevant scores can be developed. It seems reasonable to require that such sampling have a random character but it may of course operate within stratz or cells. Such a base defines a universe of items and a universe of test scores based upon appropriate samples of these items. Let us further assume that we would like to use student response data to study both the instructional program and the test development and test interpretation process. We have identified several types of studies that seem to be fruitful; not surprisingly, some of these studies are rather standard ones. I shall outline three types of studies and for each one illustrate the kind of procedure we believe will be appropriate for the purposes described above.

The notion of stability, which can be related to the concept of specific reliability, is of importance. What we would like to have is an estimate of: an appropriate stability coefficient for an item and a coefficient for a score from which one could describe generally this characteristic of the pool of items and of the universe of test scores that can be derived from these items. Let me illustrate at the level of the item. If, for a population of students whose instructional history has been controlled, an item varies markedly in difficulty (normative difficulty for this population) over administrations separated by brief time periods during which no additional instruction is given, then we have evidence that the combination of instruction and test development process yields undependable item data. Such a finding for a random sample of the items would be grounds for reworking the instruction and the test development process, with the hope of finding clues as to why the items behaved so badly. There would be several places to look. For example the item type or format may be so unfamiliar to this population of students as to introduce a factor of learning that systematically makes the item easier on the second trial.

One can use McNemar's chi square procedure (or the underlying exact procedure that employs the binomial) to test the hypothesis that the difficulty of the item is the same for the two administrations. This is very useful, and it is a proper test for this purpose; the more familiar use of chi squire for a test of independence seems to me to be all wrong here. But the McNemar test probably isn't sufficient. We would also like an estimate of the common difficulty level of the item and we would like to be able to aggregate such estimates to secure a meaningful index to describe the pool of items on the basis of a study of a sample of these items. We would also like an estimate of the degree of association (or some aspect of association) that can be aggregated in a similar manner. We are now exploring a statistic devised by Lazarsfeld and Kendall and reported by Goodman and Kruskal (1959, P. 149-150), using some Monte Carlo methods to examine its sampling distribution. In time we will know whether or not to recommend it.

A second important notion is that of equivalence, which can be related to the concept of generic reliability. I illustrate again with the case of the study of a pair of items. In such a study one may or may not expect the two item difficulties to be the same for a specified population of students for whom the instructional history has been controlled, and so a test of the hypothesis of identical difficulties may or may not be informative. Even if such a test is informative, however, one would like estimates of difficulty for the two items and some measure of association that might be meaningfully aggregated. We have found important leads in Goodman and Kruskal (1959) and in the fairly new volume by Fleiss (1973) and are looking into sampling characteristics for these measures. For both stability and equivalence item studies, the appropriate sampling design is Fleiss' Method I.

A third type of study is that of sensitivity to instruction. If the instructional program is effective ana if the test development process has yielded items and test scores that measure the outcomes of the instruction adequately, then one expects that the items and/or the test scores will be sensitive to instruction. If they are not, then again something is wrong and one must begin a search for the defect, which may be in either or both the instruction or the test development. In studying sensitivity to instruction of an item, more than one experimental and sampling design is available, and the statistic one would employ to measure sensitivity to instruction may differ with the different designs. If we choose a sample of students to whom we administer the item, whom we then teach, and to whom we then readminister the item, we have fixed the total sample size but not the marginals in the two-by-two table, and we have introduced an experimental manipulation that is intended to change the difficulty of the item. With such a design the usual chi square test of independence and the related phi coefficient are inappropriate. Instead, one would like a measure of the amount of change attributable to instruction, and this can be derived from the appropriate conditional probability which can be estimated by determining the proportion of those who failed the item on the first administration who passed it on the second administration. It also is possible to introduce a model of measurement error for the responses and develop a modified estimate of this conditional probability corrected for measurement error.

As we study the various methods that have been suggested for examining stability, equivalence, and sensitivity to instruction of both test items and test scores,we are attempting to coordinate three things: sampling procedure and experimental design, choice of a statistic, and method of aggregating the statistic so as to provide generalizations for the pool of items or the universe of test scores. We hope to have a number of specific results that can be summarized in a forthcoming issue of the Center monograph series.

REFERENCES

Fleiss, J. L. Statistical methods for rates and proportions. New York, John Wiley, 1973

Goodman, L. A. and Kruskal, H. Measures of association for cross classifications, II. Journal of the American Statistical Association 1959, 54, 123-163.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, Alkin, M. C. and W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, Center for the study of Evaluation, University of California, 1974.