DOCUMENT RESUME

ED 091 400 TM 003 609

AUTHOR Bridgeman, Brent
TITLE A Duplicate Construction Experiment.
PUB DATE [74]
NOTE 4p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; Item Analysis; Norm Referenced Tests; Standard Error of Measurement; *Test Construction; Testing Problems; *Test Interpretation; *Test Reliability

ABSTRACT

This experiment was designed to assess the ability of item writers to construct truly parallel tests based on a "duplicate-construction experiment" in which Cronbach argues that if the universe description and sampling are ideally refined, the two independently constructed tests will be entirely equivalent, and that within the limits of item sampling error any person would receive the same score on both tests. Two item writing committees developed forty item driver's license examinations based solely on the material in the driver's manual. The two independently developed tests were administered to 117 high school students who took both forms three to five days apart. The two forms were not equivalent according to Cronbach's criterion. As Cronbach suggests, inspection of individual items designed to measure the same general area, but worded differently, revealed some marked differences in item difficulty. His suggestion that the standard error of measurement be estimated from split-half reliabilities seemed unwarranted. The author states that perhaps on tests of very heterogeneous content domains test-retest coefficients would be more appropriate. (RC)

A DUPLICATE CONSTRUCTION EXPERIMENT

Brent Bridgeman

University of Virginia

Many educators, realizing the problems and limitations of norm based interpretations of test scores, have become increasingly interested in assessing students with respect to absolute performance standards. But such absolute standards are rarely as absolute as they appear. While the interpretation of a student's performance level may be independent of the performance of other students, it may be very dependent on who happened to write the test questions. Unless the content universe is very precisely defined, different item writers could construct tests on which the same student would receive quite different scores, making absolute interpretations of the scores meaningless. Cronbach (1971) has suggested an experimental method for assessing the adequacy of a content universe definition which he labels a "duplicate-construction experiment." In such an experiment two completely independent groups of item writers, given the same definition of a particular domain of tasks (or universe), and the same passing standard (in terms of percent correct), write tests of a prespecified number of items. Cronbach argues that if the universe description and sampling are ideally refined, the two independently constructed tests will be entirely equivalent, and that within the limits of item sampling error any person would receive the same score on both tests. To be more precise, Cronbach suggests that the mean of the squared differences between scores of both tests should not exceed the sum of the squared standard errors of measurement of the two tests, where the standard errors could be derived from split-half analyses of the two tests. The current experiment was designed to assess the ability of item writers to construct such truly parallel tests, and to identify any practical problems in using Cronbach's model.

Method: In his hypothetical discussion Cronbach used a test based on "knowledge of the State Motor Vehicle Code". A similar task was used in the current experiment since it allows for a fairly precise universe definition (forty-two pages of the state published Driver's Manual of Virginia), and was a topic with which all the item writers, as licensed drivers, were familiar. The two item writing committees consisted of about twelve members each from an introductory graduate tests and measurements course. They were instructed to develop forty item driver's license exams based solely on material in the manual. They were asked to suppose that the state had established 75% correct answers as the minimal competency standard. The two independently developed tests were then administered to 117 driver education students from three rural high schools. Half of the students took form A first and half of them took form B first followed three to five days later by the other form.

Results and Implications: The two forms were not equivalent according to Cronbach's criterion (sum of $(X_A-X_B)^2/N=37.00$; and the sum of the squared standard errors = 16.60). The correlation between the two forms was .60, indicating some considerable changes in relative position from one form to the other. The mean score on form A was 22.4 (S.D.=4.4) while on form B it was 26.8 (S.D.=4.8). The lack of direct comparability suggests a major difficulty of making absolute interpretations of test scores; statements that a student has mastered 75% of the relevant information because he correctly answered 75% of the test items make little sense if on another test of the same information, but constructed by a different group, the student gets 60% correct.

As Cronbach suggests it should, inspection of individual items apparently designed to measure the same general area, but worded differently, revealed some marked differences in item difficulty.

For example, both test forms have questions based on a table of maximum speed limits in the Manual. On one form 74% of the students correctly identified the speed limit on interstate highways as 70 m.p.h., but on the other form only 31% recognized that the speed limit on limited access highways was also 70 m.p.h. Only a very precise domain definition would be likely to differentiate between knowledge of these two speed limits.

Cronbach's suggestion that the standard error of measurement be estimated from split-half reliabilities seems unwarranted considering his statement that "nothing in the logic of content validation requires that the universe or the test be homogeneous in content" (1971, p. 457), and his further statements that high item intercorrelations have nothing to do with content validity. Perhaps on tests of very heterogeneous content domains test-retest coefficients would be more appropriate.

References

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