THE CONCEPT OF TEST AND ITEM RELIABILITY
IN RELATION TO FACTOR PATTERN

ROBERT J. WHERRY AND RICHARD H. GAYLORD
UNIVERSITY OF NORTH CAROLINA

It is shown that approaches other than the internal consistency method of estimating test reliability are either less satisfactory or lead to the same general results. The commonly attendant assumption of a single factor throughout the test items is challenged, however. The consideration of a test made up of \( K \) sub-tests each composed of a different orthogonal factor disclosed that the assumption of a single factor produced an erroneous estimate of reliability with a ratio of \( \frac{n-K}{n-1} \) to the correct estimate. Special difficulties arising from this error in application of current techniques to short tests or to test batteries are discussed. Application of this same multi-factor concept to item-analysis discloses similar difficulties in that field. The item-test coefficient approaches \( \sqrt{1/K} \) as an upper limit rather than 1.00 and approaches \( \sqrt{1/n} \) as a lower limit rather than .00. This latter finding accounts for an over-estimation error in the Kuder-Richardson formula (8). A new method of isolating sub-tests based upon the item-test coefficient is proposed and tentatively outlined. Either this new method or a complete factor analysis is regarded as the only proper approach to the problem of test reliability, and the item-sub-test coefficient is similarly recommended as the proper approach for item analysis.

The term reliability has been used loosely to apply to the resultant of the application of many different statistical operations. As a general rule these operations attempt to determine the verifiability of the original data and thus to establish the degree to which non-chance factors entered into the original measurements.

One group of techniques approaches the problem through the direct comparison of the observed distribution of measurements with that which would have arisen by chance in accordance with certain theories of probability. These techniques parallel the use of the critical ratio in establishing the reliability of means, sigmas, differences, and other statistical parameters by disproving the null hypothesis. One such approach is that of Jackson (7), who measures the sensitivity of the test, gamma (\( \gamma \)), by taking the ratio of the standard deviation of the capacity to the standard deviation of chance. Another writer, Hoyt (6), using the analysis of variance approach, suggests using the ratio of the true variance of the student responses to
the obtained variance among students. Edgerton and K. F. Thomson (3) suggest using the Lexis ratio to show that differences among students are greater than those among items. Hoyt shows that his results are comparable to those of Jackson, and both Hoyt and Edgerton and Thomson show their results to be comparable to those achieved by the use of certain of the Kuder-Richardson series [actually to formulas (14) and (20), which involve only the sigmas of the items and the sigma of the total test] which we will discuss later at some length. Any restrictions demonstrated to hold for these particular portions of the Kuder-Richardson series will thus apply to these probability methods as well. Allied techniques are (1) the Horst (4) maximized criterion which holds that the most reliable weighted composite is that with the largest relative variance, and (2) the item selection techniques based on the variance (difficulty) of the item. The remaining techniques attack the measurement of reliability through the verifiability of the original ranks of the members of the population. These techniques form two natural groups as the emphasis on stability is (1) regardless of time, or (2) regardless of the particular test or items used.

The first of these viewpoints—verifiability regardless of time—is exemplified by the test-retest method of measuring reliability. While appearing best to meet the operations indicated in the usual definition of reliability, this method has been widely criticised from many different viewpoints. Typical criticisms refer to the effects of differential practice, memory, inability to duplicate testing conditions, inability to sustain motivation, etc. Perhaps even more serious is the objection raised that this coefficient is affected not only by the unreliability of the test but also by the unreliability (lack of stability) of the function being tested. Paulsen (12) suggested measuring this trait fluctuation by correcting the test-retest coefficient for attenuation, using the split-half method for obtaining the reliabilities of the initial and final testings. Woodrow (18) suggested measuring this "quotidian variability" by the ratio of the actual sigma of the means of various samples to the sigma of the means as predicted from the average standard deviation of the multiple samples. Thouless (17) proposed the measurement of this "functional fluctuation" by what he called the double test-retest index, where he substituted alternate forms for the split-half approach of Paulsen. Thus in application we see that the "regardless of time criterion" in addition to many other ills requires an appeal to the preceding view—as in Woodrow—or to the remaining criterion of stability "regardless of test or items used" criterion—as in the split-half or comparable form methods of Paulsen or Thouless. Indirectly associated techniques are: (1) the original