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

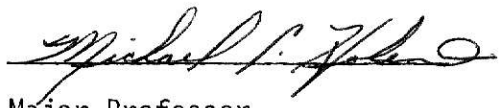
MASTER OF SCIENCE

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.

CHAPTER II

PROCEDURES AND PRESENTATION OF DATA

Subjects

Ss were obtained from Kansas State University undergraduate education and psychology classes, and from a major midwestern military installation. Completed scales were collected from 102 male and 181 female college students. Trainees from a military criminal rehabilitation facility provided 107 completed scales, and 106 useable scales were collected from non-offender active duty soldiers. All military personnel were below the enlisted grade of E-5 in order to provide age comparability with the college samples. No females were present in the military samples.

Procedure

All Ss were administered identical I-E scales and were presented with and read identical instructions. A reproduction of the instructions and instrument appear in Appendix A.

Printed at the end of the scale were three questions used to determine the liking, interest, and importance subgroups. These also appear in Appendix A.

Instruments were administered to the college Ss during the last twenty minutes of normal class periods through prior arrangement with instructors. Scales were adminstered to the trainees at the military corrections institution during the last twenty minutes of normal data collection periods scheduled by the organization's evaluation section at the completion of the trainees' rehabilitation program immediately prior to their return

to active duty. Active duty military personnel were obtained from two sources: Thirty-six young enlisted military police responded to the scale immediately following completion of their duty shift. The remaining Ss, obtained from an Engineering batallion, completed the scale at the end of scheduled Race Relations classes held by their organizations.

Instrument

The instrument was the standard Rotter 29 item (six filler item) scale with instructions taken from Wilson (1972). Three items were added at the end of the scale. The entire instrument with instructions appears in Appendix A.

Presentation of Data

KR-20 reliability coefficients were calculated using the Roscoe (1973) FUNSTAT 'item analysis for objective tests' G1 program. Data organization and ANOVA functions were performed using the Data-Text Social Data Analysis language (Armor & Couch, 1972).

Data for KR-20 analyses were organized by two procedures. Tables 1 through 7 show analysis results, means and standard deviations for subgroups as defined by their responses to the three 'Interest in Questionnaire' items as follows:

The "All 'Yes'" groups were defined by Ss' responding positively (first alternative) to all three questions.

The "Not All 'Yes', Zero 'No'" groups were defined by Ss' responding to less than all three first alternatives to the items, but not responding to any of the last (negative) alternatives

to the three items.

The "At Least One 'No'" groups were defined by Ss' responding to at least one negative alternative on the three items.

Tables begin by showing the results for these differential subgroups with all samples combined, and successively differentiating these samples, first into 'military' and 'civilian' groupings (Tables 2 and 3) and finally into the individual independent samples (Tables 4-7).

Tables 8 through 14 show results in identical format to Tables 1 through 7, but list subgroups as defined by responses to each of the three 'Interest in Questionnaire' items, with groups distinguished by positive, neutral and negative responses to each item, listed sequentially by items 1, 2 and 3.

Table 15 shows three separate simple ANOVA analyses on individual samples with analyses performed on the three subgroups as differentiated by the "All 'Yes'", "Not All 'Yes', Zero 'No'", and "At Least One 'No'" definitions listed above.

TABLE 1

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee, Normal Duty Military, College Male and College Female Samples Combined. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	118	0.7498	9.949	4.234
Not All "Yes", Zero "No"	270	0.7034	11.000	3.957
At Least One "No"	108	0.5809	11.306	3.401
Total Sample	496	0.6996	10.817	3.994

TABLE 2

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee and Normal Duty Military Samples Combined. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	55	0.7614	8.673	4.174
Not All "Yes", Zero "No"	96	0.6339	9.844	3.513
At Least One "No"	62	0.6024	11.226	3.466
Total Sample	213	0.6847	9.944	3.802

TABLE 3

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Male and College Female Samples Combined. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	63	0.7088	11.063	3.964
Not All "Yes", Zero "No"	174	0.7190	11.638	4.043
At Least One "No"	46	0.5654	11.413	3.307
Total Sample	283	0.6981	11.473	3.922

TABLE 4

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee Sample. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	38	0.7410	8.079	3.950
Not All "Yes", Zero "No"	42	0.6947	9.714	3.807
At Least One "No"	27	0.5650	10.963	3.328
Total Sample	107	0.7063	9.449	3.911

TABLE 5

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Normal Duty Military Sample. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	17	0.7834	10.000	4.352
Not All "Yes", Zero "No"	54	0.5837	9.944	3.263
At Least One "No"	35	0.6427	11.429	3.556
Total Sample	106	0.6542	10.443	3.621

TABLE 6

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Male Sample. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	25	0.6954	10.880	3.871
Not All "Yes", Zero "No"	53	0.7793	11.170	4.526
At Least One "No"	24	0.6124	11.667	3.484
Total Sample	102	0.7283	11.216	4.153

TABLE 7

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Female Sample. Three Item "Interest in Questionnaire" Response Groups and Total Sample.

Response Group	N	KR-20	Mean	SD
All "Yes"	38	0.7341	11.184	4.019
Not All "Yes", Zero "No"	121	0.6825	11.843	3.794
At Least One "No"	22	0.5185	11.136	3.079
Total Sample	181	0.6801	11.619	3.777

TABLE 8

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee, Normal Duty Military, College Male and College Female Samples Combined. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like" and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	247	0.7218	10.466	4.055
Neither Interesting Nor Boring	198	0.6951	11.212	3.923
Boring	51	0.6407	10.765	3.606
Item 2:				
Like	176	0.7267	10.102	4.094
Neither Like Nor Dislike	261	0.6928	11.192	3.882
Dislike	59	0.5803	11.228	3.430
Item 3:				
Important	211	0.7190	10.237	4.038
Neither Important Nor Unimportant	215	0.7091	10.912	3.996
Unimportant	72	0.6261	11.931	3.564

TABLE 9

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee and Normal Duty Military Samples Combined. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like" and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	96	0.7187	9.010	3.933
Neither Interesting Nor Boring	81	0.6451	10.568	3.610
Boring	36	0.6763	10.722	3.724
Item 2:				
Like	78	0.7410	8.808	4.057
Neither Like Nor Dislike	100	0.5773	10.330	3.305
Dislike	35	0.6808	11.371	3.840
Item 3:				
Important	85	0.7122	9.153	3.891
Neither Important Nor Unimportant	86	0.6649	9.663	3.649
Unimportant	43	0.6169	11.837	3.517

TABLE 10

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Male and College Female Samples Combined. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like" and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	151	0.6936	11.391	3.855
Neither Interesting Nor Boring	117	0.7191	11.658	4.066
Boring	15	0.5610	10.867	3.304
Item 2:				
Like	98	0.6812	11.133	3.822
Neither Like Nor Dislike	161	0.7320	11.727	4.110
Dislike	24	0.3375	11.167	2.718
Item 3:				
Important	126	0.7084	10.968	3.970
Neither Important Nor Unimportant	129	0.7106	11.744	4.001
Unimportant	29	0.6565	12.069	3.629

TABLE 11

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Military Corrections Trainee Sample. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like" and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	55	0.7156	8.673	3.918
Neither Interesting Nor Boring	37	0.7039	9.973	3.838
Boring	15	0.7785	10.267	4.328
Item 2:				
Like	51	0.7217	8.157	3.857
Neither Like Nor Dislike	45	0.6094	10.511	3.449
Dislike	11	0.7091	11.091	4.010
Item 3:				
Important	55	0.7199	8.782	3.911
Neither Important Nor Unimportant	37	0.7269	9.189	3.979
Unimportant	16	0.6066	11.750	3.419

TABLE 12

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: Normal Duty Military Sample. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like" and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	41	0.7291	9.463	3.908
Neither Interesting Nor Boring	44	0.5702	11.068	3.326
Boring	21	0.5671	11.048	3.184
Item 2:				
Like	27	0.7484	10.037	4.141
Neither Like Nor Dislike	55	0.5609	10.182	3.174
Dislike	24	0.6763	11.500	3.753
Item 3:				
Important	30	0.6951	9.833	3.760
Neither Important Nor Unimportant	49	0.6013	10.020	3.335
Unimportant	27	0.6478	11.889	3.573

TABLE 13

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Male Sample. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like", and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	51	0.7490	10.843	4.207
Neither Interesting Nor Boring	41	0.7341	11.707	4.233
Boring	10	0.5521	11.100	3.239
Item 2:				
Like	41	0.6718	11.073	3.815
Neither Like Nor Dislike	48	0.8069	11.313	4.744
Dislike	13	0.2051	11.308	2.493
Item 3:				
Important	39	0.7461	10.051	4.224
Neither Important Nor Unimportant	46	0.7227	11.717	4.084
Unimportant	17	0.6124	12.529	3.449

TABLE 14

KR-20 Reliability Coefficients, Means and Standard Deviations for Rotter I-E Scale: College Female Sample. Individual "Interest in Questionnaire" Item Response Groups "Interesting", "Like", and "Important".

Response Group	N	KR-20	Mean	SD
Item 1:				
Interesting	100	0.6571	11.670	3.631
Neither Interesting Nor Boring	76	0.7145	11.632	3.973
Boring	5	0.6141	10.400	3.382
Item 2:				
Like	57	0.6989	11.175	3.826
Neither Like Nor Dislike	113	0.6831	11.903	3.796
Dislike	11	0.5089	11.000	2.954
Item 3:				
Important	87	0.6822	11.379	3.779
Neither Important Nor Unimportant	83	0.7110	11.759	3.953
Unimportant	12	0.7116	11.417	3.774

TABLE 15

Comparison of Samples on Rotter I-E Scores. Three Item "Interest in Questionnaire" Response Groups Analyzed Separately.

	Military Trainee	Duty Military	College Male	College Female	F
			All "Yes"		
Mean	8.079	10.000	10.880	11.184	4.245 ($p < .001$)
SD(N-1)	4.00	4.49	3.95	4.07	
N	38	17	25	38	
			Not All "Yes", Zero "No"		
Mean	9.714	9.944	11.170	11.843	4.799 ($p < .01$)
SD(N-1)	3.85	3.29	4.57	3.81	
N	42	54	53	121	
			At Least One "No"		
Mean	10.963	11.429	11.667	11.136	0.208 (NS)
SD(N-1)	3.39	3.61	3.56	3.15	
N	27	35	24	22	

CHAPTER III

RESULTS AND DISCUSSION

With little exception, for the combined sample analyses that utilized adequate sample sizes, KR-20 reliability coefficients (KR-20s) and standard deviations (SDs) of the sample scores closely followed the predicted directions of decreasing magnitude as a function of decreasing test taking motivation. KR-20s for the combined samples, as shown in Table 1, were 0.7498, 0.7034 and 0.5809 for the positive, neutral and negative subgroups, respectively. SDs for those respective groups, also shown in Table 1, were 4.234, 3.957 and 3.401. Score means for the respective subgroups also began to approach the scale midpoint (11.5) as reliability decreased, as predicted by the hypothesis concerning random responding. Score means for the positive, neutral and negative subgroups were 9.949, 11.000 and 11.306, respectively.

The combined military samples, as shown in Table 2, followed this pattern identically. The combined college samples, shown in Table 3, revealed a slight inversion in the pattern in the neither subgroup. This may have been a function of different response patterns and attitudes being reflected by the college samples as they related to the interest items. The negative subgroup still demonstrated a substantial reduction in the magnitude of the KR-20 and SD relative to the other subgroups. KR-20s were 0.7088, 0.7190 and 0.5654, and SDs were 3.964, 4.043 and 3.307 for the college positive, neutral and negative subgroups, respectively.

College sample mean scores for all subgroups were very close to the scale midpoint, and little score change was noticeable.

Large inversions from the predicted directions of magnitude change only occurred when sample sizes were relatively small. An example can be seen in Table 5, which shows the higher KR-20 of 0.6427 for the negative subgroup, relative to the neither subgroup KR-20 of 0.5837. The positive subgroup KR-20 was 0.7834. College males (Table 6) showed an inversion between the positive subgroup KR-20 of 0.6954 and the neutral subgroup KR-20 of 0.7793. SDs reflected these inversions in all cases.

Analyses using large sample sizes, however, showed consistent support for all hypotheses. Inversions were inconsistent and did not strongly counter the hypotheses even though they did not lend support. The negative subgroup KR-20s were in only two cases larger than the positive subgroup KR-20s. Both of these occurred in individual item subgroups. As shown in Table 11, the Item 1 positive subgroup revealed a KR-20 of 0.7156 while the Item 1 negative subgroup showed a KR-20 of 0.7785. Table 14 shows the Item 3 positive subgroup KR-20 of 0.6822 while the Item 3 negative subgroup KR-20 was 0.7116.

ANOVA comparisons between the four samples, shown in Table 15, as separated into interest subgroups, provided support for the hypothesis that differences between various population samples would not be found significant with the negative interest subgroups. While it is possible that these results were a product of a real (true score) decrease in I-E differences

between the samples, as opposed to being the result of the predicted tendency for obtained scores to underestimate high scores and overestimate low scores for low motivation subgroups, the decrease in size of the SDs, in accordance with Nunnally's (op.cit.) contention concerning error variance, is supportive of the latter interpretation.

These results have generally supported the contention that test taking motivation plays an important role in determining the reliability of tests in a broad range of situations and with different populations. It is likely that reported test reliabilities in many cases have been diminished by differential test taking motivation of Ss, and that this contaminant has affected the interpretation of reliability error variance, heretofore considered primarily a function of item sampling.

These demonstrated differential subgroup reliabilities, traditionally considered to be subsumed under the heading of random error, should be considered as systematic and controllable. Reliability research may now be at the stage that it must attend to the control of this variability and begin to account for the motivation of Ss in the measurement of constructs.

BIBLIOGRAPHY

Adler, S. Data Gather : The Reliability and Validity of Test Data from Cultura y Different Children. Journal of Learning Disabilities, 1973, 6(7), 429-434.

Armor, D. J., and Couch, A. S. Data-Text Primer: An Introduction to Computerized Social Data Analysis. New York: The Free Press, 1972.

Bauer, D. H. Error Sources in Aptitude and Achievement Test Scores: A Review and Recommendation. Measurement and Evaluation in Guidance, 1973, 6(1), 28-34.

Bialer, I. Conceptualization of Success and Failure in Mentally Retarded and Normal Children. Journal of Personality, 1961, 29, 303-320.

Cherlin, A., and Bourque, L. B. Dimensionality and Reliability of the Rotter I-E Scale. Sociometry, 1974, 37(4), 565-582.

Ebel, R. L. Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965.

Gay, G., and Abrahams, R. D. Does the Pot Melt, Boil, or Brew? Black Children and White Assessment Procedures. Journal of School Psychology, 1973, 11(4), 330-340.

Gorsuch, R. L., Henighan, R. P., and Barnard, C. Locus of Control: An Example of Dangers in Using Children's Scales with Children. Child Development, 1972, 43, 579-590.

Harrow, M., and Ferrante, A. Locus of Control in Psychiatric Patients. Journal of Consulting and Clinical Psychology, 1969, 33(5), 582-589.

Hersch, P. D., and Scheibe, K. E. Reliability and Validity of Internal-External Control as a Personality Dimension. Journal of Consulting Psychology, 1967, 31(6), 609-613.

Hjelle, L. A. Social Desirability as a Variable in the Locus of Control Scale. Psychological Reports, 1971, 28, 807-816.

Lamont, J. Item Mood-Level as a Determinant of I-E Test Response. Journal of Clinical Psychology, 1972, 28(2), 190.

Lefcourt, H. M., and Ladwig, G. W. Alienation in Negro and White Reformatory Inmates. The Journal of Social Psychology, 1966, 68, 153-157.

Nowicki, S., Jr., and Strickland, B. R. A Locus of Control Scale for Children. Journal of Consulting and Clinical Psychology, 1973, 40(1), 148-154.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill Book Company, 1967.

Ray, J. J. Projective Tests Can Be Made Reliable: Measuring Need for Achievement. Journal of Personality Assessment, 1974, 38(4), 303-307.

Roscoe, J. T. The Funstat Package in Fortran IV. New York: Holt, Rinehart and Winston, Inc., 1973.

Rotter, J. B. Social Learning and Clinical Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1954.

Rotter, J. B. Generalized Expectancies for Internal versus External Control of Reinforcement. Psychological Monographs, 1966, 80(1), (Whole No. 609).

Valecha, G.K., and Ostrom, T. M. An Abbreviated Measure of Internal-External Locus of Control. Journal of Personality Assessment, 1974, 38(4), 369-376.

Wilson, K. A Further Analysis of Internal Control of Reinforcement. Unpublished Master's Thesis, Kansas State University, 1972.

APPENDIX

M_____

F_____

SOCIAL REACTION INVENTORY

This is a questionniare to find out the way in which certain important events in our society affect different people. Each item consists of a pair of alternatives lettered a or b. Please select the one statement of each pair (and only one) which you more strongly believe to be the case as far as you're concerned. Be sure to select the one you actually believe to be more true than the one you think you should choose or the one you would like to be true. This is a measure of personal belief: obviously there are no right or wrong answers.

Your answers to the items on this inventory are to be recorded by circling the letter of your choice (a or b) for each question. Do not open the booklet until you are told to do so.

Please answer these carefully but do not spend too much time on any one item. Be sure to find an answer for every choice.

In some instances you may discover that you believe both statements or neither one. In such cases, be sure to select the one you more strongly believe to be the case as far as you're concerned. Also try to respond to each item independently when making your choice; do not be influenced by your previous choices.

(1) (2) (3) (4) (5) (6)

REMEMBER

Select that alternative which you personally believe to be more true. I more strongly believe that:

(7) 1. a. Children get into trouble because their parents punish them too much.

b. The trouble with most children nowadays is that their parents are too easy with them.

(8) 2. a. Many of the unhappy things in people's lives are partly due to bad luck.

b. People's misfortunes result from the mistakes they make.

(9) 3. a. One of the major reasons why we have wars is because people don't take enough interest in politics.

b. There will always be wars, no matter how hard people try to prevent them.

(10) 4. a. In the long run people get the respect they deserve in this world.

b. Unfortunately, an individual's worth often passes unrecognized no matter how hard he tries.

(11) 5. a. The idea that teachers are unfair to students is nonsense.

b. Most students don't realize the extent to which their grades are influenced by accidental happenings.

(12) 6. a. Without the right breaks one cannot be an effective leader.

b. Capable people who fail to become leaders have not taken advantage of their opportunities.

(13) 7. a. No matter how hard you try some people just don't like you.

b. People who can't get others to like them don't understand how to get along with others.

(14) 8. a. Heredity plays the major role in determining one's personality.

b. It is one's experiences in life which determine what he is like.

I more strongly believe that:

(15) 9. a. I have often found that what is going to happen will happen.

b. Trusting to fate has never turned out as well for me as making a decision to take a definite course of action.

(16) 10. a. In the case of the well prepared student there is rarely if ever such a thing as an unfair test.

b. Many times exam questions tend to be so unrelated to course work, that studying is useless.

(17) 11. a. Becoming a success is a matter of hard work, luck has little to do with it.

b. Getting a good job depends mainly on being in the right place at the right time.

(18) 12. a. The average citizen can have an influence in government decisions.

b. This world is run by the few people in power, and there is not much the little guy can do about it.

(19) 13. a. When I make plans, I am almost certain that I can make them work.

b. It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.

(20) 14. a. There are certain people who are just no good.

b. There is some good in everybody.

(21) 15. a. In my case getting what I want has little or nothing to do with luck.

b. Many times we might just as well decide what to do by flipping a coin.

(22) 16. a. Who gets to be the boss often depends on who was lucky enough to be in the right place first.

b. Getting people to do the right thing depends upon ability, luck has little or nothing to do with it.

(23) 17. a. As far as world affairs are concerned, most of us are the victims of forces we can neither understand, nor control.

b. By taking an active part in political and social affairs the people can control world events.

I more strongly believe that:

(24) 18. a. Most people don't realize the extent to which their lives are controlled by accidental happenings.

b. There really is no such thing as "luck".

(25) 19. a. One should always be willing to admit his mistakes.

b. It is usually best to cover up one's mistakes.

(26) 20. a. It is hard to know whether or not a person really likes you.

b. How many friends you have depends upon how nice a person you are.

(27) 21. a. In the long run the bad things that happen to us are balanced by the good ones.

b. Most misfortunes are the result of lack of ability, ignorance, laziness, or all three.

(28) 22. a. With enough effort we can wipe out political corruption.

b. It is difficult for people to have much control over the things politicians do in office.

(29) 23. a. Sometimes I can't understand how teachers arrive at the grades they give.

b. There is a direct connection between how hard I study and the grades I get.

(30) 24. a. A good leader expects people to decide for themselves what they should do.

b. A good leader makes it clear to everybody what their jobs are.

(31) 35. a. Many times I feel that I have little influence over the things that happen to me.

b. It is impossible for me to believe that chance or luck plays an important role in my life.

(32) 26. a. People are lonely because they don't try to be friendly.

b. There's not much use in trying too hard to please people, if they like you, they like you.

I more strongly believe that:

(33) 27. a. There is too much emphasis on athletics in high school.

b. Team sports are an excellent way to build character.

(34) 28. a. What happens to me is my own doing.

b. Sometimes I feel that I don't have enough control over the direction my life is taking.

(35) 29. a. Most of the time I can't understand why politicians behave the way they do.

b. In the long run the people are responsible for bad government on a national as well as on a local level.

Three more questions:

Did you find that this questionnaire was:

(36) ___a. Interesting
___b. Neither interesting nor boring
___c. Boring

Did you:

(37) ___a. Like filling out the questionnaire
___b. Neither like nor dislike it
___c. Dislike filling out the questionnaire

Did you think that the questionnaire was:

(38) ___a. Important
___b. Neither important nor unimportant
___c. Unimportant

TEST RELIABILITY AS A FUNCTION OF SUBJECT ATTITUDE TOWARD TEST TAKING

by

GERALD M. EADS II

B.A., Western Washington State College, 1967

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

College of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1976

The effect of subjects' attitude toward test taking on test reliability, score means and standard deviations was assessed. Rotter's Internal-External Locus of Control scale was utilized as an instrument representative of the range of personality and attitude scales found in the literature. Results showed that when adequate sample sizes were utilized Kuder-Richardson Formula 20 reliability estimates consistently decreased as a function of indicated decreased interest in test taking. Standard deviations of the sample subgroup scores also decreased as a function of test taking interest, supporting Nunnally's (1967) contention that reductions in reliability are related to increases in the proportion of error variance relative to the proportion of true score variance. Dominance of the restricted error variance results in the reduction in size of the obtained variance. Mean I-E scale score differences were significant between the subgroups of the samples expressing positive or neutral interest in test taking, while the differences between the subgroups expressing negative interest failed to reach significance. In correspondence with decreasing sample variance and reliabilities this finding supported the hypothesis that less accurate (random) responding on the part of subjects who expressed negative test taking interest would result in the tendency for scale scores to approach the scale midpoint. Findings supported the thesis that there exists a measurable attribute of test taking interest that cannot be appropriately discounted as random error. Reliability research needs to begin to account for the impact of these systematic subject differences.