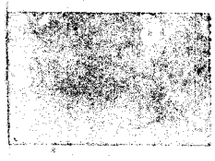
14

# GENERALIZABILITY ESTIMATES FOR DIFFERENCE SCORES:

# AN ASPECT OF THE CONSTRUCT VALIDITY OF TESTS

Hans-Magne Eikeland

University of Oslo



Oslo, March 1973

This report is a preliminary version issued for limited circulation. Corrections, criticisms, and suggestions for revision are solicited. The report should not be cited as a reference without the specific permission of the author.

# GENERALIZABILITY ESTIMATES FOR DIFFERENCE SCORES: AN ASPECT OF THE CONSTRUCT VALIDITY OF TESTS.

Hans-Magne Eikeland, University of Oslo

## Introduction.

Over the years test theory has almost exclusively been concerned with the linear combination of test form scores called the sum. One of the questions most frequently asked of test data is to what extent different test forms combined in such a sum measure the same construct or trait.

It is well known that one could also ask test data to what extent different tests measure different constructs. Until recently, however, this problem of differential construct validity, as it will be called in this paper, has been of negligible interest. The linear combination of test scores called the difference has for a long time remained somewhat obscure as a test theoretical problem. Generally, it has been little understood and thought to be of less practical significance.

In education increasing efforts are made to adapt instructional programs to individual differences. In the past students were apt to be selected into fixed treatments (programs), whereas a modern philosophy argue for adapting treatments to fit individual aptitudes. The matching of treatments to aptitudes calls for a classification rather than a selection procedure. In this mainstream of educational philosophy the measurement of differential aptitudes has come more and more in the foreground together with renewed efforts to construct adaptive treatments or programs.

The considerable interest of recent years in the phenomenon of aptitude-treatment interaction is certainly also a challenge to the test theoretician to reconsider the psychometric problems connected with the difference score. As references for this emphasis on differential validity, see, for example, Cronbach and Snow (1969), Cronbach (1971), Hills (1971), and Thorndike (1971).

The purpose of the present paper is to consider the reliability problem of difference scores within the framework of generalizability theory. It will be shown and tried to make understandable that the same test theoretical rationale as developed for sum scores is also valid for difference scores. Further, it will be made clear how the reliability formulas for difference scores are dependent upon the particular test designs employed. This is of crucial importance if one intends to generalize to defined families of difference scores.

It should be noted at the outset that the subsequent discussion does not go into the problematic character of gain or change scores, as dealt with by Cronbach and Furby (1970), and Cronbach, Gleser, Nanda, and Rajaratnam (1972). It is here assumed that difference scores between tests are logically sound and should be assessed for their dependability as measures of differential constructs, both bipolar discrete constructs (e.g. verbal/performance), and bipolar continuous constructs (e.g. satisfaction/dissatisfaction).

## Traditional formulation of the reliability of a difference.

In the discussion of a difference score, test theory literature for many years has adhered to a uniform derivation of the

formula for the reliability of such a difference. In its most simplified form this formula reads,

$$r_{1-2} = \frac{r_{11} - r_{12}}{1 - r_{12}} \qquad (1)$$

where $r_{11}$ is the average reliability for tests 1 and 2, and $r_{12}$ the correlation between the two tests. This form is the only one recommended by, for example, Gulliksen (1950), Mosier (1951), Guilford (1954), Horst (1966), Magnusson (1967), McNemar (1969), and Thorndike and Hagen (1969).

Recently, however, the generality of this formulation of the reliability of a difference score has been questioned. Formula (1) "is a considerable simplification of the exact longer formula obtained when one derives the coefficient of reliability for differences from classical measurement theory" (Stanley 1967, 249). "The traditional formula for reliability of a difference score is a special case of the correct formula" (Cronbach and Snow 1968, 20 ). Cronbach and Furby (1970) maintain that the formula has to change with different test designs.

The intricate character of formula (1) is associated with what kind of reliability to choose for $r_{11}$ and what intercorrelation between tests ($r_{12}$) to use when more than one is conceivable. For the time being, there seems to be some confusion concerning the correct formulation of the reliability of difference scores. Until the particular derivations of formulas for specified test designs are shown, this confusion is likely to persist. We shall show that only <u>one</u> particular test design

can match the traditional reliability formula for a difference score.

A prominent feature of this paper will be to approach the general problem of estimating difference score reliability in terms of intraclass correlations. In reformulating this problem by way of an analysis of variance rationale, one can much more easily deal with the different facets that might go into test designs. The flexibility of this approach will become apparent as one proceeds with complex designs where the reliability of various linear combinations, whether a sum, a difference, or a combination of both, can be of substantive interest. In such designs, formula (1) is completely out of date.

The nature of difference scores.

Estimating the generalizability of a test score, whether generated as a sum composite or a difference composite, implies finding how much of the score variance can be regarded as signal and how much as noise.

In the case of an assumedly homogeneous composite it should be clear that the difference between two random test samples going into that composite is per definition a measure of random error. Rulon (1939) saw this property of the difference scores of a homogeneous test and ingeniously utilized it in developing a new formula for the split-half reliability. If two halves of a composite supposedly measure the same construct, then the variance of the difference scores between the two halves can be taken to define the needed error variance.

When Hoyt (1941) applied the analysis of variance technique for the estimation of the internal consistency of a homogeneous composite, he used the person by test (item) interaction as the defined error variance. Later, Gulliksen (1950) and Stanley (1957) showed that the interaction variance used by Hoyt as a definition of error variance, was the average item variance minus the average interitem covariance. For a two-test composite, like the Rulon case, the sophisticated reader will see that this amounts to saying that the difference score variance and the person by half-test interaction are identical definitions. The general finding of Gulliksen and Stanley can be interpreted to mean that for a homogeneous composite the error variance is defined by the average of all possible difference score variances among items.

While the difference scores of a homogeneous composite reflect the noise property, a signal property of a difference score is defined when two tests on a rational basis are conceived to be measuring different constructs. When such scores are subtracted, whatever they might have in common is partialled out, and the residual score is a measure of differential constructs. The variance of difference scores, rationally defined, should be taken to mean that different persons obtain different composite profiles in responding to the two tests in the composite. Within a probabilistic model, one certainly has to define an error term to which the difference score variance should be related in order to assess the reliability of the differences.

The dependability problem involved in dealing with the difference score implies finding to what extent the profiles obtained by persons are consistent over comparable difference scores.

Defining a family of difference scores.

According to generalizability theory, particularly, one has to define a domain of tests in order to be able to determine to what universe of measures he wants to generalize. This seems to be evidently clear as far as a sum score is concerned.

But the same rationale should of course also be valid for a difference score. In order to estimate difference score reliability, one has to be quite explicit of what can be accepted as comparable, admissible, (nominally) parallel, or (nominally) equivalent difference scores, suitable for the particular testing problem at issue. As Guttman (1953) remarked, defining parallel or comparable measures is to a considerable extent a matter of choice.

When concerned with a sum score, one identifying aspect or facet of the test samples (items, forms) has to be defined. In deciding on a family of measures for the homogeneous composite, one is involved in a one-facet test design, having at least, say, two forms or two occasions. It appears to the author that one can not, as Cronbach and Furby (1970) seem to maintain, avoid the complications of multifacet theory in discussing difference score reliability, even in the most simple test design. It takes one facet to define one difference score, and another facet to define the family of difference scores to which one wants to generalize.

If a score is defined as the difference between two tests, the next decision to be made is to define one or more facets over which to generalize. For example, one may want to genera-

lize over forms, or over occasions, or over confounded forms and occasions. In FIGURE 1 these options for defining families of difference scores are illustrated. Let X be a test score, the

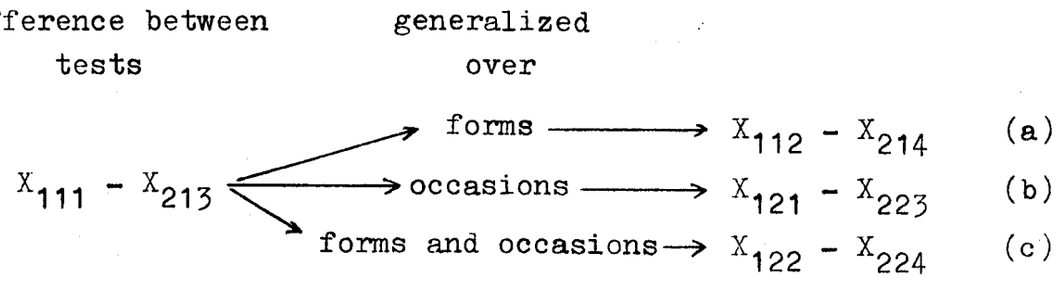



FIGURE 1. Families of difference scores

first subscript denoting test, the second occasion, and the third form. By having defined three families of difference scores over which to generalize, three test designs are simultaneously specified, a, b, and c.

When a particular family of difference scores is chosen as the one of substantive interest, the reliability problem involves estimating the consistency of the defined comparable measures.

To keep the formulations within reasonable bounds, the subsequent discussion will be restricted to designs with 2 fixed tests, two fixed occasions, and 2 or k random forms.

The interclass correlation approach to the reliability of a homogeneous difference score composite.

The correlation between two comparable measures is generally accepted as an estimate of the reliability of one of the compa-

rable measures. If two comparable difference scores are given, then their intercorrelation should be taken to be the reliability of one difference score (Stanley 1967).

As a first step in approaching the general problem of estimating difference score reliability, the rationale of an interclass correlation will be applied to show the derivation of difference score reliability formulas for the simplest case possible, only two difference scores given.

Assume that two domains of tests are defined, each domain being thought to measure different constructs. Further, assume that two forms are picked within each of the domains, such that two difference scores are available, $D_1 = X_{11} - X_{23}$ and $D_2 = X_{12} - X_{24}$. The first subscript denotes test, the second form. Thus, form 1 and 2 are comparable measures within test 1; form 3 and 4 comparable measures within test 2. A family of difference scores is now defined, being a case of design a in FIGURE 1. This means that one is interested in generalizing over forms of difference scores.

How the correlation between the two difference scores will come out, can most clearly be seen from a correlation matrix where all four test forms are considered one linear combination with both signs used, plus for the two forms of the first test and minus for the two forms of the second test, as shown in TABLE 1. Two categories of correlation coefficients should be kept separate in TABLE 1. These are the correlations between forms within one of the two defined domains and the correlations between forms between the two domains. The two categories of coefficients will be called the within test between forms cor-

relation, $r_{wb}$, and the between tests between forms correlation, $r_{bb}$. It should be noted that the terminology adopted here, is parallel to the notion of a correlation between linked and independent observations, as adopted by Cronbach and Furby (1970). The correlation within test between forms is based on two linked observations, while the correlation between tests between forms is based on independent observations. It should be obviously

TABLE 1. The correlation between difference scores

| | | $D_1$ | | $D_2$ | |
|---|---|---|---|---|---|
| | | $X_{11}$ | $-X_{23}$ | $+X_{12}$ | $-X_{24}$ |
| $D_1$ | $X_{11}$ | 1 | $-r_{bb}$ | $+r_{wb}$ | $-r_{bb}$ |
| | $-X_{23}$ | $-r_{bb}$ | 1 | $-r_{bb}$ | $+r_{wb}$ |
| $D_2$ | $+X_{12}$ | $+r_{wb}$ | $-r_{bb}$ | 1 | $-r_{bb}$ |
| | $-X_{24}$ | $-r_{bb}$ | $+r_{wb}$ | $-r_{bb}$ | 1 |

clear that the $r_{wb}$ coefficients can be expected to be considerably higher than the $r_{bb}$ coefficients if differential construct validity is indicated for the two domains.

The correlation between the two difference scores in TABLE 1 can easily be found by taking the ratio of the covariance between the difference scores to the product of the standard deviations of the two scores,

$$r_{D_1/D_2} = \frac{Cov_{D_1/D_2}}{(V_{D_1})^{\frac{1}{2}} (V_{D_2})^{\frac{1}{2}}} = \frac{\Sigma r_{wb} - \Sigma r_{bb}}{(2 - \Sigma r_{bb})^{\frac{1}{2}} (2 - \Sigma r_{bb})^{\frac{1}{2}}} \qquad (2)$$

If one makes the assumption that the correlation coefficients within tests between forms can be expected to be equal, and likewise for the correlation coefficients between tests between forms, formula (2) simplifies to

$$r_{D_1/D_2} = \frac{2\bar{r}_{wb} - 2\bar{r}_{bb}}{2 - 2\bar{r}_{bb}} = \frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bb}} \qquad (3)$$

Formula (3) is the traditional form of the difference score reliability. Syntactically it is equal to formula (1). Semantically, formula (3) is associated with a particular test design and has a clear meaning, while formula (1) as a general formulation of difference score reliability is unequivocal. As will become apparent as we proceed, this nested design, here forms nested within tests, is the only test design for which the traditional formula for the difference score reliability is valid.

Next, a completely ignored feature of the reliability of difference scores will be approached. In keeping with traditional test theory, one may want to ask what the reliability of the two difference scores combined in a sum will be. This amounts to being concerned with the reliability of the linear combination $X_{11} - X_{23} + X_{12} - X_{24}$. Intuitively, it seems reasonable to adopt the Spearman-Brown prophecy formula for this problem: In doubling the single difference score measure, what will the reliability be? By applying the simple Spearman-Brown formula for double length, one can derive formula (4), using formula (3) as the point of departure. As far as the author knows, formula (4) has never appeared in the test theory literature before. Conceptually, the formula is doubtless sound, and it certainly should

$$r_{D(2)} = \frac{2\left(\frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bb}}\right)}{1 + \left(\frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bb}}\right)} = \frac{2\bar{r}_{wb} - 2\bar{r}_{bb}}{1 + \bar{r}_{wb} - 2\bar{r}_{bb}} \qquad (4)$$

prove to be an extremely useful formula. If two comparable difference scores are available, one should not use the two scores for estimating the reliability for one of them, as Stanley (1967) recommends. Rather, he should estimate the reliability for the sum of the two difference scores, like what is done in (4), and use that linear combination as a measure in a practical testing situation, and not only one of the difference scores available.

By having elaborated the rationale for the difference score reliability in dealing with one and two difference scores, one should be ready to consider the more general problem of approaching the internal consistency of a defined family of difference scores. This, we think, will be an exact parallel to the derivation of the Hoyt-Cronbach alpha coefficient. The general form of coefficient alpha for $\underline{k}$ comparable difference scores can be defined

$$\text{alpha}_{D(k)} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma V_{d_i}}{V_D}\right) \qquad (5)$$

where k is the number of difference scores, or number of forms within each of the two tests, $\Sigma V_{d_i}$ the sum of the k difference score variances, and $V_D$ the variance of the sum of the k difference scores. Thus, (5) is in form equal to traditional coeffi-

cient alpha for the case where k test scores are defined as k difference scores.

From the correlation matrix of the sum of two difference scores, TABLE 1, it can be seen that one difference score variance has the form $2 - \Sigma r_{bb}$, where 2 is the sum of two standard score variances. By averaging the correlation coefficients between tests between forms, the sum of the k difference score variances can be written $k(2-2\bar{r}_{bb})$.

From the correlation matrix, TABLE 1, can also be observed that the covariance between difference scores has the form $\Sigma r_{wb} - \Sigma r_{bb}$. By averaging the correlation coefficients, the form will be $2\bar{r}_{wb} - 2\bar{r}_{bb}$. While there are k difference score variances, there are k(k-1) difference score covariances, such that the variance of the sum of k difference scores can be written $k(2-2\bar{r}_{bb}) + k(k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})$. Consequently, formula (5) will read in terms of the properties of the correlation matrix of the k difference scores,

$$\text{alpha}_{D(k)} = \left(\frac{k}{k-1}\right)\left(1 - \frac{k(2-2\bar{r}_{bb})}{k(2-2\bar{r}_{bb}) + k(k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})}\right)$$

$$= \left(\frac{k}{k-1}\right)\left(\frac{k(2-2\bar{r}_{bb})+k(k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})-k(2-2\bar{r}_{bb})}{k(2-2\bar{r}_{bb}) + k(k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})}\right)$$

$$= \frac{k(2\bar{r}_{wb}-2\bar{r}_{bb})}{2-2\bar{r}_{bb} + (k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})}$$

$$= \frac{k\bar{r}_{wb}-k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{wb} - k\bar{r}_{bb}} \qquad (6)$$

Formula (6) is here considered a true counterpart to coefficient alpha as traditionally conceived. It is the expected correlation of the sum of k comparable difference scores with another set of k comparable difference scores.

The form of coefficient alpha developed in formula (6) is the general Spearman-Brown prophecy formula adopted for difference scores. As far as the author knows, formula (6) is also new in test theory literature. (A similar reasoning seems to be the basis for Bereiter (1963), and Webster & Bereiter (1963) in conceiving of composite difference scores, i.e. differences as sums of change items, and a stepped-up reliability. It can be seen that formula (6) is related in form to the traditional Spearman-Brown prophecy formula: In (6) the numerator and the denominator in traditional Spearman-Brown is reduced by the common variance between the two tests (domains).

It should be noted that formula (6) could well be derived by directly applying the general Spearman-Brown to the correlation between two difference scores, or the reliability of one difference score, formula (3). But that would be a more mechanical derivation. The point of departure for developing formula (6) is believed to be more meaningful.

Both formula (3) and formula (6) pay attention to a phenomenon which has been totally ignored in dealing with difference scores (except for the two references above ): The possibility of increasing the reliability of a difference score by adding more comparable observations to the measure. This point will be emphasized throughout in the subsequent discussion.

Deriving the reliability of difference scores by way of a variance components analysis as a general approach.

It has just been shown how the rationale for the consistency of difference scores between tests over nested forms within tests can be explicated in terms of an interclass correlation approach. In the following, additional test designs of greater complexity will be presented where several difference scores can be defined and also different families of difference scores over which one may want to generalize.

The multifacet character of difference score reliability even in its most simple form makes the interclass correlation approach less suitable than an analysis of variance approach in terms of variance components. As some of the test designs to be dealt with will include more than two facets, it seems desirable to establish a more general approach. Undoubtedly, a variance components analysis is such an approach in which it will be possible to derive the correct formulas for the reliability of variously defined difference scores over different kinds of universes of generalization by taking into account the particular test designs used.

Altogether 7 test designs will be presented and analyzed by a variance components approach, emphasizing the assessment of the differential/construct validity of tests over forms and/or occasions.

One feature of the present formulation of the reliability of difference scores should be noted at the outset. The analysis of variance will be performed on standardized scores. This is done, first, in order that the sets of scores should be in comparable units, else the difference scores will have no meaning. Another

reason for choosing the standard score as the comparable unit is that the derivations of formulas via variance components analysis can be expressed in terms of the covariances of standardized scores, which are correlation coefficients. Thus the formulations to be developed in the following can be compared to previous formulations of difference score reliability, which has always been in terms of interclass correlations.

Design 1.

In analysis of variance terms, what is here called test design 1 is an n-persons-by-two-tests-by-k-forms-within-tests design, where persons and forms are random factors and tests a fixed factor. This is the same test design that underlies formula (6).

First the minimum design for finding the reliability of the difference between two tests will be presented,i.e. only two forms within each of the tests, in order to keep matters as

TABLE 2. ANOVA of test design 1.

| Sources | SS | df | MS | E(MS) |
|---|---|---|---|---|
| Persons | $SS_p$ | n-1 | $MS_p$ | $\sigma^2_{pf:t} + 2\sigma^2_{pt} + 4\sigma^2_p$ |
| Tests | 0 | 1 | 0 | |
| Forms:T | 0 | 2 | 0 | |
| P x T | $SS_{pt}$ | n-1 | $MS_{pt}$ | $\sigma^2_{pf:t} + 2\sigma^2_{pt}$ |
| P x F:T | $SS_{pf:t}$ | 2(n-1) | $MS_{pf:t}$ | $\sigma^2_{pf:t}$ |

simple as possible by way of introduction. The analysis of variance table (ANOVA table) together with the expected mean squares

(E(MS)) is shown in TABLE 2. The notation f:t means forms nested within tests. This is in accordance with the notation recommended by Millman and Glass (1967) and Cronbach,Gleser,Nanda, and Rajaratnam (1972). It should be noted that two/of the sources in TABLE 2, tests and forms, come out with zero sums of squares because the data matrix has been columncentered by the standardizing procedure.

While $MS_p$ is of crucial interest when the sum score is at issue, it is the $MS_{pt}$ that attracts the attention in the present context. The person by test interaction reflects the variance attributed to differential aptitudes on the two tests. Conceptually, the same variance can be obtained by taking the variance of the difference scores between the two tests across the two forms within each of the tests.

The components model for the person by test interaction, $E(MS_{pt})$, shows the conceptual separation of what is defined as true difference score variance/error of measurement variance. As can be seen, it is the person by form interaction within the tests that serves the function of defining the error term associated with the observed difference scores. This interaction is a measure of inconsistency of responses over forms within the tests. As such,this interaction appears to be a logically sound error of measurement variance.

In defining/reliability of difference scores, the ratio of true score variance to observed score variance is still the reasonable formulation to make. By way of the $E(MS_{pt})$, two reliability coefficients can be defined, one for the sum of the two difference scores, another for the average difference score,

denoted $r_{D(2)}$ and $r_{D(1)}$, respectively.

$$r_{D(2)} = \frac{2\sigma^2_{pt}}{\sigma^2_{pf:t} + 2\sigma^2_{pt}} = \frac{MS_{pt} - MS_{pf:t}}{MS_{pt}} \tag{7}$$

$$r_{D(1)} = \frac{\sigma^2_{pt}}{\sigma^2_{pf:t} + \sigma^2_{pt}} = \frac{MS_{pt} - MS_{pf:t}}{MS_{pt} + MS_{pf:t}} \tag{8}$$

Under formulas (7) and (8) both defining and computing forms are given.

Formulas (7) and (8) seem indeed a far cry from formulas (4) and (3). Actually, formula (3) is identical with formula (8); and formula (4) with formula (7). However, the two sets of formulas are expressed in languages that are apparently quite different.

As a matter of fact, the complete convergence of the two languages is fairly easy to show. The sources of variance associated with persons in test design 1, TABLE 2, are all linear combinations of the four forms going into the design. According to multivariate statistics one can construct four orthogonal linear combinations of the four forms that will exhaust the total variance of the four forms. These four linear combinations are given by the particular test design used.

In the present design there is one linear combination which is the sum of the four forms, one that is a difference between the two tests over the two forms within each of them, and two linear combinations, one within each of the two tests, that are the differences between the forms. The variances of the four linear combinations will be, using the notation employed in

TABLE 1 (The first subscript denotes test, the second form.):

$$V_p = (\frac{1}{n-1})\Sigma(z_{11} + z_{12} + z_{23} + z_{24})^2 \quad (9)$$

$$V_{pt} = (\frac{1}{n-1})\Sigma(z_{11} + z_{12} - z_{23} - z_{24})^2 \quad (10)$$

$$V_{pf:t} = (\frac{1}{n-1})\Sigma(z_{11} - z_{12})^2 + (\frac{1}{n-1})\Sigma(z_{23} - z_{24})^2 \quad (11)$$

By expanding formulas (9), (10), and (11), four variances and twelve covariances, which are correlation coefficients, are obtained. Of the twelve covariances, two categories can be distinguished and will be kept separate: The covariances between forms within tests, $r_{wb}$, and the covariances between forms between tests, $r_{bb}$. There are four covariances of the first category and eight of the second. By reassembling, summing, and averaging the variances and the covariances of the two categories, the variances of the linear combinations in (9), (10), and (11) can be written,

$$V_p = 4 + 4\bar{r}_{wb} + 8\bar{r}_{bb} \quad (12)$$

$$V_{pt} = 4 + 4\bar{r}_{wb} - 8\bar{r}_{bb} \quad (13)$$

$$V_{pf:t} = 2(2 - 2\bar{r}_{wb}) \quad (14)$$

A functional relationship between the variances of the actual linear combinations as developed in (12), (13), and (14) and the MS's obtained in an analysis of variance approach should be observed. The variances obtained in (12) and (13) are larger than the $MS_p$ and the $MS_{pt}$ in TABLE 2 by a factor of 4, which is the number of forms going into the linear combinations. The variance

obtained in (14) is larger than the $MS_{pf:t}$ in TABLE 2 by a factor of 2, which is the number of forms going into each of the two pooled linear combinations. These are all consequences of different conventions in defining the variance of linear combinations in a psychometric and an analysis of variance tradition.

According to the way of expressing the variances of the linear combinations in terms of the correlation matrices, as done in (12), (13), and (14), and in observing the relationship between those variances and the MS's of TABLE 2, a modified ANOVA table of test design 1 is given in TABLE 3, with components derived as functions of average correlation coefficients. Only the MS's for the three sources associated with individual differences are presented. TABLE 3 is interesting in showing the convergence of an

TABLE 3. ANOVA of standardized scores of design 1.

| | E(MS) | Obs(MS) | Variance components |
|---|---|---|---|
| $MS_p$ | $= \sigma^2_{pf:t} + 2\sigma^2_{pt} + 4\sigma^2_p$ | $= 1 + \bar{r}_{wb} + 2\bar{r}_{bb}$ | $\bar{r}_{bb}$ |
| $MS_{pt}$ | $= \sigma^2_{pf:t} + 2\sigma^2_{pt}$ | $= 1 + \bar{r}_{wb} - 2\bar{r}_{bb}$ | $\bar{r}_{wb} - \bar{r}_{bb}$ |
| $MS_{pf:t}$ | $= \sigma^2_{pf:t}$ | $= 1 - \bar{r}_{wb}$ | $1 - \bar{r}_{wb}$ |

analysis of variance of a repeated measures design with functions of the correlation matrices of the linear combinations of those repeated measures.

In TABLE 3 the information needed to translate the defined difference score reliabilities of formulas (7) and (8) into formulas in terms of observed properties of the correlation matrices of the linear combinations is provided.

$$r_{D(2)} = \frac{2\sigma^2_{pt}}{\sigma^2_{pf:t} + 2\sigma^2_{pt}} = \frac{MS_{pt} - MS_{pf:t}}{MS_{pt}} = \frac{2\bar{r}_{wb} - 2\bar{r}_{bb}}{1 + \bar{r}_{wb} - 2\bar{r}_{bb}} \qquad (4)$$

$$r_{D(1)} = \frac{\sigma^2_{pt}}{\sigma^2_{pf:t} + \sigma^2_{pt}} = \frac{MS_{pt} - MS_{pf:t}}{MS_{pt} + MS_{pf:t}} = \frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bb}} \qquad (3)$$

The derivation of the reliability of the sum of two difference scores and the reliability of one average difference score by an analysis of variance approach ends up with just the same formulas as derived by the more traditional interclass correlation approach (see pages 10-11). What is of considerable interest to learn from TABLE 3 is that the variance components can be written as functions of the correlation coefficients. As a matter of fact, what is called variance components in the terms of analysis of variance can sometimes more appropriately be called covariance components (Stanley 1961, Eikeland 1970, Cronbach, Gleser, Nanda, and Rajaratnam 1972, and Eikeland 1972).

TABLE 4. ANOVA of standardized scores of test design 1. (n x 2 x k)

| | E(MS) | Obs(MS) | Variance components |
|---|---|---|---|
| $MS_p$ | $= \sigma^2_{pf:t} + k\sigma^2_{pt} + 2k\sigma^2_p$ | $= 1 + (k-1)\bar{r}_{wb} + k\bar{r}_{bb}$ | $\bar{r}_{bb}$ |
| $MS_{pt}$ | $= \sigma^2_{pf:t} + k\sigma^2_{pt}$ | $= 1 + (k-1)\bar{r}_{wb} - k\bar{r}_{bb}$ | $\bar{r}_{wb} - \bar{r}_{bb}$ |
| $MS_{pf:t}$ | $= \sigma^2_{pf:t}$ | $= 1 - \bar{r}_{wb}$ | $1 - \bar{r}_{wb}$ |

The more general formulation of the reliability for test design 1, with k forms within each of the two tests, can readily be worked out in terms of variance components expressed as

functions of the 2k x 2k correlation matrix. This expansion is shown in TABLE 4.

In the n-persons-by-two-tests-by-k-forms-within-tests design, a set of k random differences between two forms, one from each of the tests, can be formed. The reliability of the sum of these difference scores can be established by taking the ratio of universe score variance, which is $k\bar{r}_{wb} - k\bar{r}_{bb}$, to the observed sum of difference score variance, which is $1 + (k-1)\bar{r}_{wb} - k\bar{r}_{bb}$.

$$\text{alpha}_{D(k)} = \frac{k\bar{r}_{wb} - k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{wb} - k\bar{r}_{bb}} \qquad (6)$$

By the variance components analysis formula (6) is rederived as the alpha coefficient for the sum of k random difference scores. (For the previous derivation of formula (6), see page 12.)

Design 2.

Consider next another test design with the same two facets, tests and forms, as in design 1. What is different from design 1 is that forms are thought to be crossed with tests. In an n-persons-by-two-tests-by-two-forms test design the same formats can be used under both tests. For example, one may be interested in the difference score between two concepts measured by the same two scales in a semantic differential approach. Let the two concepts be named tests and the two scales forms. In this particular design, the four orthogonal linear combinations that are established by the design matrix are somewhat different from the linear combinations established for design 1. The variances

of the four linear combinations will be,

$$V_p = (\frac{1}{n-1})\Sigma(z_{11} + z_{12} + z_{21} + z_{22})^2 \quad (15)$$

$$V_{pt} = (\frac{1}{n-1})\Sigma(z_{11} + z_{12} - z_{21} - z_{22})^2 \quad (16)$$

$$V_{pf} = (\frac{1}{n-1})\Sigma(z_{11} - z_{12} + z_{21} - z_{22})^2 \quad (17)$$

$$V_{ptf} = (\frac{1}{n-1})\Sigma(z_{11} - z_{12} - z_{21} + z_{22})^2 \quad (18)$$

where $V_p$ is the variance of the sum score across all of the four observations, $V_{pt}$ the variance of the difference score between tests across forms, $V_{pf}$ the variance of the difference score between forms across tests, and $V_{ptf}$ is the variance of a difference between two differences score; i.e., the difference between the two differences between forms for each of the two tests. The two subscripts for the standard scores denote tests and forms, respectively.

TABLE 5. ANOVA of standardized scores of test design 2.

| | E(MS) | Obs(MS) | Variance components |
|---|---|---|---|
| $MS_p$ | $= \sigma^2_{ptf}+2\sigma^2_{pf}+2\sigma^2_{pt}+4\sigma^2_p$ | $= 1 + \bar{r}_{wb}+ \bar{r}_{bw}+ \bar{r}_{bb}$ | $\bar{r}_{bb}$ |
| $MS_{pt}$ | $= \sigma^2_{ptf}+2\sigma^2_{pt}$ | $= 1 + \bar{r}_{wb}- \bar{r}_{bw}- \bar{r}_{bb}$ | $\bar{r}_{wb}-\bar{r}_{bb}$ |
| $MS_{pf}$ | $= \sigma^2_{ptf}+2\sigma^2_{pf}$ | $= 1 - \bar{r}_{wb}+ \bar{r}_{bw}- \bar{r}_{bb}$ | $\bar{r}_{bw}-\bar{r}_{bb}$ |
| $MS_{ptf}$ | $= \sigma^2_{ptf}$ | $= 1 - \bar{r}_{wb}- \bar{r}_{bw}+ \bar{r}_{bb}$ | $1-\bar{r}_{wb}-\bar{r}_{bw}+\bar{r}_{bb}$ |

By expanding formulas (15)-(18), reassembling the variances and three categories of correlation coefficients, and averaging, the Obs(MS) column of TABLE 5 is obtained by dividing each of

the variances of the linear combinations in (15)-(18) by 4, the number of observations going into each of the combinations. In passing it should be noted that the sum of the Obs(MS) column of TABLE 5 adds to 4, which is the total variance of the four standardized variables, the trace.

The three categories of correlation coefficients represented in TABLE 5 are a within test between forms correlation, $r_{wb}$; a between tests within form correlation, $r_{bw}$; and a between tests between forms correlation, $r_{bb}$.

Again it is the person by test interaction that is of interest in assessing the reliability of the difference score, i.e. the $MS_{pt}$ in TABLE 5, the observed difference score variance. From the $E(MS_{pt})$ can be seen what is considered universe score variance and what error. The two alpha coefficients for the difference score between tests will be,

$$\text{alpha}_{pt(2)} = \frac{2\sigma^2_{pt}}{\sigma^2_{ptf} + 2\sigma^2_{pt}} = \frac{MS_{pt} - MS_{ptf}}{MS_{pt}} = \frac{2\bar{r}_{wb} - 2\bar{r}_{bb}}{1 + \bar{r}_{wb} - \bar{r}_{bw} - \bar{r}_{bb}} \qquad (19)$$

$$\text{alpha}_{pt(1)} = \frac{\sigma^2_{pt}}{\sigma^2_{ptf} + \sigma^2_{pt}} = \frac{MS_{pt} - MS_{ptf}}{MS_{pt} + MS_{ptf}} = \frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bw}} \qquad (20)$$

In formulas (19) and (20) the reliabilities are given as defining formulas in terms of variance components; one set of computing formulas in terms of MS's, another in terms of correlation coefficients. It is indeed difficult on an intuitive basis to see why the formulas in terms of correlation coefficients should come out as they do. The subtle difference between formula (3) and formula (20) should be noted. This is the same distinction as made by Cronbach and Furby (1970, p.71), their formulas (6) and (7).

The general case of test design 2 will be a design with n persons, 2 tests, and k forms crossed with tests. The reliability of the difference score between the two tests across the k forms is of interest. In deriving the formula for the reliability of this particular difference score, several approaches could be undertaken. The most convenient approach is certainly by way of the $E(MS_{pt})$ in an ANOVA table for this general test design, which will give,

$$\text{alpha}_{pt(k)} = \frac{k\sigma^2_{pt}}{\sigma^2_{ptf} + k\sigma^2_{pt}} = \frac{MS_{pt} - MS_{ptf}}{MS_{pt}} \qquad (21)$$

It may also be interesting to see what the general formula will be like in terms of correlation coefficients. One could elaborate the Obs(MS) column in TABLE 5 for this purpose. More easily, this formula can be derived by applying the general Spearman-Brown prophecy formula to formula (20). By this procedure, the result is,

$$\text{alpha}_{pt(k)} = \frac{k\bar{r}_{wb} - k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{wb} - \bar{r}_{bw} - (k-1)\bar{r}_{bb}} \qquad (22)$$

There is a slight change from formula (6), which is the general case of test design 1, to formula (22), the general case of test design 2. What these changes in formulas will be from test design to test design seem not to be foreseeable on a common sense basis. A strict adherence to rules of thumb for writing out the variance components model for the particular test designs used will be a good advice in order to be able to end up with the correct reliability formulas.

Design 3.

For the first time occasions will be included in one of our /test designs. We are still interested in the differential/validity of the two test scores, but now the intention is to generalize over two fixed occasions.

Design 3 is a confounded test design in that only one form is used for each test on both occasions. Thus a test-form unit is established, making the operational definition of a test totally dependent on the one particular form chosen for each of the two tests.

Syntactically, test design 3 is identical to design 2 when k = 2. Semantically, however, they are quite different as design 2 generalizes over forms while design 3 generalizes over occasions.

Let $X_{111}$, $X_{121}$, $X_{212}$, $X_{222}$ be the four scores going into test design 3 with first subscript denoting test, second occasion, and third form. The intention is to estimate the reliability of the linear combination of the two difference scores between tests, $(X_{111} - X_{212}) + (X_{121} - X_{222})$ and also the reliability of one average difference score between tests. While forms in test design 2 are crossed with tests, in design 3 occasions are crossed with tests. Just the same three categories of correlation coefficients as specified for test design 2, $r_{wb}$, $r_{bw}$, and $r_{bb}$, can also be identified in the present design, but the meaning will be different. In design 3, $r_{wb}$ means the correlation within test between occasions, $r_{bw}$ the correlation between tests within occasion, and $r_{bb}$ the correlation between tests between occasions.

In deriving the reliability formulas for the present design, TABLE 5 is applicable, remembering that the PF interaction is

replaced by a PO, a person by occasion, interaction. Thus formulas (19) and (20) will also be correct for the reliabilities wanted for test design 3, with a slight change in the subscript for the component and the MS for the triple interaction. The correct formulas will read,

$$\text{alpha}_{pt(2)} = \frac{2\sigma^2_{pt}}{\sigma^2_{pto} + 2\sigma^2_{pt}} = \frac{MS_{pt} - MS_{pto}}{MS_{pt}} = \frac{2\bar{r}_{wb} - 2\bar{r}_{bb}}{1 + \bar{r}_{wb} - \bar{r}_{bw} - \bar{r}_{bb}} \qquad (23)$$

$$\text{alpha}_{pt(1)} = \frac{\sigma^2_{pt}}{\sigma^2_{pto} + \sigma^2_{pt}} = \frac{MS_{pt} - MS_{pto}}{MS_{pt} + MS_{pto}} = \frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bw}} \qquad (24)$$

An extremely interesting change in the syntactical feature of reliability formulas should be noted in degressing for a short while to the reliability of the change score, i.e. the difference between occasions score. From TABLE 5 it is possible to derive the two reliability coefficients for the difference between occasions. The formulas will be,

$$\text{alpha}_{po(2)} = \frac{2\bar{r}_{bw} - 2\bar{r}_{bb}}{1 - \bar{r}_{wb} + \bar{r}_{bw} - \bar{r}_{bb}} \qquad (25)$$

$$\text{alpha}_{po(1)} = \frac{\bar{r}_{bw} - \bar{r}_{bb}}{1 - \bar{r}_{wb}} \qquad (26)$$

As can be seen, the two categories of correlations, $r_{wb}$ and $r_{bw}$, has changed roles from the set of coefficients for test difference, (23) and (24), to the new set for occasion difference, (25) and (26). Certainly, it is possible to figure out on a logical basis that the change has to be made exactly this way, but it is not immediately apparent.

Design 4.

In design 4, two comparable forms for each test will be included, in addition to two tests and two occasions. However, the forms are going to be confounded with occasions, such that occasion-form units are established. Thus the separate effects of occasion and form can not be distinguished in the design. Let $X_{111}$, $X_{122}$, $X_{213}$, $X_{224}$ be the four scores going into test design 4 with first subscript denoting test, second occasion, and third form. The intention is to estimate the reliability of the linear combination of the two difference scores between tests, $(X_{111} - X_{213}) + (X_{122} - X_{224})$ and also the reliability of one average difference score between tests.

This particular design is a prominent one in the literature, as it is the one test design (among many possible others) used by Stanley (1967) and Stanley (1971) in discussing the problem of difference score reliability.

It should be more or less obvious that design 4 is syntactically identical to design 3, as two tests are crossed with two occasions. Therefore, no new formulas can be developed for this test design. Formulas (23) and (24) are valid for the difference between tests, and formulas (25) and (26) for the difference between occasions, if that particular difference should be of concern. Semantically, however, there is a slight but significant discrepancy, attributable to the different kinds of confounded effects in design 3 and design 4.

Design 5.

A much stronger test design than the two preceding ones can be generated by taking new samples of comparable forms for the

tests for each test-occasion combination. This design will include tests, occasions that are crossed with tests, and forms nested within each test-occasion combination. How can the reliability for the difference score between tests generalized over occasions and forms be worked out?

Consider a 2-tests-by-2-occasions-by-k-nested-forms-within-test-occasion-combinations design. For this three-facet test design, it is obviously clear how much can be gained by applying an analysis of variance approach. Actually, an approach to the reliability of the difference score between tests by way of interclass correlations would be extremely difficult, although not impossible.

In writing out the ANOVA table for design 5, only the structural models for those sources that involve individual differences will be specified. For the present design this means that tests, occasions, test by occasion interaction, and nested forms will be ignored. As remembered, these sources will have zero sums of squares in a columncentered matrix.

TABLE 6. Variance components model for standardized scores of test design 5

| | |
|---|---|
| P | $\sigma^2_{pf:to} + k\sigma^2_{pto} + 2k\sigma^2_{po} + 2k\sigma^2_{pt} + 4k\sigma^2_{p}$ |
| PT | $\sigma^2_{pf:to} + k\sigma^2_{pto} + 2k\sigma^2_{pt}$ |
| PO | $\sigma^2_{pf:to} + k\sigma^2_{pto} + 2k\sigma^2_{po}$ |
| PTO | $\sigma^2_{pf:to} + k\sigma^2_{pto}$ |
| PF:TO | $\sigma^2_{pf:to}$ |

In the present context, it is the structural model for PT, variance of the difference score between tests, that is of particular inter-

est. Notice that for the first time the variance components model for the difference score between tests has three terms. The new term is the weighted component for the person by test by occasion interaction, which can be interpreted to mean the inconsistency of the difference score between tests for the two occasions, or the stability of the difference score. The problem with $k\sigma^2_{pto}$ is whether it should go to error or to universe score variance. The solution is dependent upon how occasion is defined, whether as a random or a fixed factor. As there can be no meaning in generalizing to a universe of occasion, this facet has to be considered fixed,i.e. the intention is to generalize to just those two occasions chosen for the test design. Therefore, $k\sigma^2_{pto}$ will be a systematic source of variance in the observed difference score and is allocated to universe score variance. (For a discussion of this kind of problems, see Rabinowitz and Eikeland (1964), and Eikeland (1972).) Thus,as a defining formula for the reliability of the difference score for tests in design 5, the following should be the correct ones,

$$\text{alpha}_{pt(k)} = \frac{k\sigma^2_{pto} + 2k\sigma^2_{pt}}{\sigma^2_{pf:to} + k\sigma^2_{pto} + 2k\sigma^2_{pt}} \tag{27}$$

$$\text{alpha}_{pt(1)} = \frac{\sigma^2_{pto} + \sigma^2_{pt}}{\sigma^2_{pf:to} + \sigma^2_{pto} + \sigma^2_{pt}} \tag{28}$$

In terms of obtained MS's, i.e. as computing formulas, (27) and (28) should be, using TABLE 6,

$$\text{alpha}_{pt(k)} = \frac{MS_{pt} - MS_{pf:to}}{MS_{pt}} \tag{29}$$

$$\text{alpha}_{pt(1)} = \frac{MS_{pt} + MS_{pto} - 2MS_{pf:to}}{MS_{pt} + MS_{pto} + 2(k-1)MS_{pf:to}} \qquad (30)$$

In practical testing, formulas (29) and (30) are the convenient formulas to use. More as a curiosity, it might be of interest to see how formulas (27) and (28) will come out as a function of the correlation matrix of test design 5.

TABLE 7. Obs(MS) for test design 5
as a function of the correlation matrix.

| | Components |
|---|---|
| $MS_p = 1 + (k-1)\bar{r}_{ww} + k\bar{r}_{wb} + k\bar{r}_{bw} + k\bar{r}_{bb}$ | $\bar{r}_{bb}$ |
| $MS_{pt} = 1 + (k-1)\bar{r}_{ww} + k\bar{r}_{wb} - k\bar{r}_{bw} - k\bar{r}_{bb}$ | $\bar{r}_{wb} - \bar{r}_{bb}$ |
| $MS_{po} = 1 + (k-1)\bar{r}_{ww} - k\bar{r}_{wb} + k\bar{r}_{bw} - k\bar{r}_{bb}$ | $\bar{r}_{bw} - \bar{r}_{bb}$ |
| $MS_{pto} = 1 + (k-1)\bar{r}_{ww} - k\bar{r}_{wb} - k\bar{r}_{bw} + k\bar{r}_{bb}$ | $\bar{r}_{ww} - \bar{r}_{wb} - \bar{r}_{bw} + \bar{r}_{bb}$ |
| $MS_{pf:to} = 1 - \bar{r}_{ww}$ | $1 - \bar{r}_{ww}$ |

In TABLE 7, as there are three facets in design 5, a third subscript is understated, the subscript for form. The understatement is that all correlations are between forms. Else, the first subscript denotes test, the second occasion. As a check on the correctness of the derivation of variance components, it should be remembered that the sum of the unweighted components in the components column in TABLE 7 must add to 1, which is the variances in the principal diagonal of the correlation matrix.

Using TABLE 6 and TABLE 7, the reliabilities for the difference score can be worked out in terms of average correlation coefficients,

$$\text{alpha}_{pt(k)} = \frac{k\bar{r}_{ww} + k\bar{r}_{wb} - k\bar{r}_{bw} - k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{ww} + k\bar{r}_{wb} - k\bar{r}_{bw} - k\bar{r}_{bb}} \qquad (31)$$

$$\text{alpha}_{pt(1)} = \frac{\bar{r}_{ww} - \bar{r}_{bw}}{1 - \bar{r}_{bw}} \qquad (32)$$

Formula (31) undoubtedly bears a certain similarity to the general Spearman-Brown prophecy formula, but has become much more complicated. It should be compared to formula (6) and formula (22).

Formula (32) is in form equal to formula (3); however, the choice of correlation coefficients should be noted. It should also be compared to formula (20) and formula (26). The comparisons show how dependent the formulas are on test design and what kind of difference score is being examined.

Design 6.

The next test design to be discussed is thought to be a realistic one in that much test data should exist that match this design. It would be like taking test-retest for a battery consisting of tests with forms nested within them. Actually, this should be the proper test design for Irwin (1966) in his effort to assess the reliability of difference scores in WISC. Here design 6 will be presented as a 2-tests-by-two-crossed occasions-by-k-forms-nested-within-tests design. (The change from design 5 to design 6 should be noted: In design 5 forms are nested within TO, in design 6 they are nested within T.)

The variance components model for the standardized scores of test design 6 is presented in TABLE 8. It looks formidable, yet it is believed to be meaningful. Only the model for the PT interaction, the difference score for tests, will be examined. There are four components going into the theoretical structure of the

TABLE 8. Variance components model for standardized scores of test design 6

| | |
|---|---|
| P | $\sigma^2_{pof:t} + 2\sigma^2_{pf:t} + k\sigma^2_{pto} + 2k\sigma^2_{po} + 2k\sigma^2_{pt} + 4k\sigma^2_{p}$ |
| PT | $\sigma^2_{pof:t} + 2\sigma^2_{pf:t} + k\sigma^2_{pto} + 2k\sigma^2_{pt}$ |
| PO | $\sigma^2_{pof:t} + k\sigma^2_{pto} + 2k\sigma^2_{po}$ |
| PTO | $\sigma^2_{pof:t} + k\sigma^2_{pto}$ |
| PF:T | $\sigma^2_{pof:t} + 2\sigma^2_{pf:t}$ |
| POF:T | $\sigma^2_{pof:t}$ |

difference score of interest. The $\sigma^2_{pt}$ measures the consistency of the difference scores across occasions and forms, while $\sigma^2_{pto}$ is a measure of the inconsistency of the difference scores for the two occasions. The $\sigma^2_{pf:t}$ reflects the inconsistency of forms within the tests.

In the present design form is considered a random facet; test and occasion are fixed facets. Because occasion is fixed, $k\sigma^2_{pto}$ has to be regarded as part of the universe score variance, together with $2k\sigma^2_{pt}$, while the two other components define the error variance. Thus the defining formula for the reliability of the difference score for tests will be,

$$\text{alpha}_{pt(k)} = \frac{k\sigma^2_{pto} + 2k\sigma^2_{pt}}{\sigma^2_{pof:t} + 2\sigma^2_{pf:t} + k\sigma^2_{pto} + 2k\sigma^2_{pt}} \tag{33}$$

$$\text{alpha}_{pt(1)} = \frac{\sigma^2_{pto} + \sigma^2_{pt}}{\sigma^2_{pof:t} + \sigma^2_{pf:t} + \sigma^2_{pto} + \sigma^2_{pt}} \tag{34}$$

Formula (35) is the computing form of formula (33). A computing formula of (34) could be developed. But the form would be

$$\text{alpha}_{pt(k)} = \frac{MS_{pt} - MS_{pf:t}}{MS_{pt}} \qquad (35)$$

too unwieldy to be of any practical value. Formulas in terms of the correlation matrix of test design 6 for (33) and (34) could also be developed, as was done for design 5. This will be dropped in the present case because the formulations will be extremely awkward.

Design 7.

A modified design of the preceding one can be thought of, having tests, two occasions, and forms crossed with tests instead of nested. Design 7 will be an extended design 2 by adding two occasions. Thus the difference score between tests can be generalized across both occasions and forms. The variance components

TABLE 9. Variance components model
for standardized scores of test design 7

| | |
|---|---|
| P | $\sigma^2_{pfto}+2\sigma^2_{pfo}+2\sigma^2_{pft}+4\sigma^2_{pf}+k\sigma^2_{pto}+2k\sigma^2_{po}+2k\sigma^2_{pt}+4k\sigma^2_{p}$ |
| PT | $\sigma^2_{pfto}+2\sigma^2_{pft}+k\sigma^2_{pto}+2k\sigma^2_{pt}$ |
| PO | $\sigma^2_{pfto}+2\sigma^2_{pfo}+k\sigma^2_{pto}+2k\sigma^2_{po}$ |
| PTO | $\sigma^2_{pfto}+k\sigma^2_{pto}$ |
| PF | $\sigma^2_{pfto}+2\sigma^2_{pfo}+2\sigma^2_{pft}+4\sigma^2_{pf}$ |
| PFT | $\sigma^2_{pfto}+2\sigma^2_{pft}$ |
| PFO | $\sigma^2_{pfto}+2\sigma^2_{pfo}$ |
| PFTO | $\sigma^2_{pfto}$ |

model for test design 7 with 2 tests,2 crossed occasions, and k crossed forms is presented in TABLE 9.

There are theoretical structures for the variance of several kinds of scores in TABLE 9 that might be of considerable interest to examine. In the context of the present discussion, however, only the model for the difference score between tests, the variance of the PT interaction, will be analyzed.

The present design is powerful enough to provide detailed information on how the difference score behaves. As a matter of fact, the model for the difference score of tests in TABLE 9 has a clear meaning in the Thorndike (1951) sense. He described a test score as possibly influenced by general-lasting effects, general-temporary effects, specific-lasting effects, and specific-temporary effects (plus a fifth group of various random effects). Now, it is of considerable interest to look at the model for the difference score with this perspective in mind: The $\sigma^2_{pt}$ is the general component, indicating how much of the observed difference score variance can be attributed to a common source across forms and occasions (general-lasting). Next, the $\sigma^2_{pto}$ is indicating to what extent the difference score is inconsistent from the first to the second occasion (general-temporary). The $\sigma^2_{pft}$ reflects the inconsistency of forms across the two occasions, thus being a case of the specific-lasting effect. Lastly, the $\sigma^2_{pfto}$ is a measure of the specific-temporary effect in the difference score, together with a hodge-podge of random effects, because the design is an unreplicated one in the sense that there is only one observation within each of the test-occasion-form cells.

In defining the reliability of the difference score between tests, it should be remembered that test and occasion are fixed facets, while forms are considered to be a random facet. There-

fore, the universe score variance should consist of $k\sigma^2_{pto}$ and $2k\sigma^2_{pt}$, and the defining formulas for the reliability of the difference score between tests for test design 7 will read,

$$\text{alpha}_{pt(k)} = \frac{k\sigma^2_{pto} + 2k\sigma^2_{pt}}{\sigma^2_{pfto} + 2\sigma^2_{pft} + k\sigma^2_{pto} + 2k\sigma^2_{pt}} \tag{36}$$

$$\text{alpha}_{pt(1)} = \frac{\sigma^2_{pto} + \sigma^2_{pt}}{\sigma^2_{pfto} + \sigma^2_{pft} + \sigma^2_{pto} + \sigma^2_{pt}} \tag{37}$$

The computing form of formula (36) in terms of observed MS's will be,

$$\text{alpha}_{pt(k)} = \frac{MS_{pt} - MS_{pft}}{MS_{pt}} \tag{38}$$

No effort will be made to derive a computing form of formula (37) in terms of observed MS's, neither will (36) and (37) be developed as functions of the correlation matrix of test design 7. The formulations would be quite impractical and also of less theoretical interest.

An overview of the 7 test designs examined in this paper is presented in TABLE 10. For convenience, only two forms for each test or each test-occasion combination are included for designs 1,2,5,6,and 7, instead of k, which is the general case treated above. The linear combination of scores for the difference score between tests for the $i^{th}$ person is given for each of the seven designs.

TABLE 10. An overview of test designs 1-7.

| Design 1 | | | |
|---|---|---|---|
| $T_1$ | | $T_2$ | |
| $F_1$ | $F_2$ | $F_3$ | $F_4$ |
| $+X_{11i}$ | $+X_{12i}$ | $-X_{23i}$ | $-X_{24i}$ |

| Design 2 | | | |
|---|---|---|---|
| $T_1$ | | $T_2$ | |
| $F_1$ | $F_2$ | $F_1$ | $F_2$ |
| $+X_{11i}$ | $+X_{12i}$ | $-X_{21i}$ | $-X_{22i}$ |

| Design 3 | | | |
|---|---|---|---|
| $T_1$ | | $T_2$ | |
| $O_1$ | $O_2$ | $O_1$ | $O_2$ |
| $F_1$ | $F_1$ | $F_2$ | $F_2$ |
| $+X_{111i}$ | $+X_{121i}$ | $-X_{212i}$ | $-X_{222i}$ |

| Design 4 | | | |
|---|---|---|---|
| $T_1$ | | $T_2$ | |
| $O_1$ | $O_2$ | $O_1$ | $O_2$ |
| $F_1$ | $F_2$ | $F_3$ | $F_4$ |
| $+X_{111i}$ | $+X_{122i}$ | $-X_{213i}$ | $-X_{224i}$ |

| Design 5 | | | | | | | |
|---|---|---|---|---|---|---|---|
| $T_1$ | | | | $T_2$ | | | |
| $O_1$ | | $O_2$ | | $O_1$ | | $O_2$ | |
| $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | $F_6$ | $F_7$ | $F_8$ |
| $+X_{111i}$ | $+X_{112i}$ | $+X_{123i}$ | $+X_{124i}$ | $-X_{215i}$ | $-X_{216i}$ | $-X_{227i}$ | $-X_{228i}$ |

| Design 6 | | | | | | | |
|---|---|---|---|---|---|---|---|
| $T_1$ | | | | $T_2$ | | | |
| $O_1$ | | $O_2$ | | $O_1$ | | $O_2$ | |
| $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_3$ | $F_4$ |
| $+X_{111i}$ | $+X_{112i}$ | $+X_{121i}$ | $+X_{122i}$ | $-X_{213i}$ | $-X_{214i}$ | $-X_{223i}$ | $-X_{224i}$ |

| Design 7 | | | | | | | |
|---|---|---|---|---|---|---|---|
| $T_1$ | | | | $T_2$ | | | |
| $O_1$ | | $O_2$ | | $O_1$ | | $O_2$ | |
| $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ |
| $+X_{111i}$ | $+X_{112i}$ | $+X_{121i}$ | $+X_{122i}$ | $-X_{211i}$ | $-X_{212i}$ | $-X_{221i}$ | $-X_{222i}$ |

Discussion.

Altogether 7 test designs have been examined with a view to the generalizability of differences between test scores. It is believed that the designs chosen will cover most of test designs actual for such purposes. Yet, the designs should be looked upon as illustrative and suggestive of a general procedure /hopefully and not as exhaustive of possible test designs, being diverse enough to enable the informed reader to proceed on his own with test designs that are appropriate for his specific objective.

No deep and thorough discussion of the meaning of difference scores has been aimed at in the present paper. In developing the various models for assessing the generalizability of difference scores it has though been assumed that such an undertaking is meaningful and worth while. Recently, Cronbach and Furby (1970), and Cronbach,Gleser,Nanda and Rajaratnam (1972) have questioned difference and gain scores as constructs. It seems to the author that there might be more problems involved in gain scores than in differences between rationally defined constructs, operationalized in two separate tests. In a simultaneous administration of a differential aptitude test, there is hardly any more problems connected with difference scores than sum scores. They are both linear combinations of part scores going into the composite. The interpretation of empirically demonstrated differential constructs has to be closely linked to the content and format of tests being employed.

It should though be admitted that interpreting difference scores may have some of the problematic character as bipolar factors in factor analysis. This is particularly the case when

a linear combination like a difference score is an a posteriori construction. The preceding derivation of generalizability estimates for difference scores has elaborated on the assumption of a priori rationally defined constructs, to be critically assessed by analyzing data generated by proper test designs. It is also apparent that in restricting the issue to differences between two tests, rather than to differences among more than two tests, i.e. to profiles generally, the interpretability of such scores has been considerably facilitated.

Seemingly, characteristic for studying difference score reliability in the past has been a freedom of choice of values to insert in formula (3). Often, the values have been taken from other sets of data than just that being analyzed. As $r_{wb}$ in formula (3) is a reliability estimate, this freedom of choice seems to have implied that whatever reliability estimate at hand, or conveniently reached, could be put into the formula. This is certainly not correct, which can easily be seen from the differences between formulas (23) and (24) as contrasted with formulas (25) and (26), where difference scores between tests and difference scores between occasions are assessed, respectively.

A prominent feature of the development of generalizability estimates in the present paper is that test designs are complete in the sense that all information needed for estimating generalizability is available in test data generated by the design. What is evidently clear from the presentation of test designs above, is that by taking into account the statistical properties of the lowest unit scores, i.e. the scores on the level of forms, there is no need to go outside test data on hand to fill in the formula for difference score reliability.

The approach to generalizability estimates for difference scores developed here,is to a very great extent concerned with the internal consistency of sets of comparable, or nominally parallel, difference scores. It is thus closely related to a kind of construct validation procedure, where one intends to assess to what extent the difference scores are tapping one construct, so to say, a bipolar construct. Thus the problem is relatively complex in the case of difference scores between two tests, in that the difference scores imply two distinct constructs, if demonstrated to be reliable. On differential constructs persons tend to be high on the one and low on the other. As convergent tests indicate a form of construct validity, so do divergent tests. Divergent tests indicate discriminant validity, which repeatedly is called differential construct validity above.

Design 7 is an example of a very general design that is powerful enough to indicate to what extent differential constructs are measured consistently across both forms and occasions.In a real sense, design 7 gives distinct estimates of equivalence and stability of difference scores, while Stanley's (1967) test design can only give an estimate of equivalence and stability that is confounded.

A totally overlooked aspect of difference score reliability seems to be that also difference scores obey the Spearman-Brown prophecy formula. In the past, difference score reliability is always given as the reliability of <u>one</u> difference score. The demonstration that the Spearman-Brown prophecy formula applies to difference scores as well as sum scores, should make it possible to increase test length in order to obtain satisfactory

reliability coefficients for composite difference scores, provided the reliability of one difference is promising enough.

In generalizability theory, the notion of a defined universe of scores to which one intends to generalize is of crucial importance. He has to be quite specific about what should be considered comparable scores. In the case of differences between tests constructed to measure differential aptitudes, one has to bring in another facet in order to be able to specify a family of difference scores over which to generalize. FIGURE 1 should remind us that one difference score can be defined into several families of such scores, each serving particular testing objectives.It is up to the test user to specify what universe of difference scores is appropriate for his purpose, and construct test designs that meet his requirements.

The multi-facet character of difference scores should be noted. The minimum test design is a two-facet one. But frequently more complex designs are needed,and should not be avoided. Certainly, much test data are placed on file that contain much unexploited information on the generalizability of difference scores. For example, the Irwin (1966) test data on WISC could most profitably be analyzed according to a test design much more complex than any of those examined in the present paper. Actually, his data would fit a four-facet test design with tests (verbal and performance), subtests within tests, items within subtests, and two occasions as the facets. This would indeed prove to be a very sophisticated model for the structure of the difference score variance with altogether six different components. However, this would be the model that best preserves the information in test

data concerning the consistency of the difference scores between verbal and performance tests across subtests,items,and occasions. A simpler model would spoil some information on this consistency.

In the following two sets of real-world test data of two different designs will be reanalyzed in order to illustrate how data can be exploited concerning the internal consistency of composite difference scores.

Irwin (1966) analyzed WISC data to estimate reliabilities for subtests,for the verbal scale,the performance scale,and for the full scale. He also estimated the reliabilities of the difference scores between subtests, both within and between the verbal and the performance scales. His estimates of difference score reliabilities are not quite meaningful according to the formulations of such estimates as given in the present paper. This was also the conclusion reached by Stanley (1967). In the context of the present paper it is not clear which family of difference scores Irwin is generalizing to (cfr.FIGURE 1).

While Irwin was concerned with estimating the reliability of differences between subtests, it might seem of even more substantive interest to estimate the reliability of the difference between composite verbal and composite performance scales. This should indicate to what extent verbal and performance tests as operationalized by Wechsler represent differential constructs.

The reanalysis subsequently to be performed will elaborate on the matrix of intersubtest correlations for age level 11, which is part of Irwin's (1966) Table 3, p.291. The complete correlation matrix is presented here in TABLE 11. Subtest Digit Span has been omitted from the verbal scale in order to have 5 subtests within

each of the two scales, verbal and performance. This will be more convenient for the analysis. The appropriate test design for the WISC data as found in TABLE 11, is our test design 1, as subtests might be considered forms nested within the tests.

TABLE 11. Intercorrelation matrix of WISC subtests, age 11.

| | | Ve | | | | | Pe | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | In | Co | Ar | Si | Vo | Pc | Pa | Bd | Oa | Cd | |
| Ve | In | 1.- | .83 | .74 | .84 | .90 | .54 | .59 | .68 | .45 | .63 | |
| | Co | .83 | 1.- | .66 | .72 | .82 | .65 | .56 | .77 | .49 | .64 | |
| | Ar | .74 | .66 | 1.- | .69 | .75 | .62 | .63 | .58 | .44 | .52 | |
| | Si | .84 | .72 | .69 | 1.- | .80 | .41 | .45 | .68 | .34 | .47 | |
| | Vo | .90 | .82 | .75 | .80 | 1.- | .57 | .48 | .71 | .46 | .65 | |
| Pe | Pc | .54 | .65 | .62 | .41 | .57 | 1.- | .53 | .62 | .59 | .51 | Sum=65.24 |
| | Pa | 59 | 56 | .63 | .45 | .48 | .53 | 1.- | .53 | .53 | .55 | |
| | Bd | .68 | .77 | .58 | .68 | .71 | .62 | .53 | 1.- | .61 | .68 | |
| | Oa | .45 | .49 | .44 | .34 | .46 | .59 | .53 | .61 | 1.- | .71 | |
| | Cd | .63 | .64 | .52 | .47 | .65 | .51 | .55 | .68 | .71 | 1.- | |

Note.-- Ve=verbal, Pe=performance, In=information, Co=comprehension, Ar=arithmetic, Si=similarities, Vo=vocabulary, Pc=picture completion, Pa=picture arrangement, Bd=block design, Oa=object assembly, Cd=coding.

Thus, formula (6) should be the correct estimate for the generalizability of the difference between the verbal and the performance scales.

The needed values to be inserted in formula (6) can all be found in TABLE 11 by averaging the sum of the correlation coefficients for each of the two categories of correlation, the between-tests-between-subtests and the within-tests-between-subtests correlations. The $\bar{r}_{bb}$ is found to be 0.560, the $\bar{r}_{wb}$ 0.680, and $k = 5$, as there are 5 subtests within each of the tests.

$$\text{alpha}_{D(5)} = \frac{k\bar{r}_{wb} - k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{wb} - k\bar{r}_{bb}} = \frac{5.0,680-5.0,560}{1+4.0,680 - 0,560} = \underline{0,652}$$

$\text{Alpha}_{D(5)}$ estimates the reliability of the difference scores between the verbal and the performance scales (each of them a sum of 5 subtests) to 0,652. This means that 65% of the difference score variance can be considered universe score variance.

The reliability for one average difference between tests can be estimated by formula (3),

$$\text{alpha}_{D(1)} = \frac{\bar{r}_{wb} - \bar{r}_{bb}}{1 - \bar{r}_{bb}} = \frac{0,680 - 0,560}{1 - 0,560} = \underline{0,273}$$

On the average, the difference scores between subtests between tests correlate 0,273, which is the reliability of one difference score. It should be recalled that by employing the general Spearman-Brown prophecy formula to $\text{alpha}_{D(1)}$, $\text{alpha}_{D(5)}$ is obtained,

$$\text{alpha}_{D(5)} = \frac{5.0,273}{1 + 4.0,273} = \underline{0,652}$$

While $\text{alpha}_{D(1)}$ can be regarded as an expected correlation between differences of two subtests, one from each of the tests; $\text{alpha}_{D(5)}$ is the expected correlation between two sum composites of 5 subtest differences.

It might perhaps be concluded that $\text{alpha}_{D(5)}$ is indicating that WISC comes out with a fairly good differential construct validity for the verbal and performance scales. If the battery can be lengthened by adding another 5 subtests to each of the tests, the difference score reliability will increase in accordance with the Spearman-Brown formula to 0,798.

It will be of considerable interest to examine the variance structure of the sum score across subtests and tests for the WISC data, to see to what extent the differential constructs influence the sum score variance. This is important to know in order to interpret the sum score. For this purpose the components model for the sum score variance ($MS_p$) will be used, where

$$E(MS_p) = \sigma^2_{pf:t} + 2\sigma^2_{pt} + 4\sigma^2_p$$

for the particular test design used there. In the present case, that model will be changed in the coefficients for the components as a result of employing 5 subtests (forms) instead of 2. Thus, for the WISC data in TABLE 11,

$$E(MS_p) = \sigma^2_{psub:t} + 5\sigma^2_{pt} + 10\sigma^2_p$$

According to TABLE 3 the components can be estimated as functions of the correlation matrix, such that

$$\sigma^2_p = \bar{r}_{bb} = \underline{0,560}$$

$$\sigma^2_{pt} = \bar{r}_{wb} - \bar{r}_{bb} = 0,680 - 0,560 = \underline{0,120}$$

$$\sigma^2_{psub:t} = 1 - \bar{r}_{wb} = 1,000 - 0,680 = \underline{0,320}$$

$$MS_p = 0,320 + 5.0,120 + 10.0,560$$

$$= 0,320 + 0,600 + 5,600 = 6,520$$

The $MS_p = 6,520$ is $1/2k = 1/10$ of the sum of the correlation matrix in TABLE 11, which is 65,24. In setting $MS_p$ to unit variance, the following proportions of variance is obtained for respective weighted variance components,

$$MS_p = 1,000 = 0,049 + 0,092 + 0,859$$

In the context of the present discussion,it is the contribu-

bution of $5\sigma^2_{pt}$ to sum score variance which is of particular concern. As shown, 9% of the sum score variance can be attributed to the differential constructs as measured by the differences between verbal and performance scales. More concretely, this means that persons with the same sum score have different scale profiles. If these profiles are very different,it becomes difficult to interpret individual differences in sum score meaningfully; even the same sum scores have different meanings.

In the present case, the contribution of the differential traits of 9% is small compared to the contribution yielded by the general trait as represented by $10\sigma^2_p$, which is 86%. This general trait is common to both verbal and performance tests. Thus, it might be concluded that the sum score is substantially loaded with a general factor running through all of the WISC subtests.

A somewhat changed picture of test score variance is obtained by examining the variance structure of an average subtest, which will be a sum of unweighted variance components,

$$E(V_{sub}) = \sigma^2_{psub:t} + \sigma^2_{pt} + \sigma^2_p$$
$$= 0{,}320 + 0{,}120 + 0{,}560 = 1{,}000,$$

which is the variance of the subtests in the correlation matrix of TABLE 11. The difference in relative contribution to score variance by the components for average subtest score compared to sum score across subtests, is certainly dependent on the hierarchical structure imposed by the variance components model. The general component, $\sigma^2_p$, is defined as common to all parts of the test battery, while the more specific component, $\sigma^2_{pt}$, is common to either the verbal or the performance scales, but not

to both. Therefore, by lengthening the battery 2k times, the general component will according to the model increase by a factor of 2k, while the differential component will increase by a factor of k.

It might be tentatively concluded from the reanalysis of the Irwin (1966) data that WISC for the 11 year level might be used both as a meaningful one-dimensional measure to tap a general construct and/as a measure of differential constructs, for some purposes as represented by the verbal and the performance scales. A somewhat misleading picture of the WISC battery as a one-dimensional measure might be obtained by wrongly applying traditional alpha to the correlation matrix of TABLE 11. For this purpose the overall average intersubtest correlation, $\bar{r}_{ij}$ (where $i \neq j$), ignoring the differentiation previously made between $\bar{r}_{bb}$ and $\bar{r}_{wb}$. With an $\bar{r}_{ij} = 0{,}614$, the assumed "homogeneous" alpha will be,

$$\text{alpha}_{\text{sum}(2k)} = \frac{5.0{,}614}{1 + 4.0{,}614} = \underline{0{,}888}$$

By applying traditional alpha to the WISC data in TABLE 11, the contribution of the general component will be overestimated by ignoring the differential effects of the verbal and the performance scales.

Recently, Lauvås (1973) in a study of college dropouts used difference scores as measures of students' experienced satisfaction, or dissatisfaction, of several kinds in the college environment. He adopted Pervin's (1967) approach, applying a semantic differential as the method for measuring these attitudes. For example, the concepts might be College and Ideal College, and the consonance or dissonance between the concepts are tapped

by employing the same bipolar adjectives as scales for both concepts. Thus, in this particular test design scales are crossed with concepts, such that it matches our test design 2, where forms (scales) are crossed with tests (concepts).

Lauvås, like Pervin, correctly used the difference between the same scales for the two concepts as the measure of satisfaction. He used 20 scales in his instrument, thus generating altogether 20 difference scores, all of them intended to measure the same bipolar dimension, reflecting both satisfaction and dissatisfaction in the students' experience of the real college they met as compared to the college they saw as an ideal one.

The matrix of N persons by 20 difference scores constituted data from which the values relevant for formula (5) were taken to estimate the internal consistency of these 20 difference scores by way of a traditional alpha coefficient. In accordance with the logic of deriving generalizability estimates for difference scores undertaken in the present paper, formula (22) would be the correct one to apply in assessing the reliability of the difference between the two concepts, say, College and Ideal College, across the 20 scales. For this particular difference, Lauvås got $\text{alpha}_{D(20)} = 0{,}82$, by using raw score formula (5). Almost the same result is obtained by inserting the average correlation coefficients from the intercorrelation matrix of the 40 original scores into formula (22) going into the test design.[1] Recalling that 3 categories of correlation coefficients were defined for test design 2, the crossed design, the following coefficients were obtained: $\bar{r}_{bb} = -0{,}014$ $\bar{r}_{bw} = 0{,}115$ and $\bar{r}_{wb} = 0{,}172$. With $k = 20$, the reliability of the composite difference between College and Ideal College

1) A discrepancy in results is solely due to the possibility of non-homogeneous variances in scales when raw difference scores are used, as in formula (5).

for the Lauvås data becomes,

$$\text{alpha}_{D(20)} = \frac{k\bar{r}_{wb} - k\bar{r}_{bb}}{1+(k-1)\bar{r}_{wb}-\bar{r}_{bw}-(k-1)\bar{r}_{bb}}$$

$$= \frac{20.\ 0,172 - (20.-0,014)}{1 + 19.\ 0,172 - 0,115 - (19.-0,014)} = \underline{0,84}$$

It should be noted that formula (5) applies to test design 1, which is shown in the derivation of formula (6), and also to test design 2. In order to derive formula (22) from formula (5), one has to be quite explicit about how to write the sum of the difference score variances and the composite difference score variance in terms of the correlation matrix of test design 2 as compared to the corresponding variances for test design 1. By observing this precaution, formula (22) is derived from formula (5) this way,

$$\text{alpha}_{D(k)} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma V_{d_i}}{V_D}\right) \qquad (5)$$

$$= \left(\frac{k}{k-1}\right)\left(1 - \frac{k(2-2\bar{r}_{bw})}{k(2-2\bar{r}_{bw})+k(k-1)(2\bar{r}_{wb}-2\bar{r}_{bb})}\right)$$

$$= \frac{k\bar{r}_{wb} - k\bar{r}_{bb}}{1 + (k-1)\bar{r}_{wb} - \bar{r}_{bw} - (k-1)\bar{r}_{bb}} \qquad (22)$$

The fairly high difference score reliability obtained by Lauvås is a result of an almost ideal combination of low correlations among different scales for the two concepts and relatively high correlations among scales within the concepts. What is implied in the correlation coefficients reported from the Lauvås data, is that the

average interdifference correlation is 0,208 which is even higher than the average interscale correlation within concepts. The value of 0,208 is found by solving for the average interdifference correlation in the general Spearman-Brown formula with k = 20 and $alpha_{D(20)} = 0,84$.

The substantial difference score reliability of 0,82 does not in any way suggest that the score is valid. This is another story. It should be mentioned, though, that the difference scores generated in the Lauvås study, correlated on the average higher with the dependent variable, dropouts versus not-dropouts, as compared to the correlations obtained by using more traditional predictors, like sex, age, average high school mark, etc.

The two real-world studies reviewed above seem to indicate that difference scores can be meaningfully interpreted in a substantive context. Undoubtedly, aside from the problem of meaning, difference score reliability should be considered a needed contribution to the assessment of discriminant validity of constructs, thus being an aspect of the construct validation procedure. In this perspective, the present paper may be regarded as a continued and extended discussion of the problems raised by Campbell and Fiske (1959) concerning convergent and discriminant validity. It is also believed to be congenial with Cronbach's (1971) basic outlook on test validation.

References.

Bereiter, C. 1963. Some persisting dilemmas in the measurement of change. In Harris, C.W. (Editor). Problems in Measuring Change. Madison, Wisconsin: The University of Wisconsin Press.

Campbell, D. T. and Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cronbach, L. J. 1971. Test validation. In Thorndike, R. L. (Editor). Educational Measurement. Washington D. C.: American Council on Education.

Cronbach, L. J. and Snow, R. E. 1968. Project on Individual Differences in Learning Ability as a Function of Instructional Variables. Annual Report No. 2. School of Education, Stanford University.

Cronbach, L. J. and Snow, R. E. 1969. Individual Differences in Learning Ability as a Function of Instructional Variables. Final Report. ERIC Document Reproduction Service ED-029-001.

Cronbach, L. J. and Furby, L. 1970. How we should measure "change" - or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley.

Eikeland, H. H. 1970. Coefficient alpha and the expected variance-covariance matrix of random composite measurements. Oslo: Institute for Educational Research. Mimeographed.

Eikeland, H. M. 1972. The structure of generalizability theory for hierarchically stratified tests. Oslo. Institute for Educational Research. Mimeographed.

Guilford, J. P. 1954. Psychometric Methods. New York:McGraw-Hill.

Gulliksen, H. 1950. Theory of Mental Tests. New York: John Wiley.

Guttman, L. 1953. A special review of Harold Gulliksen's Theory of Mental Tests. Psychometrika, 28, 123-130.

Hills, J. R. 1971. Use of measurement in selection and placement. In Thorndike, R. L. (Editor). Educational Measurement. Washington D. C.: American Council on Education.

Horst, P. 1966. Psychological Measurement and Prediction. Belmont, California: Wadsworth Publishing Company.

Hoyt, C. 1941. Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Irwin, D. O. 1966. Reliability of the Wechsler Intelligence Scale for Children. Journal of Educational Measurement, 3, 287-292.

Lauvås, P. 1973. Student, miljø og studiefrafall. (Student, environment, and wastage.) Oslo: Institute for Educational Research. Mimeographed.

Magnusson, D. 1967. Test Theory. Reading, Mass.: Addison-Wesley.

McNemar, Q. 1969. Psychological Statistics. Fourth Edition. New York: John Wiley.

Millman, J. and Glass, G. V. 1967. Rules of thumb for writing the ANOVA table. Journal of Educational Measurement, 4, 41-51.

Mosier, C. I. 1951. Batteries and profiles. In Lindquist, E. F. (Editor). Educational Measurement. Washington D. C.:American Council on Education.

Pervin, L. A. 1967. A twenty-college study of student X college interaction using TAPE (Transactional Analysis of Person and Environment). Journal of Educational Psychology, 58, 290-302.

Rabinowitz, W. and Eikeland, H. M. 1964. Estimating the reliability of tests with clustered items. Pedagogisk Forskning (Scandinavian Journal of Educational Research), 8, 85-106.

Rulon, P. J. 1939. A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9, 99-103.

Stanley, J. C. 1957. KR 20 as the stepped-up mean item intercorrelation. 14th Yearbook of the National Council on Measurement in Education, 78-92.

Stanley, J. C. 1961. Analysis of unreplicated three-way classifications with applications to rater bias and treatment independence. Psychometrika, 26, 205-219.

Stanley, J. C. 1967. General and specific formulas for reliability of differences. Journal of Educational Measurement, 4, 249-252.

Stanley, J. C. 1971. Reliability. In Thorndike, R. L. (Editor). Educational Measurement. Washington D. C.: American Council on Education.

Thorndike, R. L. 1951. Reliability. In Lindquist, E. F. (Editor). Educational Measurement. Washington D. C.: American Council on Education.

Thorndike, R. L. 1971. Educational measurement for the seventies. In Thorndike, R. L. (Editor). Educational Measurement. Washington D. C.: American Council on Education.

Thorndike, R. L. and Hagen, E. 1969. Measurement and Evaluation in Psychology and Education. Third Edition. New York: John Wiley.

Webster, H. and Bereiter, C. 1963. The reliability of changes measured by mental test scores. In Harris, C. W. (Editor). Problems in Measuring Change. Madison, Wisconsin: The University of Wisconsin Press.