TEST RELIABILITY AS A FUNCTION OF SUBJECT ATTITUDE TOWARD TEST TAKING

by

GERALD M. EADS II

B.A., Western Washington State College, 1967

A MASTER'S THESIS

submitted in partial fulfullment of the

requirements for the degree

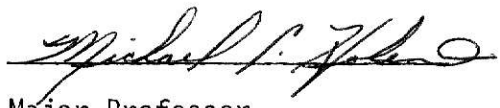
MASTER OF SCIENCE

College of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1976

Approved by:



Major Professor

ACKNOWLEDGEMENTS

There are literally hundreds of individuals who have significantly touched my life in the nine years culminating in the completion of this exercise. The vast majority of them deserve thanks for their impact upon me both positive and negative.

In a more immediate sense, deserving of appreciation are those who provided assistance and support in the accomplishment of the task itself. I wish to thank Drs. Theresa Chang, Bob Newhouse, Leon Rappoport and Carole Urbansok for allowing access to their classes, and Major Ted Slifer for providing access to the correctional program trainees. A special note of thanks must be given to Lieutenant Carl Steiner, without whose assistance there would not have been a military sample.

To a fellow named Tom Gooch I owe the debt of my understanding of the Black Box that digested and regurgitated the numbers resulting in the following pages. I am honored to have known and learned from him.

Significant growth forces have also been bestowed by such good people as Drs. E. Robert Sinnet, Mike Rohrbaugh, Evelyn Gauthier, Dick Wampler, Steve Handel, Steve Carmean, B. L. Kintz, Louis Lippman and Kit and Saundra Taylor. Without their influence and support I surely would not have arrived here through my incredibly circuitous journey.

My committee, of course, has earned my gratitude for their support and assistance -- Drs. Fred Bradley, Mike Holen and Leon Rappoport. Mike, my primary mentor at this point in time, has managed the provision of a compendious assemblage of most useful knowledge and skill through his direction of this task. To Leon must go my appreciation and indebtedness beyond the scope of language for his unfailing support, direction, assistance, nurturance, and when appropriate, admonition, during the last six years of my frustrating search for purpose.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION AND REVIEW OF THE LITERATURE

Self-report methods of collecting information about individual behavior are common alternatives to laboratory measurement in psychology and education. Many instruments have been developed, refined, and utilized for the assessment of constructs ranging from ability and aptitude to personality, attitudes and values. For most of these instruments, both published and those still considered 'research' instruments, the adequacy of the reliability information available has been questioned too infrequently.

Nunnally (1967) provided the following definition of the reliability coefficient: "By convention, the average correlation of one test, or one item, with all tests or items in the domain is called the reliability coefficient." "The square root of the average correlation is equal to the correlation of the first item or first test with true scores in the domain (p. 179)." The domain is considered to be the infinite population of test items or tests from which the item or test is drawn. There are two major models discussed in Nunnally: (1) the 'domain sampling' model, which assumes that any item or test is a random sample of the items or tests from a hypothetical domain of infinite size, and (2) the 'parallel test' model, which assumes that any test is parallel to other tests measuring the same attribute. More precisely, "two tests are parallel if (1) they have the same standard deviation, (2) they correlate the same with a set of

true scores and (3) the variance in each test which is not explainable by true scores is because of purely random error" (Nunnally, op.cit., p.181).

These three assumptions made from the latter model force the fourth assumption that the correlation between any two tests in a domain is a precise determination of the reliability coefficient, rather than an estimate. Furthermore, only the first assumption can be empirically determined, and no way is available to determine the correctness of the latter assumption.

Precision of estimation is accepted as an issue by the domain sampling model; a reliability coefficient is only an estimate of an average correlation with the domain. The parallel test model is much more restrictive than the domain sampling model. Nunnally continues to argue in favor of the domain sampling model, demonstrating that the parallel test model is only a special case of the more general domain sampling model.

There are several methods commonly utilized to estimate the reliability of tests. Ebel (1965) lists the test-retest, equivalent forms and split-half methods as the easiest and most common approaches to reliability estimation. All of these methods require only the application of a parametric correlation formula to two sets of scores on the same subjects, with a correction for length in the case of the split-half estimate..

The test-retest method requires the readministration of an instrument after a suitable period of time. The correlation between the set of scores obtained on the first administration

and that obtained on the second administration provides the test-retest reliability coefficient. Several problems are inherent in the method. The correlation is between the administrations of the same set of items, and hence does not provide any indication of the relationship between the test items and other items purported to be from the same item domain. In addition, subjects' answers to the second administration are not independent of the first: memory and intercommunication between subjects can influence the data from the second administration.

The parallel forms method of reliability estimation is not subject to the above concerns. The method demands, however, the construction of two tests measuring the same construct and trait continuum which are then correlated subsequent to the administration of both forms. Parallel forms, though, are ordinarily not available to the investigator.

The split-half method is considered by Ebel (op.cit.) to be a practical alternative to the above routines and their consequent problems. To obtain the reliability estimate, the items from a single test are divided into two reasonably equivalent halves (usually by scoring the odd-numbered and even-numbered items separately). The scores from these independent subtests are then correlated in the same manner as the methods described above. Reliability coefficients are affected by test length; the estimates derived by this method are corrected by the Spearman-Brown formula which predicts the coefficient that would be obtained from the correlation of tests (in this case) twice as long as the split-halves.

There are numerous ways to split the items for the above approach to reliability estimation, of course, and the obtained coefficients will vary somewhat depending on the splitting method. This raises some question as to what the reliability is. Nunnally (op.cit.) proposes that the corrected (for length) correlation between any two halves of a test can be considered an estimate of the coefficient alpha, a reliability estimation procedure which is based on the average correlation among test items (usually referred to as 'internal consistency') and the number of items in the test.

Coefficient alpha and a special version applicable to dichotomous items (Kuder-Richardson Formula 20 (KR-20)) are the formulas used to determine reliability based on internal consistency. These formulas set an upper limit to the reliability obtainable from an instrument (Nunnally, op.cit.). A low alpha or KR-20 coefficient is indicative that either the test is too short or that the items have very little in common. Although there are several sources of measurement error not considered by coefficient alpha and KR-20 (e.g., mood changes or actual changes on the trait continuum over time or content differences between alternate forms), these estimates are usually very close to other kinds of estimates in most situations, since "the major source of measurement error is because of the sampling of content" (Nunnally, op.cit., p. 211).

Nunnally also demonstrated that internal consistency reliability estimates consider not only sampling of item content, per se, but

also consider "sources of measurement error that are present within the testing session" (op.cit., p. 224). Neither in the literature nor in the measurement texts, however, was the issue of subject (S) variability, as it affects reliability, approached in any depth. Reliability theory limits the investigator to the consideration of certain groups as defined by such determinants as demographics and personality and performance variables. Other uncontrolled variables may produce results at variance with what otherwise might occur; such variables are usually not directly addressed in the literature. How Ss respond to an instrument may be important; the 'set' or 'approach' that Ss take to completing an instrument may lead to substantial differences in the 'reliability' with which they respond to the items. Nunnally discussed the need to control the error caused by S variation in test performance. In order to minimize the impact of S error sources, Nunnally suggested that a minimum of 300 should be used in reliability research. This policy ought, he suggested, allow the random assignment of errors such as illness, misunderstanding of instructions and inadvertent clerical errors to all items within a test. In discussing variation between tests (as with test-retest reliability) Nunnally discussed change on the attribute and 'change in health' as possible sources of measurement error.

In no case was the possibility of S motivation to take the test addressed. Most discussion of testing has concerned the assessment of achievement measures where it has been assumed that

the S has some investment in test outcome. In that this assumption may be questionable (e.g., Bauer, 1973), it may even be less reasonable to assume that Ss will be equally invested in being accurately measured by attitude or personality instruments: there frequently are no strong reinforcers for responding accurately to items in such scales.

There are several possible effects of Ss' 'not caring' to accurately respond to an instrument. Such an attribute as 'not caring' -- which may or may not be consistent or chronic -- is perhaps systematic rather than random, in that the attribute could be viewed as a continuous trait as is, for example, 'social desirability'. In reference again to Nunnally's (op.cit.) discussion of reliability theory models and internal consistency: "All errors that occur within a test can be easily encompassed by the domain sampling model. The assumptions of the model can be extended to the case where situational influences are randomly 'assigned' to the items. Thus not only would each person be administered a random sample of items from the domain, but also each item would be accompanied by a random set of situational factors. Then whether or not a person passes any item drawn at random from the domain is a function partly of the happenstance of which item is selected and partly of the happenstance of the situational factors that accompany the item. All such sources of error will tend to lower the average correlation among items within the test, but the average correlation is all that is needed to estimate the reliability" (p. 208). If a S 'does not

care' about responding accurately, however, it seems that a substantial contribution may be made to the error variance by the attribute rather than the item sampling, in which case the reliability estimate will underestimate the 'true' reliability. Furthermore, if, for example, 15% of the sample population tends to 'not care' then it matters little whether the sample is 30, 300 or 3000; that portion of the sample will not be measured accurately on the attribute in question.

If the S 'does not care' -- is not invested in being measured accurately -- he can be presumed not to consistently answer items in accordance with his true position on whatever dimension (continuum) is in question. The result may be a response pattern which may range from early 'fatigue' and low response consistency late in the instrument administration, to completely random responding. Such error is surely not 'random' in the sense the domain sampling model assumes.

This inconsistency in response pattern leads to a concern for what is being measured; it is no longer a function of the appropriateness of the item sample, which Nunnally contends is the major source of internal consistency reliability error. Even should the items be a highly accurate representation of the domain they cannot be utilized as an accurate measure unless the S chooses to respond to that measure accurately.

The problem is similar to that of measuring the level of learning of a rat in the classic and time-honored 'T-maze'. The Experimenter (E) normally starves the rat for a predetermined

number of hours in order to provide the appropriate level of motivation for the rat to respond accurately to the measure of learning. Should the rat not be hungry for some reason (for example, the laboratory technician's mistaken feeding of the animal immediately prior to the experimental seesion), it will simply 'not care' to rush to the appropriate goal box for food, and hence E cannot possibly accurately measure the rat's performance in reference to its level of learning.

This lack of investment leads to several diversions from Nunnally's assumptions concerning reliability. He has predicted that the obtained variance of a measure will be restricted as a result of low reliability. The obtained variance is a combination of both true and error variance. When reliability is low, the true variance will be reduced and more of the obtained variance will be due to error. In that error variance is much more restricted in range than true variance, the obtained variance (that observed) will be reduced. This ought to happen in the case of low S motivation to be accurately measured. High and low true score Ss will tend to produce scores towards the scale midpoint rather than at or near their 'true' scores. The obtained variance ought to reflect this tendency.

Other problems do not meet Nunnally's predictions. The distribution of the obtained scores should be skewed away from the mean and the scale midpoint, such that they overestimate the true scores. If a S disallows accurate measurement, however, it should be expected that the obtained scores will underestimate the true

scores of higher and lower scoring Ss who have no motivation to be accurately measured. "Coefficient alpha" (and hence KR-20) "is sensitive not only to the sampling of items but also to sources of measurement error that are present within the testing session" (Nunnally, op.cit., p. 224).

The other major factor in test reliability, of course, is test length. "The primary way to make tests more reliable is to make them longer" (Nunnally, *op.cit.*, p. 223). An increase in test length should not alleviate the problem of motivation, however; if a S 'does not care' to fill out a short test, he almost certainly will not become escited at the prospect of filling out a longer one. If anything, a longer test will aggravate the problem.

In order to explore the impact of possible differential subgroup reliability on personality and attitude measures, the Rotter Internal-External Locus of Control (I-E) scale was chosen as representative of the many scales available in this area.

Review of the Literature

Rotter's Social Learning Theory (1954) suggested that individuals develop a generalized expectancy concerning their ability to control events. Those persons who believe that their actions can affect the course of their lives are said to have an expectancy of internal control. Those who believe that chance, fate or luck determines the outcome of events are identified as having an expectancy of external control.

The amount of attention that behavioral scientists have recently given to the I-E construct is substantial. There are now in excess of twelve scales purporting to measure locus of control. Throop & McDonald (1971) published a bibliography containing 339 references and later work indicates a continuing increase in published interest in the construct (Robinson & Shaver, 1973). Within the literature, however, little effort is evident concerning issues of reliability of the instrument.

Most of the I-E research has relied heavily on the reliability data reported in Rotter's original monograph (1966). Test-retest reliability estimates encompassing one- and two-month intervals ranged from .65 to .79. KR-20 reliability estimates calculated from data collected from high school and college students were from .69 to .76. Rotter defended these levels suggesting that KR-20 results were somewhat limited by the forced-choice format in which the attempt was made to "balance alternatives so that probabilities of endorsement of either alternative do not include the more extreme splits" (p. 10).

Hersche & Scheibe (1967) reported further I-E scale test-retest reliability data using student volunteer mental health workers. Reliability coefficients ranged from .43 to .84 over two month intervals. These estimates were based on data collected from the volunteers while they were serving as helpers in four state-operated mental institutions. Data were collected before and after an eight-week training program. Matched control-group reliabilities did not differ significantly from those of the

experimental groups. Data collected from eighteen students who participated in the same program for two consecutive years revealed a one-year test-retest reliability of .72, based on the correlation of the sum of their 1964 pre- and post-scores with the sum of their comparable 1965 scores. Estimates of the reliability of the summed pre- and post-scores for the various groups ranged from .60 to .91.

Harrow & Ferrante (1969) reported test-retest reliabilities for the Rotter I-E scale administered to several groups of acute psychiatric patients. Reliability for the total sample was .75, comparing favorably with Rotter's (1966) student samples. Test-retest reliabilities computed on subgroups of the psychiatric sample showed similar results.

Numerous other studies approach issues of validity of the various I-E scales, and in some instances make the inclusion of reliability data. Some concern is expressed in the literature as to the applicability of the original I-E scale to other populations; this argument has served to justify the development of several subsequent scales. Valecha & Ostrum (1974) reported an abbreviated scale deemed useful for certain testing conditions and included data on several psychometric properties of the new scale including internal consistency reliability as measured by coefficient alpha. Reliabilities on data from their eleven item scale for the Caucasian, Negro and total samples were .66, .49 and .62, respectively. Nowicki & Strickland (1972) reported the development of an I-E scale for children, using the justification

that the original Rotter scale was too difficult for grade levels below high school. Split-half reliabilities were reported for different grade level groups and ranged from .63 for the grade 3-5 group to .81 for the grade 12 group. Gorsuch, Henighan & Barnard (1972) addressed strong concern over the impact of certain individual differences unrelated to the construct of I-E in so far as they would impact research results. Specifically, these authors demonstrated a relationsnip between verbal ability and I-E scores obtained with Bialer's (1961) I-E scale for children. Estimates of scale reliability were generally nonexistent for low-verbal-ability children, while estimates of the reliability of the scale used on high-verbal-ability children (4th and 5th graders) reached .60.

In addition to the recent concern relative to the impact of individual ability on research results, a substantial amount of emphasis has been directed at the influence of other 'trait continuums' on I-E research results. One of the simpler approaches to this problem has been to correlate the scores of a measure of one construct with those of a measure of internal-external control. Hjelle (1971), for example, addressed the influence of social desirability as measured by the Marlowe-Crowne Social Desirability Scale on the scores of the Rotter I-E scale. Reliability statistics utilizing social desirability as an independent variable, however, were not included.

Other work has addressed the possible influence of other ability and personality dimensions but has not always included

supportive reliability data. Adler (1973) and Gay & Abrams (1973) published essays concerning the impact of intra-cultural differences on data gathering and testing procedures in general. Both articles contended that different cultural upbringing can cause misleading results due to variable ability, motivation and reaction to testing procedures in different ethnic groups.

Lefcourt & Ladwig (1966), in a study of reformatory inmates and the construct of alienation, and Lamont (1972), in a brief report of mood-level and impact on I-E, addressed their data in terms of constructs and trait continuums, but their data begged the question of less stable individual differences effects on scale scores.

In a study reporting both factor analysis and reliability data on the original I-E scale, Cherlin & Bourque (1974) attacked the viability of Rotter's (1966) supposition that the original I-E scale is unidimensional, and performed both factor analyses and reliability estimations on the data collected from college and non-college samples. Both approaches demonstrated that the Rotter scale measures different aspects of the construct with non-college populations and should be used with caution on samples from other than the college population on which it was developed. The non-college population sampled in the study were California residents who had recently experienced a severe earthquake, but the authors mentioned this variable almost in passing and did not address the possible acute impact such an experience of crisis proportions might have on a population and the data gathered from them.

The Rotter I-E scale is appropriate for use in the present study in light of its adequate but not high reliability and the broad range of obtained reliabilities reported in the literature. This range (approximately .40 to .90) allows the possibility of very low S related reliability. A highly reliable test might not produce widely divergent differential subgroup reliabilities. In addition, the literature suggests that there is a substantial range of population variability in response to the test.

Very little work has been published in the area of S differential reliability although some investigators have discussed changes in mood and S motivation with achievement and projective tests. Ray (1974), in addressing the development of reliability in projective tests, spoke of the problem of 'mood swings' and the fluctuations in Rorschach test scores, but avoided confrontation with the problem by apparently hoping to transcend it. "... Rorschach scores do fluctuate in cycles as would be expected if mood swings were involved. This line of argument is, however, a potentially limiting one: it is very often desirable to use projective test scores as indices of traits, i.e., as indices of chronic rather than of momentary dispositions. If a projective test is to be used to measure consistent traits in people, it does seem an indispensable requirement that the measurements it provides should also be consistent" (p. 303). He went on to talk primarily of test length as the solution to low projective test-retest reliability.

Bauer (1973) discussed possible sources of error in aptitude

and achievement test scores; in confronting researchers in their avoidance of measuring motivation, he stated that "although a number of investigators have noted that differences among individuals in the acceptance of societal criteria for success and failure influence behavior in testing situations, few studies have included this variable in their design" (p. 32). His contention was that there is great variability in individuals' willingness to do well in testing situations.

Adler (1973) acknowledged the issue of individual differences in motivation and the impact on test reliability and validity, but addressed only the impact on individual administration situations. His discussion was limited to the concern of improving test-taking attitude on an individual basis prior to testing. No mention was given to the impact of hostility or the lack of motivation on the reliability of instruments in group testing situations.

These authors have either directly or indirectly acknowledged the existence of and problems with attitude and emotional variables in the collection of test data. No researchers, however, have undertaken the task of assessing the impact of S attitude toward test taking on test reliability.

Intent of the Study

If a measurable attribute exists which reflects the degree of motivation an individual possesses for responding to a scale accurately, then identifiable subgroups should exist that will demonstrate differential reliabilities as a function of their position on such a continuum. It was determined that for

exploratory purposes several brief questions could serve to identify groups of Ss demonstrating differential levels of investment in responding to a scale.

It was assumed that Ss who like filling out a scale, or those who find a particular scale interesting, or even those who for whatever reason feel that a particular instrument is important ought to fill out that instrument with a greater degree of care than those Ss who have a dislike for filling out instruments, are not interested by such activity, or who see no importance to the scale. This difference should be reflected by lower reliabilities obtained from subgroups expressing negative feelings on these dimensions.

It was furthermore predicted that, as postulated by Nunnally (op.cit.), those groups expressing such negative feelings would also produce restricted score variances due to the restricted nature of the dominant error variances of the obtained scores.

If a tendency also exists such that the scores obtained from the groups reflecting negative feelings underestimate high scores and overestimate low scores, mean scores from different population samples that are normally significantly different should fail to reach significance for negative attitude subgroups.

Because of the exploratory nature of the study and the need to collect data from several populations including some other than the traditional college pool, sample sizes were smaller than many reliability theorists would propose. In that the study

is exploratory it was further decided to differentiate between small subgroups within the samples in order to at least note indications of the effect of differential subgroup motivation on reliability.