R E P O R T R E S U M E S

ED 019 714 24 CG 002 093

AN INVESTIGATION OF NON-INDEPENDENCE OF COMPONENTS OF SCORES ON MULTIPLE-CHOICE TESTS. FINAL REPORT.

BY- ZIMMERMAN, DONALD W. BURKHEIMER, GRAHAM J., JR.
EAST CAROLINA COLL., GREENVILLE, N.C.

REPORT NUMBER BR-6-8209 PUB DATE 8 MAR 68
CONTRACT OEC-2-7-068209-0389
EDRS PRICE MF-$0.25 HC-$1.44 34P.

DESCRIPTORS- *TEST RELIABILITY, *MATHEMATICAL MODELS, COMPUTERS, ITEM ANALYSIS, *OBJECTIVE TESTS,

INVESTIGATION IS CONTINUED INTO VARIOUS EFFECTS OF NON-INDEPENDENT ERROR INTRODUCED INTO MULTIPLE-CHOICE TEST SCORES AS A RESULT OF CHANCE GUESSING SUCCESS. A MODEL IS DEVELOPED IN WHICH THE CONCEPT OF THEORETICAL COMPONENTS OF SCORES IS NOT INTRODUCED AND IN WHICH, THEREFORE, NO ASSUMPTIONS REGARDING ANY RELATIONSHIP BETWEEN SUCH COMPONENTS NEED BE MADE. UTILIZING THIS MODEL, AN EXAMINATION OF THE DEPENDENCE OF RELIABILITY ON GROUP HETEROGENEITY REVEALS THAT THE NECESSARY AND SUFFICIENT CONDITIONS UNDER WHICH THE CLASSICAL EQUATION DESCRIBING THIS RELATIONSHIP HOLDS. MODELS IN WHICH ASSUMPTIONS OF THESE CONDITIONS ARE NOT APPLICABLE ARE EXAMINED. COEFFICIENT ALPHA AND OTHER RELIABILITY ESTIMATES ARE EXAMINED. IT IS FOUND THAT THE NECESSARY AND SUFFICIENT CONDITIONS FOR THE RELIABILITY OF A COMPOSITE TEST OF N PARTS TO EQUAL COEFFICIENT ALPHA ARE IDENTICAL TO THOSE UNDER WHICH RELIABILITY OF A TEST LENGTHENED N TIMES IS GIVEN BY THE SPEARMAN-BROWN FORMULA. FURTHER, THESE CONDITIONS ARE ANALOGOUS TO THOSE UNDER WHICH THE RELIABILITY OF A TEST OF N ITEMS IS EQUAL TO KUDER-RICHARDSON FORMULA 20. THE INCREASE IN TEST RELIABILITY WITH INCREASE IN NUMBER OF ALTERNATIVES PER ITEM IS CONSIDERED. THE DERIVED EQUATION DESCRIBING THIS RELATIONSHIP IS EXPRESSED IN A FORM SIMILAR TO THE SPEARMAN-BROWN FORMULA FOR INCREASE IN RELIABILITY WITH INCREASE IN TEST LENGTH. (AUTHOR)

Final Report

Project No. 6-8209
Contract No. OEC2-7-068209-0389

AN INVESTIGATION OF NON-INDEPENDENCE OF COMPONENTS OF SCORES ON MULTIPLE-CHOICE TESTS

Donald W. Zimmerman
and
Graham J. Burkheimer, Jr.

East Carolina University
Greenville, North Carolina

8 March 1968

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

CONTENTS.

SUMMARY

The problem investigated in this research is that of the non-independent error component introduced into scores on multiple-choice tests due to chance guessing success. This class of error operates somewhat differently than the random, independent error described in classical test theory. It differs in two important aspects. First, it operates unidirectionally, in that it can only increase an individual's score on a test; secondly, the usual assumption of independence of the error component of a score can not be made in this case. This is because such error is negatively correlated with an individual's true score, and positively correlated with the same class of error on parallel forms of the test. Since this type of error violates the assumptions made in the derivation of many of the equations of classical test theory, new equations must be derived to adequately describe the effects of this error as reflected in an individual's score.

The method used to investigate the effects of non-independent error is similar to that which has been employed in the past to investigate other classes of error. Basic models are adopted and equations are derived from these models. The major difference in procedure for the purpose of this research is that the assumption of independence is not made. Further, the derived equations are validated against the results of computer simulation techniques. These techniques begin with a prepared distribution of true scores. An IBM 1620 is then programmed to generate one or more classes of error corresponding to a particular true score. The sum of the true score and the error components is then representative of an observed score on a test. This procedure is repeated to generate data that is a simulation of results on parallel forms of a test. The resultant data is then subjected to statistical analysis to obtain descriptive quantities of the scores and their components. Finally, the equations that have been derived to describe the effects of the particular class or classes of error generated by the program are validated against the results.

Three distinct problems were investigated in this research. The first problem is that of the dependence of test reliability upon the heterogeneity of the group tested. In classical theory, this relationship is described as

$$\rho_{oo_B} = 1 - \frac{\sigma^2_{o_A}}{\sigma^2_{o_B}} (1 - \rho_{oo_A}). \quad \ldots$$

In this equation, the reliability of some test administered to a particular group, A , is related to the reliability of the same test administered to another group, B; the quantities in the equation are related to group A or group B by subscripts. A cursory examination of the equation reveals that the reliabilities will be the same if the variances are the same, thus reducing the ratio of the variances to 1.

The derivation of this equation is based on certain assumptions concerning error variance that do not necessarily hold for certain classes of error. One such class of error is the non-independent error under consideration. Another class of error that does not allow such assumptions is the error connected with item sampling as described by Lord (1955). A general equation is derived to describe this relationship, and the equation indicates that the classical equation is valid if and only if the average individual error variance is the same for the two groups.

The second problem investigated is the applicability of coefficient alpha and other reliability estimates to testing situations in general and specifically to testing situations where non-independent error may be present. It is determined that the necessary and sufficient conditions for coefficient alpha to equal test reliability is the same regardless of the class or classes of error in operation. It is shown that coefficient alpha is equal to the reliability of a test if and only if, as the number of measurements increases without limit, the mean over persons of the variance over part-tests of the mean part-test scores over repeated measurements is equal to the variance over part-tests of the means over persons of the mean part-test scores over repeated measurements. Further, the Kuder-Richardson formula 20 (Kuder & Richardson, 1937) is equal to the reliability of a test if and only if all test items satisfy this same condition. Also, the Kuder-Richardson formula 21 is equal to the reliability of a test if and only if all test items satisfy this condition and if in addition, both of the quantities are zero. It is further shown that the conditions under which a composite test of N parts is equal to coefficient alpha are identical to the conditions under which the reliability of a test lengthened N times is given by the Spearman-Brown formula. An equation describing test reliability when these conditions are not met is derived. These results support recent findings by Novick & Lewis (1967).

The final problem considered is that of the dependence of test reliability on the number of alternatives per item. It has been observed that as the number of alternatives per item increases, test reliability likewise increases (Lord, 1944; Carroll, 1945; Plumlee, 1952; Zimmerman & Williams, 1965). It has further been suggested that the relationship can be described by the Spearman-Brown formula, which was originally developed to describe increase in test reliability with increase in test length (Remmers, Karslake & Gage, 1940). It is found that increase in test reliability with increase in number of alternatives is indicated only approximately by the Spearman-Brown formula. An equation is derived that indicates the relationship much more accurately.

The results of this study are probably not directly applicable to the solution of the day-to-day problems of testing; however, the findings do provide a solid foundation for the understanding of the operation of this particular class of error. Due to the widespread use of multiple-choice tests, it is probable that non-independent error due to chance guessing success is present to a large degree in the results of the

majority of tests used in the field of education. An understanding of this class of error is therefore highly desirable. It is also important in the development of test theory that one understand the differential operation of the different classes of error. Hopefully this research will serve not only as a base but also as a stimulation for further research in this area.

INTRODUCTION

Traditional mental test theory has been largely founded on the assumption that errors of measurement are independent random variables with an expected value of zero. When these assumptions are made, certain intercorrelational terms relating the theoretical components of test scores are zero, and resultant equations are considerably simplified.

Two basic models have been employed to derive equations in test theory; however, since the assumption of independence has been made in the use of both models, similar results have been obtained. The first model is based on a definition of an observed score on a test as the sum of two theoretical components, true score and an error component. This relationship is expressed as $O = T + E$. The alternate model defines true score as a limit:

$$T_j = \lim_{H \to \infty} \frac{\sum_{g=1}^{H} O_{gj}}{H} - \lim_{H \to \infty} \frac{\sum_{g=1}^{H} E_{gj}}{H} ,$$

Where T_j represents the true score for some individual j, O_{gj} represents the observed score of that individual on the g^{th} form of H parallel forms of a test, and E_{gj} represents the error component of that observed score. Since, in the limit, the sum of the error components for this individual over repeated testings is zero; the equation simplifies to a definition of an individuals true score as the limiting value of the mean of the frequency distribution of his observed scores over repeated testings with parallel forms of a test (Gulliksen, 1950).

As an example, independence will be assumed under the specifications of the first model. Such an assumption implies that the magnitude and direction of the error component is in no way influenced by the magnitude of its true score counterpart. It follows from this implication that the correlation between true scores and corresponding error components is zero, or $r_{te} = 0$; it further follows that the correlation between error components on parallel forms of a test is zero, or $r_{ee} = 0$. With this conclusion, it is relatively easy to develop the classical equations for the reliability of a test:

$$r_{oo} = \frac{s_t^2}{s_o^2} , \text{ and } r_{oo} = 1 - \frac{s_e^2}{s_o^2} ,$$

where r_{oo} is reliability, s^2 is variance, and the subscripts e, t, and o refer to observed scores, true scores, and error components respectively. Other equations of classical test theory can likewise be derived with ease once the assumption of independence has been made.

Many tests presently used are of the multiple-choice type, and it has been widely recognized that scores on such measuring instruments will reflect to some extent chance guessing success. Such chance guessing success is thus a class of error component of the score. It has not been recognized as widely that this class of error is non-independent. The relationship of this class of error to true score can be seen in a brief example. An individual with a true score of ten on a ten-item true-false test is faced with no items on which to guess the answer. The error component of his observed score as a result of chance guessing success must therefore be zero. Another individual with a true score of zero on the same test may guess on all items, and the error component of his observed score can range from zero to ten, with an expected value of five. Stated generally, the lower an individuals true score on a multiple-choice test, the more items are available for which he can guess the answer; and the greater the number of items on which he guesses, the greater is his expected number of successful guesses. From this it follows that the correlation existing between true score and this class of error is a negative one. It also follows that such error components will be positively correlated with like error components on parallel forms of the test. A further characteristic of this class of error is that it operates unidirectionally. That is, such error can only increment an individual's score.

Due to the non-independent nature of this class of error and to its unidirectional character, the effects of such error can not be effectively described by the equations of the classical test theory. A considerable amount of research has been conducted recently to examine some of the basic properties of this class of error (Burkheimer, 1965; Burkheimer, Zimmerman & Williams, 196?; Williams & Zimmerman, 1966; Zimmerman & Williams, 1965, 1967; Zimmerman, Williams & Rehm, 1966). The present research is an extension of this ongoing investigation.

METHODS

The investigation into the properties of non-independent error proceeded along two lines. The first line was the development of models from which equations describing the effects of this class of error could be derived; the second was the development of computer simulation programs to generate data against which the derived equations could be validated.

In the first stage of the research, the basic models of classical test theory, as described above, were used as a foundation for equation derivation. The major difference in terms of deriving equations from these models in this research was the fact that the assumptions of independence and zero-value expectancy were not made. Early results indicated that the classical equations for reliability could be stated in slightly modified form as follows:

$$r_{oo} = \frac{(a - 1)^2 s_t^2}{a^2 s_o^2} \quad \text{, and} \quad r_{oo} = 1 - \frac{s_e^2}{s_o^2}(1 - r_{ee}),$$

where a is the number of alternatives per item and r_{ee} is the correlation existing between non-independent error components. Other basic equations developed from these models likewise differed from the analogous equations of classical test theory. Some of the effects of non-independent error, however, led to results that were identical to those previously described in classical theory (e.g. the Spearman-Brown formula; Zimmerman & Williams, 1966).

An interesting aspect of the derived equations is that they easily reduce to the classical equations when the number of alternatives increases without limit, thus reducing the probability of a successful guess, 1/a, to a limiting value of zero and eliminating this particular class of error. In the two equations above, for example, it can be readily observed that the limit of the ratio $(a - 1)^2/a^2$, as $a \rightarrow \infty$, is 1; and the equation reduces to the classical equation. Considering the second equation above together with the equation for r_{ee},

$$r_{ee} = \frac{s_t^2}{a^2 s_e^2},$$

it is obvious that the limit, as $a \rightarrow \infty$, of the term $(1 - r_{ee})$ is 1, and again the equation reduces to the classical equation.

In the later stages of the research, an additional model was utilized. This model resembles the second model described above in that it considers the frequency distribution of scores over repeated measurements on parallel forms of a test and in that it considers the parameters of this distribution as the number of repeated measurements increases without limit. The model

differs from the classical model, however, in that it does not introduce the theoretical concepts of true score and error components. The model, which is considered in greater detail in Parts I and II of the following section, considers the mean of a distribution of observed scores for some individual j, $\bar{O}_j$, and the variance of this distribution, $(\sigma_{o_j})^2$. Since no assumptions are made under this model of either independence or non-independence, the model is quite general; and the results obtained from the model are applicable under either assumption.

A computer simulation technique which had been used quite effectively previously was further utilized in the present research. The first program developed was a duplication of the program previously described (Zimmerman & Williams, 1965). An IBM 1620 was used to generate scores on a "multiple-choice test" with the probability of a successful guess set at a value equal to the reciprocal of the number of alternatives per item on "tests" of varying length.

The procedure was simple in concept. A hypothetical distribution of true scores was prepared, and for each true score on a "test" of $\underline{N}$ items, the computer simulated N-T guessing trials. "Correct guesses" were summed, and this sum represented the non-independent error component of an observed score, which was obtained by adding the generated error component to the respective true score. This procedure was repeated to generate results comparable to repeated testings on parallel forms of a test. Finally, the intercorrelation matrix of all possible combinations of observed scores, true scores, and error components was obtained. With these results, the equations derived under the assumption of non-independence were validated.

The second program was an extension of the first, and all steps through the generation of a non-independent error component were the same. In this program, however, an additional error component was generated. The second error component was obtained by a random sampling technique in such a way that the resulting error component complied with the assumptions of independence. Observed score was obtained as before with the exception that it now consisted of the sum of a true score, an independent error component, and a non-independent error component. An intercorrelation matrix was obtained as before. The results of this program were used to validate the equations previously derived for the case in which several classes of error components are reflected in an individual's score on a test (Zimmerman, Williams & Rehm, 1966).

An additional program was placed in operation, which generated error components of scores of the item-sampling class described by Lord (1955). The final results of this program became available only recently and have not yet been fully analyzed.

RESULTS

Part I: Dependence of Test Reliability Upon Heterogeneity of Individual and Group Score Distributions

In the classical test theory the equation

$$(1) \qquad \rho_{oo_B} = 1 - \frac{\sigma^2_{o_A}}{\sigma^2_{o_B}} (1 - \rho_{oo_A})$$

relates the reliability coefficient of a test obtained in a group, A, with a certain variance of observed scores, $\sigma^2_{o_A}$, to the reliability coefficient of that test obtained in a group, B, with a different variance of observed scores, $\sigma^2_{o_B}$ (Gulliksen, 1950). Derivation of the equation depends upon equating error variance in the two groups, In the classical theory the distribution of test scores for a given person over repeated measurements is regarded as the same for all persons and also the same as total error variance for a group of persons. The purpose of this paper is to determine the necessary and sufficient conditions under which the above equation is valid and to examine modifications of the equation which are obtained when these conditions are not met.

Definitions

The subscript g, taking on values from 1 to H, will refer to repeated measurements, and the subscript j, taking on values from 1 to K, will refer to persons. We consider the distribution of observed scores, O_{gj}, for person j, over H independent repeated measurements. We consider the distribution of observed scores, O'_{gj}, for this same person j over H independent repeated parallel measurements. Finally, we consider the means, $\bar{O}_j$ and $\bar{O}'_j$, and the variances, $(\sigma^2_{o_j})$ and $(\sigma^{2'}_{o_j})$ of the distributions of scores for person j, where $\bar{O}_j = \lim_{H \to \infty} \frac{\sum_{g=1}^{H} O_{gj}}{H}$, where $(\sigma^2_{o_j}) = \lim_{H \to \infty} \frac{\sum_{g=1}^{H} O^2_{gj}}{H} - \bar{O}^2_j$, and where similar expressions can be written in prime notation.

Test reliability is defined as follows:

$$(2)\qquad \rho_{oo} = \lim_{H\to\infty} \frac{\dfrac{\sum_{g=1}^{H}\sum_{j=1}^{K} O_{gj}\, O'_{gj}}{HK} - \dfrac{\sum_{g=1}^{H}\sum_{j=1}^{K} O_{gj} \sum_{g=1}^{H}\sum_{j=1}^{K} O'_{gj}}{H^2K^2}}{\sigma_o\, \sigma'_o},$$

where $\sigma_o = \lim_{H\to\infty} \sqrt{\dfrac{\sum_{g=1}^{H}\sum_{j=1}^{K} O^2_{gj}}{HK} - \left(\dfrac{\sum_{g=1}^{H}\sum_{j=1}^{K} O_{gj}}{H^2K^2}\right)^2}$

and where σ'_o is defined by a similar expression in prime notation.

That is, reliability as determined for these K persons is the product-moment correlation between pairs of repeated parallel measurements, as the number of pairs increases without limit.

The frequency distributions of O_{gj} and O'_{gj} over repeated measurements for any person j are assumed to be the same, such that $(\sigma^2_{o_j}) = (\sigma^{2'}_{o_j})$ and $\bar{O}_j = \bar{O}'_j$. It is not assumed, however, that such distributions are the same for different persons. It follows from the above that $\sigma_o = \sigma'_o$, $\sigma_{\bar{o}_j} = \sigma_{\bar{o}'_j}$, and $\rho_{\bar{o}_j\bar{o}'_j} = 1$, where $\sigma_{\bar{o}_j}$ and $\sigma_{\bar{o}'_j}$ refer to the standard deviations of the $\bar{O}_j$ values over all K persons and $\rho_{\sigma_j\sigma'_j}$ is the correlation between $\bar{O}_j$ and $\bar{O}'_j$ for these K persons.

We can consider the covariance term of (2) as the sum of two quantities,

$$(3)\qquad \rho_{oo}\,\sigma_o\,\sigma'_o = \frac{\sum_{j=1}^{K} \rho_{o_jo'_j}\,(\sigma_{o_j})\,(\sigma'_{o_j})}{K} + \rho_{\bar{o}_j\bar{o}_j}\,\sigma_{\bar{o}_j}\,\sigma_{\sigma'_j}.$$

Since the pairs of repeated parallel measurements for person j are independent, it follows that $\rho_{o_j o'_j}(\sigma_{o_j})(\sigma_{o'_j})$, the covariance of O_{gj} and O'_{gj} for person j, is zero for all K persons. Therefore, $\rho_{oo}\sigma_o\sigma'_o = \rho_{\sigma_j\sigma'_j}\sigma_{\sigma_j}\sigma_{\sigma'_j}$. Using the equalities established above, together with a standard theorem,

$$\sigma_o^2 = \overline{(\sigma_{o_j}^2)} + \sigma_{\bar{o}_j}^2 \text{ , where } \underline{(\sigma_{o_j}^2)} = \frac{\sum_{j=1}^{K}(\sigma_{o_j}^2)}{K} ,$$

we arrive at the result

$$(4) \qquad \rho_{oo} = \frac{\sigma_{\bar{o}_j}^2}{\sigma_o^2} = 1 - \frac{\overline{(\sigma_{o_j}^2)}}{\sigma_o^2} .$$

Dependence of Reliability Upon Group Heterogeneity

From (4), $\sigma_{o_A}^2(1 - \rho_{oo_A}) = \overline{(\sigma_{o_j}^2)}_A$ and $\sigma_{o_B}^2(1 - \rho_{oo_B}) = \overline{(\sigma_{o_j}^2)}_B$.

Subtracting the second of these equations from the first and solving for ρ_{oo_B}, we obtain

$$(5) \qquad \rho_{oo_B} = 1 - \frac{\sigma_{o_A}^2}{\sigma_{o_B}^2}(1 - \rho_{oo_A}) + \frac{\overline{(\sigma_{o_j}^2)}_A - \overline{(\sigma_{o_j}^2)}_B}{\sigma_{o_B}^2} .$$

From (5) it is evident that

$$\rho_{oo_B} = 1 - \frac{\sigma_{o_A}^2}{\sigma_{o_B}^2}(1 - \rho_{oo_A}) \Leftrightarrow \overline{(\sigma_{o_j}^2)}_A = \overline{(\sigma_{o_j}^2)}_B \text{ , that is,}$$

change in reliability with change in group heterogeneity is given by (1) if and only if the arithmetic mean, over the K_A persons in group A, of the K_A variances of observed scores over repeated measurements is

equal to the same quantity for the K_B persons in group B.

In classical test theory ($\sigma^2_{o_j}$) is identified with total error variance, a quantity which has been regarded as the same for any two groups. For models in which this condition does not hold, the relationship is given by (5). It should be noted that in the above derivation $\bar{O}_j$ has not been identified with the concept of true score. Therefore, the conclusions reached hold even if O_{gj} is regarded as the sum of components which are linearly correlated (Zimmerman & Williams, 1965, 1966). We will now consider two models in which the frequency distribution of observed scores over repeated measurements varies from person to person.

Item Sampling Model

Following the model presented by Lord (1955) for sampling of test items from a population of items, we can write $\left(\sigma^2_{o_j} \right) = \frac{\bar{O}_j(N - \bar{O}_j)}{N}$, where N is the number of items. Lord identifies the quantity $\bar{O}_j$ with true score, T_j, representing the product of the proportion of items in the population which are known and the number of test items. For any two groups, A and B, we have

$$\overline{\left(\sigma^2_{o_j} \right)_A} = \frac{\mu_{o_A}(N - \mu_{o_A})}{N} - \frac{\sigma^2_{\bar{o}_{jA}}}{N} \quad \text{and} \quad \left(\sigma^2_{o_j} \right)_B = \frac{\mu_{o_B}(N - \mu_{o_B})}{N} - \frac{\sigma^2_{\bar{o}_{jB}}}{N},$$

where μ_{o_A} and μ_{o_B} are the means over all K persons and all H measurements of the O_{gj_A} and O_{gj_B} values. Substituting these results in (5) and factoring, we obtain

$$(6) \qquad \rho_{oo_B} = 1 - \frac{\sigma^2_{o_A}}{\sigma^2_{o_B}}(1 - \rho_{oo_A}) + \frac{\left[(\mu_{o_A} + \mu_{o_B} - N)(\mu_{o_B} - \mu_{o_A})\right] - \left[\sigma^2_{\bar{o}_{jA}} - \sigma^2_{\bar{o}_{jB}}\right]}{N\sigma^2_{o_B}}$$

Only if the right-hand term is zero is (1) valid. Otherwise, we can substitute in (6) for $\sigma^2_{\bar{o}_{jB}}$ its equivalent $\rho_{oo_B}\sigma^2_{o_B}$ and for $\sigma^2_{\bar{o}_{jA}}$ its

equivalent $\rho_{oo_A}\sigma^2_{o_A}$, solve for ρ_{oo_B} and simplify to obtain

$$(7) \qquad \rho_{oo_B} = \frac{N}{N-1}\left[1-\frac{\sigma^2_{o_A}}{\sigma^2_{o_B}}\left(1-\frac{N-1}{N}\rho_{oo_A}\right)+\frac{(\mu_{o_A}+\mu_{o_B}-N)(\mu_{o_B}-\mu_{o_A})}{N\sigma^2_{o_B}}\right].$$

It should be noted that the right hand term becomes zero if $\mu_{o_A}=\mu_{o_B}$ or if $\mu_{o_A} = (N - \mu_{o_B})$, in which case (7) reduces to

$$\rho_{oo_B} = \frac{N}{N-1}\left[1-\frac{\sigma^2_{o_A}}{\sigma^2_{o_B}}\left(1-\frac{N-1}{N}\rho_{oo_A}\right)\right]. \quad \text{Further, if the}$$

right hand term is not zero, but $\sigma^2_{o_{jA}} = \sigma^2_{o_{jB}}$, (7) reduces to

$$\rho_{oo_B} = 1-\frac{\sigma^2_{o_A}}{\sigma^2_{o_B}}\left(1-\rho_{oo_A}\right)+\frac{(\mu_{o_A}+\mu_{o_B}-N)(\mu_{o_B}-\mu_{o_A})}{N\sigma^2_{o_B}}.$$

Chance Success Model

Using the model for chance success presented previously (Zimmerman & Williams, 1965; Burkheimer, Zimmerman & Williams, 1967) we can write $(\sigma^2_{o_j}) = p(N - \bar{o}_j)$, where p, the probability of chance success, is assumed to be the same for all persons. Finding the arithmetic mean for groups A and B as before gives $(\sigma^2_{o_j})_A = p(N-\mu_{o_A})$ and $(\sigma^2_{o_j})_B = p(N-\mu_{o_B})$. Substituting these results in (5) we have

$$(8) \qquad \rho_{oo_B} = 1 - \frac{\sigma_{o_A}^2}{\sigma_{o_B}^2}\left(1 - \rho_{oo_A}\right) + \frac{p\left(\mu_{o_B} - \mu_{o_A}\right)}{\sigma_{o_B}^2} .$$

Only under the conditions that $\mu_{o_A} = \mu_{o_B}$ or that $p = 0$ is (1) obtained.

Part II:

Conditions Under Which Coefficient Alpha Equals Test Reliability: the Case of Heterogeneous Score Distributions and Correlated Score Components

Since the original derivation by Kuder and Richardson (1937) of the formula which Cronbach designated as coefficient alpha the assumptions required have been more fully explicated (Jackson & Ferguson, 1941, Guttman, 1945, Gulliksen, 1950, Cronbach, 1951, Novick & Lewis, 1967). Derivations have been based on the classical test theory model in which true scores and error scores are uncorrelated and error scores on parallel tests are uncorrelated. The present derivation of coefficient alpha is somewhat different from those which have appeared in the literature, reveals the necessary and sufficient conditions under which this quantity is equal to the reliability of the test, and indicates that the formula is valid under certain conditions where the classical test theory model is not applicable.

Definitions

The subscript i, taking on values from 1 to N, will refer to the parts of a test.

We consider the distributions of observed part-test scores, O_{gji}, and total test scores, O_{gj}, for person j over H independent repeated measurements. We further consider the distributions of observed part-test scores, O'_{gji}, and total test scores, O'_{gj}, for this same person j over H independent repeated parallel measurements. Also, we consider the quantities $\bar{O}_j$ and $(\sigma^2_{\bar{O}j})$, which have been defined above. Analogous quantities, $\bar{O}_{ji}$ and $(\sigma^2_{\bar{O}ji})i$, define the mean and variance of part-test scores for person j over repeated measurements.

The reliability of a test is defined as above. The reliability of a part-test, $(\rho_{OO})i$, is defined in the same way, in terms of O_{gji} and O'_{gji}.

It is assumed that, as H increases without limit, the frequency distributions of O_{gj} and O'_{gj} are the same for any person j, such that $(\sigma^2_{Oj}) = (\sigma^2_{Oj}')$ and $\bar{O}_j = \bar{O}'_j$. It is not assumed that such distributions are the same for different persons (cf. Zimmerman, Williams & Burkheimer, 1968). The same concept applies to the part-test scores, O_{gji} and O'_{gji}. It follows from the above that $\sigma_O = \sigma'_O$, $\sigma_{\bar{O}j} = \sigma_{\bar{O}j}'$, and $\rho_{\bar{O}j\bar{O}j}' = 1$. Likewise, these relationships hold for the part-test scores.

Another expression for test reliability is given in (4) above, and an expression for the reliability of part tests can be written as

$$(\rho_{oo})_i = \frac{(\sigma^2_{\bar{o}_{ji}})_i}{\sigma^2_{o_i}} = 1 - \frac{\overline{(\sigma^2_{o_{ji}})_i}}{\sigma^2_{o_i}}$$. Here, the terms used are

analogous to their counterparts defined above, where the subscript i indicates that they refer only to the ith part of the test.

The following symbols will also be used:

$$\alpha = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N} \sigma^2_{o_i}}{\sigma^2_o}\right]$$

$$= \bar{O}_j = \frac{\sum_{i=1}^{N} \bar{O}_{ij}}{N} = \frac{\bar{O}_j}{N}$$

$$= \bar{O}_i = \frac{\sum_{j=1}^{K} \bar{O}_{ij}}{K}$$

$\sigma^2_{\bar{o}_j}$ is the variance of $\bar{\bar{O}}_j$ over all K persons.

$\sigma^2_{\bar{o}_i}$ is the variance of $\bar{\bar{O}}_i$ over all N part-tests.

$(\sigma^2_{\bar{o}_{ij}})_j$ is the variance of $\bar{O}_{ij}$ for person j over all N part-tests.

$\sigma^2_{\bar{o}_{ij}}$ is the variance of $\bar{O}_{ij}$ over all K persons and all N part-tests.

$\rho_{o_n o_m}\sigma_{o_n}\sigma_{o_m}$ is the covariance of O_{gji} for any two part-tests.

$\rho_{\bar{o}_{jn}\bar{o}_{jm}}\sigma_{\bar{o}_{jn}}\sigma_{\bar{o}_{jm}}$ is the covariance of $\bar{O}_{ij}$ for any two part-tests.

Derivation of Coefficient Alpha

Total observed variance can be written as follows:

$$\sigma^2_o = \sum_{i=1}^{N} \sigma^2_{o_i} + \sum_{\substack{n=1 \\ n\neq m}}^{N} \sum_{m=i}^{N} \rho_{o_n o_m}\sigma_{o_n}\sigma_{o_m} \cdot \text{Variance of } \bar{O}_j$$

can likewise be written $\sigma^2_{\bar{o}_j} = \sum_{i=1}^{N} (\sigma^2_{\bar{o}_{ij}})_i + \sum_{\substack{n=1 \\ n\neq m}}^{N} \sum_{m=1}^{N} \rho_{\bar{o}_{jn}\bar{o}_{jm}}\sigma_{\bar{o}_{jn}}\sigma_{\bar{o}_{jm}}$.

Subtracting the second equation from the first yields

$$\sigma_o^2 - \sigma_{\bar{o}_j}^2 = \sum_{i=1}^{N}\left[\sigma_{o_i}^2 - (\sigma_{\bar{o}_{ij}}^2)_i\right] + \sum_{\substack{n=1 \\ n\neq m}}^{N}\sum_{m=1}^{N}(\rho_{o_n o_m}\sigma_{o_n}\sigma_{o_m} - \rho_{\bar{o}_{jn}\bar{o}_{jm}}\sigma_{\bar{o}_{jn}}\sigma_{\bar{o}_{jm}}).$$

Relating the second expression in the right hand member of this equation to equation (3) and following the same argument, it is evident that this expression is zero. Substituting $\rho_{oo}\,\sigma_o^2$ for its equivalent value $\sigma_{\bar{o}_j}^2$ and reducing, we arrive at

$$(9)\qquad \rho_{oo} = 1 - \frac{\sum_{i=1}^{N}\sigma_{o_i}^2}{\sigma_o^2} + \frac{\sum_{i=1}^{N}(\sigma_{\bar{o}_{ij}}^2)_i}{\sigma_o^2} = 1 - \frac{\sum_{i=1}^{N}\sigma_{o_i}^2\left[1-(\rho_{oo})_i\right]}{\sigma_o^2}$$

Since $\sigma_{\bar{o}_{ij}}^2 = \overline{(\sigma_{\bar{o}_{ij}}^2)_i} + \sigma_{\bar{\bar{o}}_i}^2$ and $\sigma_{\bar{o}_{ij}}^2 = \overline{(\sigma_{\bar{o}_{ij}}^2)_j} + \sigma_{\bar{\bar{o}}_j}^2$, it follows that $\overline{(\sigma_{\bar{o}_{ij}}^2)_i} = \overline{(\sigma_{\bar{o}_{ij}}^2)_j} + \sigma_{\bar{o}_j}^2 - \sigma_{\bar{\bar{o}}_i}^2$. Substituting this result in (9) gives

$$(10)\qquad \rho_{oo} = 1 - \frac{\sum_{i=1}^{N}\sigma_{o_i}^2}{\sigma_o^2} + \frac{\overline{N(\sigma_{\bar{o}_{ij}}^2)_j}}{\sigma_o^2} + \frac{N\sigma_{\bar{\bar{o}}_j}^2}{\sigma_o^2} - \frac{N\sigma_{\bar{\bar{o}}_i}^2}{\sigma_o^2}.$$

From the above definition of $\bar{\bar{O}}_j$ it is evident that

$$\frac{N\sigma^2_{\bar{\bar{O}}_j}}{\sigma^2_O} = \frac{\sigma^2_{\bar{O}_j}}{N\sigma^2_O} = \frac{\rho_{oo}}{N} .$$ Substituting in (10) and solving for

ρ_{oo} leads to the following result:

$$(11) \quad \rho_{oo} = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N}\sigma^2_{O_i}}{\sigma^2_O} + \frac{N\left\{\overline{(\sigma^2_{O_{ij}})_j} - \sigma^2_{\bar{\bar{O}}_i}\right\}}{\sigma^2_O}\right] ,$$

which, except for the right hand term, is identical to α.

We can write $\alpha = \rho_{oo} \Leftrightarrow \overline{(\sigma^2_{O_{ij}})_j} = \sigma^2_{\bar{\bar{O}}_i}$. In other words,

coefficient alpha is equal to the reliability of the test if and only if, as the number of measurements increases without limit, the mean over persons of the variances over part-tests of the mean part-test scores over repeated measurements is equal to the variance over part tests of the means over persons of the mean part-test scores over repeated measurements.

In the classical test theory model $O_{gj} = T_j + E_{gj}$

and $T_j = \lim_{H\to\infty} \frac{\sum_{g=1}^{H} O_{gj}}{H} = \bar{O}_j$, where T_j and E_{gj} are true and

error components of scores. In the present derivation

the quantity $\lim_{H \to \infty} \frac{\sum_{g=1}^{H} O_{gj}}{\quad}$ has not been identified with the concept of true score. If, however, $\bar{O}_j$ is a linear function of T_j the conclusions reached by Novick & Lewis (1967) hold even when true and error components of scores are correlated and when the distributions of scores over repeated measurements are different from person to person (Burkheimer, Zimmerman & Williams, 1967, Williams & Zimmerman, 1966, Zimmerman & Williams, 1965, 1966). That is, even in this case the necessary and sufficient condition for coefficient alpha to equal test reliability is that all part-tests be essentially t-equivalent, as defined by Novick & Lewis. We can now write these conditions in the following form: for every j, $\bar{O}_{nj} = \bar{O}_{mj} + C_{nm}$, for any n, m - or, that the mean part-test scores over repeated measurements for person j differ at most by constants, which must be the same for all K persons, but not necessarily the same for all part-tests. It follows, then, that for all n and m, $\rho_{O_n O_m} = (\rho_{oo})_n = (\rho_{oo})_m$, $\sigma^2_{O_n} = \sigma^2_{O_m}$, and $\rho_{\bar{O}_{jn}\bar{O}_{jm}} = 1$. In other words, the variances of all parts are the same, the intercorrelations among all parts are the same and equivalent to the reliability of each part, and the correlation between the $\bar{O}_{ij}$ values is unity for all pairs of part-tests. Thus, all parts of the test must be parallel measurements in the usual sense of "parallel" in test theory, except for differences in scores by a constant.

The present results are consistent with those obtained by Jackson and Ferguson (1941) and Gulliksen (1950). The above condition may be expressed by saying that over repeated measurements the average covariance between parts within a test must equal the average covariance for each part. Still another way of expressing the condition is this: a single administration of a test can yield an estimate of reliability only if the degree of correlation between parts within the test provides information as to the correlation which would be obtained over repeated measurements.

When individual items are considered α is equivalent to the Kuder-Richardson formula 20. That formula, then, is equal to the reliability of a test if and only if all test items satisfy the above conditions.

Equivalence of Coefficient Alpha and the Spearman-Brown Formula

The conditions under which the reliability of a composite test of N parts is equal to coefficient alpha are identical to the conditions under which the reliability of a test lengthened N times is given by the Spearman-Brown formula. The latter condition is sometimes expressed

by saying that the parts added to the original test must all be parallel. But this is just the condition required for coefficient alpha to equal the reliability of a composite test. No requirement need be made in either case as to the degree of correlation between true and error components of scores (Zimmerman & Williams, 1966).

In the case in which N refers to items, the following statement can be made: the conditions under which the reliability of a test of N items is given by the Kuder-Richardson formula 20 are identical to the conditions under which the reliability of a one-item test lengthened N times is given by the Spearman-Brown formula. The relationship can be seen by letting $(\rho_{oo})_i$ be the reliability of one part-test, or one item, before the test is lengthened. If the above conditions are met $\sum_{i=1}^{N} \sigma_{o_i}^2 = N\sigma_{o_i}^2$ and we can substitute in α its equivalent expression in terms of part-test variances and covariances and write

$$(12) \qquad \rho_{oo} = \frac{N}{N-1}\left[1 - \frac{N\sigma_{o_i}^2}{N\sigma_{o_i}^2\left[1+(N-1)(\rho_{oo})_i\right]}\right] = \frac{N(\rho_{oo})_i}{1+(N-1)(\rho_{oo})_i} .$$

Relation of the Kuder-Richardson Formula 20 to the Kuder-Richardson Formula 21

Assume that the conditions for $\alpha = \rho_{oo}$ are met. Then, when individual test items are considered,

$$\alpha = \rho_{oo} = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N} = \quad = \quad (1-0)}{\sigma_o^2}\right] .$$ Expanding this expression and reducing, we obtain

$$(13) \qquad \rho_{oo} = \frac{N}{N-1}\left[1 - \frac{\mu_o(N - \mu_o)}{N\sigma_o^2} + \frac{N\sigma^2_{\bar{o}_i}}{\sigma_o^2}\right],$$

where μ_o is the mean of observed scores over all K persons and all H repeated measurements. When $\sigma^2_{\bar{\bar{o}}_i} = 0$, this result is equivalent to the Kuder-Richardson formula 21. In other words, the condition required for the K.-R. formula 20 to equal the reliability of the test is that $\overline{(\sigma^2_{\bar{o}_{ij}})_j} = \sigma^2_{\bar{\bar{o}}_i}$, and the condition required for K.-R. 21 is that $\overline{(\sigma^2_{\bar{o}_{ij}})_j} = \sigma^2_{\bar{\bar{o}}_i} = 0$.

Derivation of Coefficient Alpha from Item Sampling Model

An interpretation of the Kuder-Richardson formula 21 can be based on the model presented by Lord (1955) for the sampling of test items from a population of items. Lord demonstrated that an individual's standard error of measurement is estimated by $\frac{\bar{o}_{gj}(N - \bar{o}_{gj})}{N-1}$ and that when that quantity is averaged over all persons and substituted in the classical equation $\rho_{oo} = 1 - \frac{\sigma_e^2}{\sigma_o^2}$, the K.-R. formula 21 is obtained. The K.-R. formula 21, in other words, can be interpreted as an equation for test reliability for the case in which the observed score distribution over repeated measurements is identified with variations in score resulting from item sampling.

It should be noted that the above requirement that $\overline{(\sigma^2_{\bar{o}_{ij}})_j} = \sigma^2_{\bar{\bar{o}}_i} = 0$ is implicit in this model. From the concept of item sampling employed by Lord, it follows that, as H increases without limit, for any j, $\bar{\bar{o}}_j = \bar{o}_{ij}$ for all i, such that $(\sigma^2_{\bar{o}_{ij}})_j$ becomes zero for all j. Further, $\bar{\bar{o}}_i$ becomes the same for all i, such that $\sigma^2_{\bar{\bar{o}}_i} = 0$.

It should be noted that K.-R. 20 would be equally applicable to this model.

Let us assume now, that the $\underline{N}$ items of a test are drawn from $\underline{N}$ distinct populations and that the values of $\bar{0}_{ij}$ for person j may differ. Then, the distribution of observed scores over repeated measurements is given by a Poisson sequence of trials, $(\sigma^2_{o_j}) = N\bar{\bar{0}}_j(1 - \bar{\bar{0}}_j) - N(\sigma^2_{\underline{o}_{ij}})_j$.

Averaging this quantity over all K persons, substituting the result in equation (4), and simplifying we obtain (11).

In other words, equation (11) gives test reliability for the case in which the $\underline{N}$ items of a test are sampled from $\underline{N}$ distinct populations of items. The K.-R. formula 20 gives test reliability only if

$\overline{(\sigma^2_{\underline{o}_{ij}})_j}$ and $\sigma^2_{o_i}$ are equal. Finally, the Kuder-Richardson formula 21 gives test reliability only if both these quantities are zero.

Part III: Dependence of Reliability of Multiple-Choice Tests Upon Number of Choices Per Item: Prediction From the Spearman-Brown Formula

It has been known for some time that the reliability of multiple-choice tests is influenced by the number of choices per item (Remmers, Karslake & Gage, 1940; Lord, 1944; Carroll, 1945; Plumlee, 1952). Since the probability of chance success on an item is $\frac{1}{a}$, where a is the number of choices per item, it is to be expected that error variance introduced by chance success is a decreasing function of number of choices and test reliability is an increasing function of number of choices.

Remmers and his associates suggested the relationship could be described by the Spearman-Brown formula, which is known to indicate increase in reliability with increase in test length. The formula is

$$r_{noo} = \frac{nr_{oo}}{1 + (n - 1)\, r_{oo}} \quad , \tag{14}$$

where r_{oo} is the original reliability, r_{noo} is the reliability of the test of increased length, and n is the number of times the test is increased in length. Remmers showed empirically that the reliability of various tests is approximated by this function, when n refers to increase in number of choices instead of test length. It has been pointed out, however, that there is no theoretical basis for predicting this result (Lord, 1944; Guilford, 1950; Gulliksen, 1950).

Computer Simulated Results

In a previous paper (Zimmerman & Williams, 1965) a computer program was used to simulate guessing error in multiple-choice tests. Distributions of assumed true scores were prepared, and error scores were generated on the basis of the probabilities to be expected from chance success due to guessing. The error scores were added to true scores to obtain observed scores. Finally, product-moment correlations between different sets of observed scores obtained by repeating the procedure several times gave an indication of test reliability.

The results of this procedure for tests differing in length and number of choices are shown in Table 1. The data in this table can be used to examine the effect of increased test length, as well as increased number of choices, upon reliability. Apparently, there is an interaction between the effects of test length and number of choices.

TABLE 1

COMPUTER SIMULATED RESULTS FOR RELIABILITY

	N = 10 a = 2	N = 10 a = 5	N = 100 a = 2	N = 100 a = 5
r_{oo}*	.44	.74	.89	.97
r_{oo}**			.89	.97
r_{oo}***		.76		.97
r_{oo}****		.66		.95

*Reliability given by computer program.
**Reliability given by substituting .44 or .74 in Equation 14.
***Reliability given by substituting .44 or .89 in Equation 18.
****Reliability given by substituting .44 or .89 in Equation 14.
where n = 2.5.

For short tests (N = 10) reliability increases greatly with increase in number of choices (.44 to .74). For long tests (N = 100) reliability increases slightly with number of choices (.89 to .97). Also, for 2 choices, reliability increases greatly with test length (.44 to .89). And for 5 choices reliability increases to a lesser degree with test length (.74 to .97).

From the table it is seen that the Spearman-Brown formula describes the increase in reliability with increase in test length for both the 2-choice test and the 5-choice test (Zimmerman & Williams, 1966). Consider, now, Remmers' suggestion that the same formula describes increase in reliability with increase in number of choices. The results in the table show that there is a greater discrepancy, although the predicted value for the longer test is close to that indicated by the program.

Increased Reliability As a Function of Increased Number of Choices

It is possible to derive a simple equation showing the effect of increasing the number of choices upon reliability for the case in which only error due to guessing is present. Reliability is given by

$$(15) \qquad r_{oo} = \frac{(a - 1)\, s_t^2}{(a - 1)\, s_t^2 + N - \bar{T}} \quad ,$$

where a is the number of choices, s_t^2 is the variance of true scores, N is the number of items, and $\bar{T}$ is the mean of true scores. This equation gives the value which is approximated by the computer simulation method described above (Burkheimer, 1965; Burkheimer, Zimmerman, & Williams, 1967). When the number of choices is increased, we can write

$$(16) \qquad r_{oo}' = \frac{(a' - 1)\, s_t^2}{(a' - 1)\, s_t^2 + N - \bar{T}} \quad .$$

where r_{oo}' is the reliability for the test with increased number of choices, a is the original number of choices, a' is the increased number of choices, and the other symbols are as defined above. Solving (15) for s_t^2 gives

$$(17) \qquad s_t^2 = \frac{(N - \bar{T})\, r_{oo}}{(a - 1)\, (1 - r_{oo})} \quad .$$

Substituting this result in (16) and simplifying, we have

$$(18) \qquad r_{oo}' = \frac{(a' - 1)\, r_{oo}}{(a' - 1) + (a - a')\, r_{oo}} \quad .$$

The data presented in Table 1 show that substitution in this equation yields results close to those indicated by the computer program. The accuracy is greater than that obtained by using (14) and of the same order as that obtained by using (14) for increased test length.

If the method employed by Remmers were valid, the ratio $\frac{a'}{a}$ would be comparable to n in (14), which could be written in this form:

$$(19) \qquad r_{oo}' = \frac{\left(\frac{a'}{a}\right) r_{oo}}{1 + \left(\frac{a'}{a}\right) - 1 \quad r_{oo}} \quad .$$

Simplifying, we obtain the following result

$$(20) \qquad r_{oo'} = \frac{a' r_{oo}}{a + (a' - a) r_{oo}} \quad ,$$

which can be compared to (18). It is seen, therefore, that equation (18) differs from the modification of the Spearman-Brown formula suggested by Remmers only by subtraction of 1 from the a' factor in the numerator and the a term in the denominator. If both a' and a were large (14) and (18) would give nearly the same results. For multiple-choice tests, however, a' and a are relatively small, and some discrepancy can be expected.

Dividing both numerator and denominator of (18) by $a - 1$ gives

$$(21) \qquad r_{oo'} = \frac{\frac{(a' - 1)}{(a - 1)} r_{oo}}{\frac{(a - 1)}{(a - 1)} + \frac{(a' - a)}{(a - 1)} r_{oo}} \quad .$$

If, now, we define A as the ratio $\frac{(a' - 1)}{(a - 1)}$ and simplify, we have

$$(22) \qquad r_{oo'} = \frac{A r_{oo}}{1 + (A - 1) r_{oo}} \quad ,$$

which has the same form as the Spearman-Brown formula. In other words, Remmers' suggestion is valid if we employ the ratio $\frac{(a' - 1)}{(a - 1)}$ in the Spearman-Brown formula, but not if we employ the ratio $\frac{a'}{a}$. It should be noted that the above equations apply only to the case in which differences in reliability result from chance success due to guessing.

Dependence of Correlation Between Error Scores On Parallel Forms Upon Number of Choices

It is of interest that an equation showing the dependence of the correlation between error scores on parallel forms of a test upon number of choices can also be derived. This quantity has been assumed to be zero in the classical theory of mental tests. However, when chance success due to guessing is present, as in the case of most multiple-choice tests, it can be shown that it is positive in value, that it

decreases with number of choices, and that the relationship is indicated by an equation similar to (18).

Correlation between error scores on parallel forms is in fact given by the following equation:

$$(23) \qquad r_{ee} = \frac{s_t^2}{s_t^2 + (a - 1)(N - \bar{T})} \quad ,$$

where the symbols are as defined above (Burkheimer, 1965; Burkheimer, Zimmerman & Williams, 1967). When number of choices is increased, we can write

$$(24) \qquad r_{ee'} = \frac{s_t^2}{s_t^2 + (a' - 1)(N - \bar{T})} \quad .$$

Solving (23) for s_t^2 gives

$$(25) \qquad s_t^2 = \frac{r_{ee}\,(a - 1)(N - \bar{T})}{(1 - r_{ee})} \quad .$$

Substituting (25) in (24) and simplifying leads to this result:

$$(26) \qquad r_{ee'} = \frac{(a - 1)\, r_{ee}}{(a' - 1) - (a' - a)\, r_{ee}} \quad .$$

Dividing both numerator and denominator of (26) by a' - 1 gives

$$(27) \qquad r_{ee'} = \frac{\frac{(a - 1)}{(a' - 1)}\, r_{ee}}{\frac{(a' - 1)}{(a' - 1)} - \frac{(a' - a)}{(a' - 1)}\, r_{ee}} \quad .$$

If we define $B = \frac{1}{A} = \frac{(a - 1)}{(a' - 1)}$ and simplify, we have

$$(28) \qquad r_{ee}' = \frac{Br_{ee}}{1 + (B - 1)\, r_{ee}}$$

which, again, has the same form as the Spearman-Brown formula. There exists no analogue of this equation in the classical theory of mental tests. From (26) and (28) it is clear that the degree of correlation between error scores on parallel forms decreases with an increase in the number of choices.

The results given by the computer program for r_{ee} are shown in Table 2. Equation (26) predicts accurately the effect of increasing number of choices upon r_{ee}. Another fact of interest shown in the table is that, if r_{ee} is treated as a reliability coefficient, the Spearman-Brown formula indicates accurately the change in its value with change in test length (Zimmerman & Williams, 1966). For longer tests the correlation between error scores on parallel forms becomes higher in value, and the degree of change is indicated by the Spearman-Brown formula.

TABLE 2

COMPUTER SIMULATED RESULTS FOR CORRELATION BETWEEN ERROR SCORES ON PARALLEL FORMS

	N = 10 a = 2	N = 10 a = 5	N = 100 a = 2	N = 100 a = 5
r_{ee}*	.46	.17	.89	.65
r_{ee}**			.90	.67
r_{ee}***		.18		.66

*Value given by computer program.
**Value given by substituting .46 or .17 in Equation (14).
***Value given by substituting .46 or .89 in Equation (26).

CONCLUSIONS

The equations derived as a result of this research are theoretical in nature. Even using the model which does not introduce the theoretical concepts of true score and error components, it is necessary to rely on a theoretical distribution of observed scores that would be obtained if one were able to measure individuals repeatedly on an extremely large number of parallel forms of some test. Further, the validation procedure used is a gross simplification of the empirical testing situation. No matter how complex a computer simulation technique - in this research the complexity was greatly limited by the limited capacity of the computer used - it is highly unlikely that one could program all the determinents of human test-taking behavior.

For these reasons, it is not likely that the results of this research will have any immediate application to the day-to-day problems of testing. It is interesting to note, however, that equations derived from the model of non-independence have been tested on the results obtained in an empirical situation and that the implications of the theory of non-independence were supported (Zimmerman, et al., 1966). Certain conclusions can be drawn, of course, from the results of this work that are applicable to an empirical situation. The classical equation relating reliability of a test as established for one group to the reliability that should be obtained for that test with another group with differing score variability should be used with caution when the test is of the multiple-choice type. One should be aware of the limiting assumptions necessary for coefficient alpha and related equations to be appropriate estimates of test reliability; and certainly the users of multiple-choice tests should be aware that tests with a greater number of alternatives per item are generally more reliable measuring instruments.

It is also considered quite important that one have an understanding of the implications of the presence of this particular class of error. Use of multiple-choice tests is widespread, and it is precisely this type of test that is subject to the appearance of non-independent error due to chance guseesing success. Since this type of error operates somewhat differently than the error that has traditionally been considered in test theory, new equations describing the effects of such error are needed. Such equations taken with those previously developed in classical theory should lead to a fuller understanding of an individual's score on a test.

This research is a beginning toward a more general theory of mental tests that takes into consideration the many classes of error reflected in an individual's score. Hopefully it will serve as a foundation for further research into non-independent error operating both alone and in combination with other classes of error. More important, it may hopefully serve as a catalyst in precipitating further research.

REFERENCES

Burkheimer, G. J. Some effects of non-independent error in multiple-choice tests: A binomial model. Unpublished M.A. thesis, East Carolina College, Greenville, N. C., 1965.

Burkheimer, G. J., Zimmerman, D. W. & Williams, R. H. The maximum reliability of a multiple-choice test as a function of number of items, number of choices, and group heterogeneity. Journal of Experimental Education, 1967, 35, 89-94.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10, 1-19.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Guilford, J. P. Psychometric methods. (2nd ed.) New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Jackson, R. W. B. & Ferguson, G. A. Studies on the reliability of tests. Bulletin #12, Dept. of Educational Research, University of Toronto, 1941.

Kuder, G. F. & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lord, F. M. Reliability of multiple-choice tests as a function of number of choices per item. Journal of Educational Psychology, 1944, 35, 175-180.

Lord, F. M. Sampling fluctuations resulting from the sampling of test items. Psychometrika, 1955, 20, 1-22.

Novick, M. R. & Lewis, C. Coefficient alpha and the reliability of composite measurements. Psychometrika, 1967, 32, 1-13.

Plumlee, L. B. The effect of difficulty and chance success on item-test correlation and on test reliability. Psychometrika, 1952, 17, 69-86.

Remmers, H. H., Karslake, R. & Gage, N. L. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula: I. Journal of Educational Psychology, 1940, 31, 583-590.

Williams, R. H. & Zimmerman, D. W. An extension of the Rulon formula for test reliability: The case of correlated true and error components of scores. Research Bulletin RB-66-55, Educational Testing Service, 1966.

Zimmerman, D. W. & Williams, R. H. Chance success due to guessing and non-independence of true scores and error scores in multiple-choice tests: Computer trials with prepared distributions. Psychological Reports, 1965, 17, 159-165.

Zimmerman, D. W. & Williams, R. H. Generalization of the Spearman-Brown formula for test reliability: The case of non-independence of true scores and error scores. British Journal of Mathematical and Statistical Psychology, 1966, 19, 271-274.

Zimmerman, D. W. & Williams, R. H. Independence and non-independence of true scores and error scores in mental tests: Assumptions in the definition of parallel forms. Journal of Experimental Education, 1967, 35(3), 59-64.

Zimmerman, D. W., Williams, R. H. & Burkheimer, G. J. Dependence of reliability of multiple-choice tests upon number of choices per item: Prediction from the Spearman-Brown formula. Psychological Reports, 1966, 19, 1239-1243.

Zimmerman, D. W., Williams, R. H. & Burkheimer, G. J. Dependence of test reliability upon heterogeneity of individual and group score distributions. Educational and Psychological Measurement, 1968, 28(1), 41-46.

Zimmerman, D. W., Williams, R. H. & Rehm, H. H. Test reliability when error scores consist of independent and non-independent components. Journal of Experimental Education, 1966, 35(1), 74-78.

Zimmerman, D. W., Williams, R. H., Rehm, H. H. & Elmore, W. Empirical estimates of intercorrelations among the components of scores on multiple-choice tests. Psychological Reports, 1966, 19, 651-654.

(TOP)

ERIC REPORT RESUME

CLEARINGHOUSE ACCESSION NUMBER	RESUME DATE	P.A.	T.A.
	04-10-68		

IS DOCUMENT COPYRIGHTED? YES [X] NO [ ]

ERIC REPRODUCTION RELEASE? YES [ ] NO [X]

TITLE

AN INVESTIGATION OF NON-INDEPENDENCE OF SCORES ON MULTIPLE-CHOICE TESTS

PERSONAL AUTHOR(S)

Donald W. Zimmerman & Graham J Burnheimer, Jr.

INSTITUTION (SOURCE) | SOURCE CODE

East Carolina University, Greenville, N. C., Department of Psychology

REPORT/SERIES NO.

OTHER SOURCE | SOURCE CODE

OTHER REPORT NO.

OTHER SOURCE | SOURCE CODE

OTHER REPORT NO.

PUB. DATE 10-Apr-68 | CONTRACT/GRANT NUMBER OEC2-7-068209-0389

PAGINATION, ETC.

32p.

RETRIEVAL TERMS

Non-independent error; computer simulation; reliability and group heterogeneity; item sampling model; chance success model; coefficient alpha and reliability; coefficient alpha and Spearman-Brown formula; Kuder-Richardson formula 20 and Kuder-Richardson formula 21; reliability and number of choices per item.

IDENTIFIERS

ABSTRACT

Investigation is continued into various effects of non-independent error introduced into multiple-choice test scores as a result of chance guessing success. A model is developed in which the concept of theoretical components of scores is not introduced and in which, therefore, no assumptions regarding any relationship between such components need be made. Utilizing this model, an examination of the dependence of reliability on group heterogeneity reveals that the necessary and sufficient conditions under which the classical equation describing this relationship holds. Models in which assumptions of these conditions are not applicable are examined. Coefficient alpha and other reliability estimates are examined. It is found that the necessary and sufficient conditions for the reliability of a composite test of n parts to equal coefficient alpha are identical to those under which reliability of a test lengthened n times is given by the Spearman-Brown formula. Further, these conditions are analogous to those under which the reliability of a test of n items is equal to Kuder-Richardson formula 20. The increase in test reliability with increase in number of alternatives per item is considered. The derived equation describing this relationship is expressed in a form similar to the Spearman-Brown formula for increase in reliability with increase in test length.