DOCUMENT RESUME

ED 103 577 CE 003 088

AUTHOR Pollack, Robert M.
TITLE Design for Reliability and Validity Testing of Assessment Instruments--Phase 2.
INSTITUTION Mountain-Plains Education and Economic Development Program, Inc., Glasgow AFB, Mont.
PUB DATE [74]
NOTE 9p.; For other documents describing aspects of the Mountain-Plains program, see CE 003 082-091

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Affective Tests; *Attitude Tests; *Cognitive Tests; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS Mountain Plains Program

ABSTRACT

Plans are outlined for developing measurement techniques and testing programs to determine the success or failure of the program objectives in the cognitive and attitude areas of the Mountain-Plains program. In the cognitive area a criterion-referenced approach will be used, in spite of problems associated with reliability. Item pools will be developed at the Learning Activity Packages (LAP) level. Then unit tests will be developed and matched with a performance test which will be scored dichotomously. A phi coefficient will be computed to establish the degree of correlation between the two tests. In those cognitive areas for which performance tests would be difficult or impossible to develop, attitude scales will be administered before and after classroom work. Finally a comprehensive diagnostic test will be developed to be used as a pre- and posttest, and validated through the performance test at the unit level. Attitude scales will be developed for the testing program in the affective area. Initial items should be administered to all applicants to the Mountain-Plains program, thus providing scores for students and controls. Students should be retested when they leave, controls after the same time interval, and followup retesting carried out later on both groups. (SA)

Mountain-Plains Education & Economic Development Program, Inc.

POST OFFICE BOX 3078 · GLASGOW AFB, MONTANA 59231 · TEL: (406) 624-6221

DESIGN FOR RELIABILITY AND VALIDITY TESTING OF ASSESSMENT INSTRUMENTS – PHASE II

Phase I - First generation tests constructed and implemented 12/31/73 for purposes of program formative evaluation.

Phase II - a. Restructure of assessment instruments to meet requirements for refined validity testing, complete 3/31/74.
b. Test validity assessment, start 4/1/74.

Submitted to:

Mr. Harold Johnson
NIE Project Officer

Author

David A. Coyle
Director of
Research Services

Robert Pollack
Test and Measurements
Specialist

Design for Reliability and Validity Testing of Assessment Instruments - Phase II

The Mountain-Plains project is essentially concerned with two areas of human functioning. One is the cognitive area which involves the imparting of vocational skills to the student population. The other involves changing the attitude of the student population with regard to such areas as child raising, home management, and the desirability of maintaining steady employment, etc.

Since these two areas are extremely divergent, the measurement techniques employed to determine the success or failure of the program objectives in these areas will necessarily be quite different. As a result, essentially two testing programs must be developed and implemented. One for the cognitive area, and one for the affective or attitudinal area.

Cognitive Area Testing Program

The Mountain-Plains project has been developed using a curriculum based on individualized instruction. The goal has been to impart a certain body of knowledge to each student without reference to the capabilities of other students in the same program. Consequently, the proper model for developing tests in this area is the "criterion-referenced" approach.

Unfortunately, all the theoretical work done in the field of measurement has been based on the normative approach to testing. Although the "Domain-sampling" model[1] is the basis for the development of normative testing instruments, and it's applicability to the criterion-referenced approach is debatable, I feel that by adhering to this model as closely as possible, we will be able to develop superior instruments.

The major problem with criterion-referenced instruments is with the concept of reliability, since they are essentially dichotomous in nature, the subject has either met the criterion (passed) or not met the criterion (failed). Coefficient alpha cannot be used as a measure of reliability because it is a function of the ratio of item variances to the variance of the test scores. Since everyone familiar with the material should be able to answer all the questions correctly, the item variances will necessarily be small and the inter-correlations between items, high. Also, since the tests are scored dichotomously, the test variance will be small. This combination of factors would probably make the result of computing coefficient alpha uninterpretable to say the least.

Test-retest reliability would also not apply as a meaningful concept, because theoretically everyone who meets the criterion on the initial testing would continue to do so, no matter how many times he took the test. However, those who did not meet the criterion on the first administration, would be expected to do so some time in the future.

---

[1]For a complete discussion of the domain-sampling model, see Nunnally 1967, Chapter Six.

Consequently, I do not believe there is any meaningful concept of reliability with regard to criterion-referenced instruments.

The overall design of the cognitive area testing program will originate with the development of item pools at the LAP level. The item type will consist of objective multiple choice items. The reasons for using objective items are: 1) that they guarantee high inter-scorer reliability; 2) a wider range of material can be sampled with them than with essay type items; 3) objective tests are generally more reliable than other types of tests; and 4) they fit the domain-sampling model much more readily than other types of items. The reason for the development of item pools to measure specific behavioral objectives is so that the tests will be constructed by randomly sampling items and to facilitate the construction of alternate forms.

The unit tests will be developed by taking a random sample of items from the LAP level item pools, and by the formulation of item pools to measure behavioral objectives formulated at the Unit level. Each Unit level test should be matched with a performance test that will be used as a criterion to validate the paper and pencil tests. The performance test should be so constructed that everyone taking it knows exactly what is required. It should also contain examples of good and bad performance so as to make it as objective as possible. It will also be scored dichotomously. The results of the paper and pencil test will be put into contingency table form with the results of the performance test. A phi coefficient will then be computed as soon as the sample size is sufficiently large to guarantee

stability. (Approximately N=25) When the correlation between the paper and pencil test and the performance test reaches .85 or above, the Unit will be considered a finished product. Until that point is reached, both the test and the learning materials will be investigated to uncover the reasons for the unsatisfactory correlation between the written test and performance.

In some cognitive areas such as health education, home management and parent involvement, it may be difficult, if not impossible to develop satisfactory performance tests with which to validate the cognitive tests. In these areas, attitude scales will be administered before and after the cognitive classroom work. The attitude change registered will then be used as the criterion for the cognitive tests. Difference in attitude can be translated into a dichotomous variable by assigning a score of 1 to all those whose attitude scores change in the desired direction, and a score of 0 to those who do not show the desired change. Due to the basic instability of attitudinal data, a correlation of .7 or above would be more than sufficient to demonstrate the validity of the cognitive test.

Finally, a comprehensive test will be developed, by sampling items from both the LAP and the Unit item pools. This test will be used as a pre-test/post-test and will be diagnostic. This test will be validated through the performance tests at the Unit level.

This program will be implemented by using in-house workshops dealing with basic measurement theory and item writing. Everyone involved with the construction of testing instruments in the cognitive area will be asked

to attend at least three training sessions. The first will deal with basic theory of test construction and basic item writing techniques. The second will be used to review and critique sample items, and the third will consist of a general review of tests that have been constructed.

The next step will consist of gathering data accumulated from the use of the instruments and reviewing both the tests and the curriculum until the desired correlations with the performance criteria are reached.

Affective Area Testing Program

The construction of attitude scales is a specialized field in and of itself, but fortunately, a sound theoretical basis for their construction has already been developed.[2] For theoretical reasons it has been found that a six point scale without a neutral point tends to be the most efficient. Also, there is no problem concerning the computation of coefficient alpha. Items can be chosen on the basis of item-total correlations until a minimum reliability of .80 is reached. After the initial item analysis, all the scales relating to a specific area can be factored to determine what specific constructs are involved. To obtain an estimate of their validity, they can be correlated with descriptions of overt behavior. The final step would be to correlate the obtained results with subjects asked to "fake-good" or "fake-bad" so as to determine how big a factor the element

---

[2]For a discussion of measurement theory with regard to the formulation of attitude scales, see Nunnally 1967, Chapters Eight and Twelve.

of social-desirability plays in determining the obtained responses.

This program will be implemented by having weekly meetings with the people involved in the development of attitude scales. These meetings will be to discuss problems in clearly defining the attitudes to be measured, and in writing items to uncover these attitudes.

When the initial items are developed they should be administered in the field to everyone applying to the program. In this manner we will have scores for both students and controls. The students should be re-tested when they leave the program, and the controls should be re-tested at approximately the same length of time. Follow-up re-testing can then be accomplished on both groups at specified intervals.

Bibliography

Nunnally, Jum C., Psychometric Theory, McGraw-Hill, New York, New York, 1967.