✦ ✦ ✦ ✦

SOME RECENT DEVELOPMENTS IN TEST CONSTRUCTION

Elmer B. Royer, *Stillwater, Oklahoma*

For years the standard practice in the building of a battery of tests has been somewhat as follows. The test maker chose perhaps ten or twelve tests which might be sames-opposites, analogies, arithmetic problems, number series, directions, geometric figures, and so forth. Enough items were included in each test to give it the required reliability, usually .90 or .95. Combining these single tests in a mimeographed edition, he administered all of them to a group of persons on whom he obtained fairly comprehensive criterion scores. All the possible intercorrelations

of the tests and the criterion were computed and the best system of weights for combining the tests found by the least-squares method.

This method pays a premium on low intercorrelations of tests. It is axiomatic among test makers that a new test to be added to a battery must have not only a high correlation with the criterion but low correlations with the tests already in the battery. Statistically minded persons can easily see the reason for this by studying the formula for multiple correlation.

One of the recent developments in the theory of test construction is the conception of the item as a test in itself. With this conception has come a change in the statistical methods and treatment. Each item must be validated individually. Any item with low coefficient or index of validity is eliminated.

Logically the items should be combined just as tests formerly were, and possibly given differential weights. That this is not yet done by test makers is due to the extreme magnitude of the task. Even so, the fact that this method is the logical procedure raises a paradox, which may be stated: "Should the items on a test have low intercorrelations or high intercorrelatoins?" The traditional answer is, of course, that they should have high intercorrelations because high intercorrelations tend to raise the reliability of the test, particularly when that reliability is computed by the odds-evens or split-halves technic. The answer which follows logically from the concept of an item as a test in itself is that items, just as the tests of old, must have high correlations with the criterion but low correlations among themselves.

A study of the literature gives little help in resolving this paradox. Dunlap recommends the dividing of the test into four equivalent parts and making them satisfy Spearman's tetrad criterion before letting the test be called by a single name. Turney, also strongly influenced by Spearman, recommends the use of a *g* in item selection for intelligence tests, but fails to follow this argument in recommending the selection of items from the field to be measured as the sole criterion for validating achievement test items. Willoughby, in what is perhaps the most philosophical attack on the problem, concludes that a test in which the items are highly correlated is therefore highly valid in the sense that it is unitary.

It is this very error that Tryon and Lorge have been attacking, rightly pointing out that consistencies of reaction on items of an interest blank or psychoneurotic inventory do not necessarily correspond to actual traits in the individual (that is, a factor determined by the mathematical method of factor analysis cannot be *assumed* to correspond to real or psychological traits.)

My answer to this dilemma is suggested by the work of Tyler, who requires the teachers of the subject to define the objectives of their course in terms of student behavior. A very comprehensive measure of student behaviors in test situations is obtained. This measure is taken as the criterion. If later short-answer pencil and paper tests can be constructed with sufficiently high correlation with the criterion, they may be used in lieu of the criterion. Otherwise no short cut for the criterion is used.

Instead of requiring high intercorrelations among the items of the test let us require high intercorrelations among the student behaviors. If these cannot be found let us conclude that the trait we had assumed does not exist. If we do find high intercorrelations among the trait behaviors then let us conclude that here is a real trait, not established by fiat, but actually observed intercorrelations of students' behaviors in the trait

situations. Let us then secure a sufficient number of trait behaviors to have a reliable measure of the trait. Lastly, let us proceed to find shortcuts for the laborious and time-consuming task of securing comprehensive measures of the student behaviors in the trait situations. If we can find or make a pencil-and-paper test that will correlate highly enough with the *g* factor in the observed behaviors let us use it instead of the more laborious method of measuring the behaviors directly, and in selecting the items in the pencil-and-paper test let us use the regular statistical devices which will select items with low intercorrelations, even though the pencil-and-paper test so constructed will not have high reliabilities as measured by the split-halves or odds-evens correlation.

✦ ✦ ✦ ✦