Psychological Bulletin
1970, Vol. 74, No. 1, 68–80

HOW WE SHOULD MEASURE "CHANGE"—OR SHOULD WE?[1]

LEE J. CRONBACH[2] AND LITA FURBY[3]

Stanford University

Procedures previously recommended by various authors for the estimation of "change" scores, "residual" or "basefree" measures of change, and other kinds of difference scores are examined. A procedure proposed by Lord is extended to obtain more precise estimates, and an alternative to the Tucker-Damarin-Messick procedure is offered. A consideration of the purposes for which change measures have been sought in the past leads to a series of recommended procedures which solve research and personnel-decision problems *without* estimation of change scores for individuals.

A persistent puzzle in psychometrics has been "the measurement of change." Many investigators have felt, for reasons good or bad, that their substantive questions required a measure of gain in ability or shift in attitude. "Raw change" or "raw gain" scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement. Although the unsuitability of such scores has long been discussed, they are still employed, even by some otherwise sophisticated investigators.

At the end of this paper the authors argue that gain scores are rarely useful, no matter how they may be adjusted or refined. The authors also distinguish four kinds of inquiry for which such scores have been used, and conclude that only one of these purposes is well served by any kind of gain score. This argument applies not only to changes over time, but also to other differences between two variables.

The first part of the paper proposes superior ways of estimating true change and true residual change scores. It may seem pointless to discuss such matters when in the end we recommend against their use (save in a few kinds of investigation). However, the development of formulas clarifies the model, providing a base for the final recommendations, and allows us to explain the limitations of previous papers on the subject. Furthermore, it develops superior estimators for the kinds of problem where we do recommend using a measure of change. Very likely some investigators will decide to obtain change or difference scores, even for problems where we consider such measures inappropriate. Such a person will often find one of our estimation formulas better than those now suggested in the literature.

Methods of handling data from successive measurements have been offered by several writers (Harris, 1963). We shall be particularly interested in the related proposals of Lord (1956, 1958, 1963) and McNemar (1958), who estimate an individual's "true change"; their approach is clearly superior to conventional techniques. This paper extends the Lord-McNemar reasoning to get a still better estimate.

DuBois (1957) and other investigators recommend a "residual gain" score as a substitute for the "raw gain" score. A gain is residualized by expressing the posttest score as a deviation from the posttest-on-pretest regression line. The part of the posttest information that is linearly predictable from the pretest is thus partialled out. Tucker, Damarin, and Messick (1966) draw attention to the "true residual gain," which they refer to as a "basefree measure of change."

FORMULATION

The Lord-McNemar argument considers measures X and Y, obtained by applying the same operation to the subject on two occasions. The subject has an observed difference score $D = Y - X$. Scores X_∞ and Y_∞, representing the person's "true" status at these times, are postulated. The "true difference" D_∞ equals $Y_\infty - X_\infty$. The key topic of the McNemar paper and the Lord papers is the determination of regression coefficients for an estimator of the form:

$$\hat{D}_\infty = \beta_1 X + \beta_2 Y + \text{constant} \qquad [1]$$

As Lord has written more extensively than McNemar on this matter, we shall refer only to Lord hereafter, though many of the comments also apply to what McNemar has said.

The development holds so long as X and Y are referred to the same numerical scale. That is to say, the investigator must be willing to say what score on Y (or Y_∞) is comparable to a given score on X (or X_∞). The relation must be reciprocal in the sense that, if the given X is mapped into a certain Y, that value of Y is mapped into the given X. It is particularly important to note that this formulation sidesteps the philosophically troublesome question, Are pretest and posttest "measuring the same variable"? A common metric is the only requirement. Even this requirement is dispensed with when we turn to a regression estimate of outcome.

The model applies to any two measures that can be sensibly expressed on the same numerical scale. This might be, for example, a standard-score scale or an age-equivalence scale. The data might be ratings of two distinct traits expressed on the same reference scale. Hence statements made about change scores can be extended to any kind of difference score. Among the more famous lines of research employing difference scores are work on "overachievement," "empathy and insight," "self-concept" versus "ideal self," and "differential aptitudes." Indices of the same psychometric character are also involved in studies of retention and transfer.

Assumptions. There are two variables and X Y. For each variable, the expected value over independent observations of the same person defines a true score: $E(X) = X_\infty$ and $E(Y) = Y_\infty$. Then $D_\infty = Y_\infty - X_\infty$. The investigator who identifies Y with X as "the same operation" must keep the true scores distinct. In a study of change, X_∞ is the person's average score over measurements of X that might be made at Time 1; Y_∞ is the average over observations by the same procedure at Time 2.

Errors are uncorrelated with true scores. But we do not assume that all correlations ρ_{XY} are equal. Sometimes X and Y observations are "linked," as when the two scores are obtained from a single test or battery administered at one sitting, or when observations on different occasions are made by the same observer. The correlation between linked observations will ordinarily be higher than that between independent observations. This distinction has not been made in the literature on change, but it does appear in a paper by Stanley (1967) on difference scores.

We develop a formal mathematical model so that some parts of our argument can be explicit rather than intuitive. We retain the classical concept of strictly parallel observations, but modify the classical concept of independence, as suggested above.

1. Let X_{pi} represent the observed score of person p on the X variable observed under condition i. The condition i may be, for example, a particular form of the test that measures X. In studies where there are several sources of "error" such as test form, observer, and short-term fluctuations in the state of the person, we assume that these sources are completely confounded in the design used to determine reliability coefficients and correlations.

2. Where the classical theory considers true score as the sum of observed score and error, we introduce two random errors: $X_{pi} = X_{\infty p} + e_{pi} + f_{pi}$. Likewise, $Y_{pi} = Y_{\infty p} + \mathbf{e}_{pi} + \mathbf{f}_{pi}$. This is required to formalize the concept of independence adequately.

If, as in the classical concept of parallel measures, one assumes that the several measures of X (or Y) are strictly interchangeable, there can be no linkage. Machinery for explicitly defining independence can be developed with some concepts from generalizability theory. There is a universe of possible conditions of observation of X. (While observations may vary with respect to test form, occasion, observer, etc., we shall avoid here the complications of multifacet theory [Gleser, Cronbach, & Rajaratnam, 1965]. That theory recognizes, as this paper does not, that one may have two measurements with the same test form on different occasions, or two measurements on different test forms on the same occasion, or

[1] Initial work on this paper was conducted under a contract with the Cooperative Research Program of the United States Office of Education. We thank many colleagues for helpful comments, particularly Janet D. Elashoff, Gene V. Glass, Ingram Olkin, and Julian C. Stanley, Jr.

[2] Requests for reprints should be sent to Lee J. Cronbach, School of Education, Stanford University, Stanford, California 94305.

[3] Now at the University of Poitiers, France.

two measurements where both form and occasion differ.) There is a universe of observations of Y, and we shall assume that these may be made under the same set of conditions i that are used for X observations. Then observations X_i and Y_i made under the same condition i are said to be linked, and observations X_i and $Y_{i'}$ made under different conditions are said to be independent.

Formally, the model specifies that conditions of observations are drawn from the universe. If a single i is drawn and used to obtain both scores X_{pi} and Y_{pi} for person p, we have linkage; if i and i' are drawn independently to give scores X_{pi} and $Y_{pi'}$, we have independence. Errors are random and independent, except that when X and Y are both observed under condition i, the f and $\mathbf{f}$ components are sampled simultaneously. It will *not* be true, in general, that $\sigma(f_{pi},\mathbf{f}_{pi}) = 0$. The following assumptions are made regarding all e components: Their mean over persons is zero for every condition; their variance over persons is the same for every condition; their intercorrelations with other components are zero. The $\mathbf{e}$ components satisfy the same assumptions and $\sigma(e_{pi},\mathbf{e}_{pi}) = \sigma(f_{pi},\mathbf{e}_{pi}) = 0$. With regard to the f components, zero means, equal variances, and zero correlations with true scores are assumed (likewise for $\mathbf{f}$). Also, $\sigma(f_{pi},\mathbf{f}_{pi'}) = 0$ for all pairs where i is not identical to i'.

3. The measures of X made under different conditions are parallel. It follows from the assumptions above that the measures have equal means, equal variances, and equal intercorrelations. The same is true for Y measures.

4. It follows, now, that

$$\sigma(X_{pi},Y_{pi'}) = \sigma(X_{\infty p},Y_{\infty p}),$$

hence this covariance is the same for all independent observations of X and Y.

For linked observations the covariance

$$\sigma(X_{pi},Y_{pi}) = \sigma(X_{\infty p},Y_{\infty p}) + \sigma(f_{pi},\mathbf{f}_{pi})$$

We assume that $\sigma(f_{pi},\mathbf{f}_{pi})$ is the same for all i, hence that the linked covariance is the same for all linked X, Y observations. The covariance of f with $\mathbf{f}$ may be large or small, depending on the extent to which the condition i influences the X and Y performances.

We shall simplify and compress notation in several convenient ways. We write $\rho_{XX'}$ for a reliability coefficient, and ρ_{XY} for a correlation. For emphasis, when linked observations are correlated or used to form a difference score we may refer to X_a and Y_a; their covariance and correlation we shall designate ${}^{\bullet}\sigma_{XY}$ and ${}^{\bullet}\rho_{XY}$. We shall similarly write X_a and Y_b (or the like) for a pair of independently observed scores, and for the covariance and correlation shall write ${}^{\circ}\sigma_{XY}$ and ${}^{\circ}\rho_{XY}$.

Ordinarily, $|{}^{\bullet}\rho_{XY}| > |{}^{\circ}\rho_{XY}|$. It is possible that ${}^{\bullet}\rho_{XY} < {}^{\circ}\rho_{XY}$, where X_a and Y_a have some complementary relation. For instance, if rate of reading and comprehension are measured on the same selections simultaneously, one score is likely to rise at the expense of the other and $\sigma(f_{pi},\mathbf{f}_{pi})$ will be negative.

5. It is assumed that the population parameters are known. All parameters considered must be for the same population.

A regression estimate of a true score is usually improved if the data permit one to derive an equation for a subpopulation—for example, for ninth-grade boys in a certain school rather than for the national ninth-grade population. The investigator using actual data will often have made no reliability study on his sample. If he uses a reliability coefficient or a value of ρ_{XY} from the published study he must adjust it to take into account the variance of his sample.

Reliability of a Difference Score

As Stanley (1967) pointed out, linkage must be taken into account in defining a reliability coefficient for differences. Classical theory, ignoring linkage, defines a reliability

$$\rho_{DD'} = \frac{\sigma_{DD'}}{\sigma^2_D} = \frac{\sigma^2_X\rho_{XX'} + \sigma^2_Y\rho_{YY'} - 2\sigma_X\sigma_Y\rho_{XY}}{\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y\rho_{XY}}, \quad [2]$$

where $D = Y - X$. The reliability coefficients $\rho_{XX'}$ and $\rho_{YY'}$ are correlations of independent observations. Likewise, because of the independence assumption, classical theory can only interpret ρ_{XY} as what we have called ${}^{\circ}\rho_{XY}$.

The reliability of a difference score is defined as the correlation of the score with an independently observed difference. Unlike the classical papers, Stanley distinguishes $\rho_{(Y_a-X_a)(Y_b-X_b)}$ from $\rho_{(Y_a-X_b)(Y_b-X_a)}$. These are distinct kinds of reliability, one for a difference of linked X and Y and one for a difference of independent X and Y.

If the observed difference is $D_{ab} = Y_a - X_b$ (i.e., if X and Y are experimentally independent), the covariance with an independently observed difference D_{cd} is

$$\sigma^2_X\rho_{XX'} + \sigma^2_Y\rho_{YY'} - 2\sigma_X\sigma_Y{}^{\circ}\rho_{XY}. \quad [3]$$

Even if $D_{aa} = Y_a - X_a$ (i.e., X and Y are experimentally linked), the covariance with a similar but independently observed linked difference $D_{bb} = Y_b - X_b$ is the same. Hence Formula 3 is the appropriate numerator for Formula 2 in both cases.

The variance of D for the independent case is

$$\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y{}^{\circ}\rho_{XY}. \quad [4]$$

For the linked case, however, $D_{aa} = Y_a - X_a$ and the variance equals

$$\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y{}^{\bullet}\rho_{XY}. \quad [5]$$

Hence for the linked case the reliability coefficient $\rho_{(Y_a-X_a)(Y_b-X_b)}$ is obtained by dividing Formula 3 by Formula 5:

$${}^{\bullet}\rho_{DD'} = \frac{\sigma^2_X\rho_{XX'} + \sigma^2_Y\rho_{YY'} - 2\sigma_X\sigma_Y{}^{\circ}\rho_{XY}}{\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y{}^{\bullet}\rho_{XY}} \quad [6]$$

whereas for the independent case the reliability coefficient $\rho_{(Y_a-X_b)(Y_c-X_d)}$ is Formula 3 divided by Formula 4:

$${}^{\circ}\rho_{DD'} = \frac{\sigma^2_X\rho_{XX'} + \sigma^2_Y\rho_{YY'} - 2\sigma_X\sigma_Y{}^{\circ}\rho_{XY}}{\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y{}^{\circ}\rho_{XY}} \quad [7]$$

Since ${}^{\circ}\rho_{XY} < {}^{\bullet}\rho_{XY}$ in most instances, ${}^{\circ}\rho_{DD'}$ for the independent case will most likely be smaller than ${}^{\bullet}\rho_{DD'}$ for the linked case. Both reliability coefficients are meaningful. They describe the correlation between differences observed according to different experimental designs. Distinctions like that between Equations 6 and 7 have to be made in considering the reliability of any composite, weighted or unweighted.

In the discussion that follows formulas are written in terms of the linked case; the substitution to fit the fully independent case will be obvious.

Estimators of True Change

Primitive Formulas

Among the possible estimators of D_{aa} are three simple formulas.

Raw gain. The simplest formula is:

$$D = Y - X \quad [8]\,[1]$$

Correction by simple regression for error in X. If $\hat{X}_\infty$ is a regression estimate of X_∞ from X,

$$\hat{D}_\infty = Y - \hat{X}_\infty = Y - \rho_{XX'}X + \text{constant} \quad [9]\,[2]$$

In this and all other equations through Equation 20, the constant is one that makes the mean (over persons) of the estimates equal to the mean raw gain. The foregoing estimate has occasionally been suggested (e.g., Trimble & Cronbach, 1943), but it has seen little use. Closely related concepts appear in Lord's comments on analysis of covariance (1960) and in a series of Swedish papers on the effect of schooling on intelligence (see Härnqvist, 1968).

Correction by simple regression for error in X and in Y. To take a further step, let $\hat{Y}_\infty$ be an estimate of Y_∞ from Y. Then

$$\hat{D}_\infty = \hat{Y}_\infty - \hat{X}_\infty = \rho_{YY'}Y - \rho_{XX'}X + \text{constant} \quad [10]\,[3]$$

This procedure does not take the X, Y correlation into account.

The Lord Procedure

Lord pointed out that unless $\rho_{X_\infty Y_\infty} = 0$, both X and Y yield information about X_∞. A multiple regression procedure can be used to obtain

$$\hat{X}_\infty = \rho_{XX'}X + \beta_{X_\infty(Y\cdot X)}(Y\cdot X) + (1-\rho_{XX'})\bar{X} \quad [11]$$

Here $Y\cdot X$ is a partial variate, the deviation of Y from the value predicted by the regression of Y on X in the population to which the other parameters in Equation 11 apply.

For the linked case we have

$$Y_a\cdot X_a = Y_a - \frac{{}^{\bullet}\sigma_{XY}}{\sigma^2_X}(X_a - \bar{X}) - \bar{Y} \quad [12]$$

We know that

$$\beta_{X_\infty(Y_a\cdot X_a)} = \frac{\sigma_{X_\infty(Y_a\cdot X_a)}}{\sigma^2_{(Y_a\cdot X_a)}} \quad [13]$$

Now $\sigma_{X_\infty Y} = \sigma_{X_a Y_b} = {}^{\circ}\sigma_{XY}$ (not ${}^{\bullet}\sigma_{XY}$). Using Equations 12 and 13, the numerator of β becomes

$$\sigma_{X_\infty(Y_a\cdot X_a)} = \sigma_X\sigma_Y({}^{\circ}\rho_{XY} - {}^{\bullet}\rho_{XY}\rho_{YY'}) \quad [14]$$

and the denominator becomes

$$\sigma^2_{(Y_a\cdot X_a)} = \sigma^2_Y(1 - {}^{\bullet}\rho^2_{XY}) \quad [15]$$

Substituting in the regression equation, we

arrive at

$$\hat{X}_\infty = \frac{\rho_{XX'} - {}^{\bullet}\rho_{XY}{}^{\circ}\rho_{XY}}{1 - {}^{\bullet}\rho^2_{XY}} X + \frac{\sigma_X({}^{\circ}\rho_{XY} - {}^{\bullet}\rho_{XY}\rho_{XX'})}{\sigma_Y(1 - {}^{\bullet}\rho^2_{XY})} Y + \text{constant} \quad [16]$$

Then

$$\hat{D}_\infty = \hat{Y}_\infty - \hat{X}_\infty, \quad [17]\,[4]$$

where the estimates on the right-hand side come from Equation 16 and its analog for $\hat{Y}_\infty$. Expanding the expression we have

$$\hat{D}_\infty = \frac{1}{1 - {}^{\bullet}\rho^2_{XY}} \left[\frac{Y}{\sigma_Y} (\sigma_Y \rho_{YY'} - \sigma_Y {}^{\bullet}\rho_{XY} {}^{\circ}\rho_{XY} + \sigma_X {}^{\bullet}\rho_{XY}\rho_{XX'} - \sigma_X {}^{\circ}\rho_{XY}) - \frac{X}{\sigma_X} (\sigma_X \rho_{XX'} - \sigma_X {}^{\bullet}\rho_{XY} {}^{\circ}\rho_{XY} + \sigma_Y {}^{\bullet}\rho_{XY}\rho_{YY'} - \sigma_Y {}^{\circ}\rho_{XY}) \right] + \text{constant} \quad [17a]\,[4]$$

For independent observations, one substitutes ${}^{\circ}\rho_{XY}$ wherever ${}^{\bullet}\rho_{XY}$ appears in Equation 17a; this is Lord's estimator of D_∞. In an experiment, all calculations must be made within a treatment group.

Estimator 4 is as good or better than any of those listed ahead of it, giving a smaller mean square of $(\hat{D}_\infty - D_\infty)$ and a larger correlation between estimate $\hat{D}_\infty$ and true score D_∞. Ordinarily Estimator 3 is better than Estimator 2 and both are better than Estimator 1. Our main point in presenting Estimators 2 and 3 is to show the Lord estimator as an elaborated form of a more conventional estimator. This lays a base for the further refinement.

It may seem anachronistic in Formulas 11 and 16 to use a posttest score to "predict" a pretest score. But the logic is clear. Within a treatment, persons higher on the posttest than others having the same observed pretest score tend to be those for whom the true pretest score is higher than the observed score. The Y receives at least nominal weight in the regression equation when ρ_{XY} is not zero or one, and $\rho_{XX'} < 1.00$. The weight given to Y increases with larger ρ_{XY} and smaller $\rho_{XX'}$.

Taking group membership into account. Previous papers have implicitly assumed that all persons come from a single population, but often there are several distinct subgroups. These groups may be distinguished by demographic characteristics or by past experience, or they may be groups receiving different treatments between the X and Y observations. One could pool all groups and determine a single within-group value for each parameter in the equations above, but parameters calculated within subgroups will give a better estimate of X_∞, Y_∞, and D_∞. There is a limit to how far subdivision of samples can profitably be carried, however.

Correlations and Regression Slopes for D_∞

Sometimes an investigator wishes to know the correlation of D_∞ with another variable, say Q. He should note, then, that the correlation of $\hat{D}_\infty$ with Q is *not* a sound estimate of the correlation of D_∞ with Q. The correlation should be determined directly from the covariance of D_∞ with the second variable of interest. This covariance takes a form such as

$$\sigma_{D_\infty Q} = \sigma_{Y_\infty Q} - \sigma_{X_\infty Q} = \sigma_{YQ} - \sigma_{XQ}.$$

To get the correlation coefficient one divides by σ_Q and σ_{D_∞} (not $\sigma_{\hat{D}_\infty}$). Since D_∞ equals $Y_\infty - X_\infty$, its variance is given by Equation 3. All parameters must be those for the same group.

The investigator is often interested in the regression of D_∞ (or Y_∞) on another variable. The slope of the regression of D_∞ on (e.g.) X_∞ is $\sigma_{D_\infty X_\infty}/\sigma^2_{X_\infty}$.

Attention must be paid to linkages. Let us write Q_e to indicate that the observation of Q is independent of X_a, Y_a, and Y_b.

$$\sigma_{D_\infty Q_e} = \sigma_{Y_b Q_e} - \sigma_{X_a Q_e} = \sigma_{Y_b Q_e} - \sigma_{X_a Q_e}. \quad [18]$$

This cannot be determined from data where Q is linked to X or Y.

A variance-covariance algorithm. A simple computational routine can be suggested for problems of this character. One may form a variance-covariance matrix of observed values as in Figure 1. Linked and independent covariances are carefully distinguished. The matrix may be augmented as shown, adding rows and copying forward covariances. Reliability information is taken into account at certain points. Columns may be added to the matrix also, in a symmetric manner. Thus all entries in the X_∞ row can be copied into the X_∞ column. The full extension carried out in this way gives a square matrix, from which such values as $\sigma_{D_\infty Q_\infty}$ and $\sigma^2_{D_\infty}$ can be read out.

			X_a	Y_a	Q_e	Y_b	Q_e	X_∞
	Row 1	X_a	σ^2_X	${}^{\bullet}\sigma_{XY}$	${}^{\bullet}\sigma_{XQ}$	${}^{\circ}\sigma_{XY}$	${}^{\circ}\sigma_{XQ}$	$\sigma^2_X\rho_{XX'}$
	Row 2	Y_a	${}^{\bullet}\sigma_{XY}$	σ^2_Y	${}^{\bullet}\sigma_{YQ}$	$\sigma^2_Y\rho_{YY'}$	${}^{\circ}\sigma_{YQ}$	${}^{\circ}\sigma_{XY}$
	Row 3	Q_e	${}^{\bullet}\sigma_{XQ}$	${}^{\bullet}\sigma_{YQ}$	σ^2_Q	${}^{\circ}\sigma_{YQ}$	$\sigma^2_Q\rho_{QQ'}$	
	Row 4	Y_b	${}^{\circ}\sigma_{XY}$	$\sigma^2_Y\rho_{YY'}$	${}^{\circ}\sigma_{YQ}$	σ^2_Y	${}^{\circ}\sigma_{YQ}$	
	Row 5	Q_e	${}^{\circ}\sigma_{XQ}$	${}^{\circ}\sigma_{YQ}$	$\sigma^2_Q\rho_{QQ'}$	${}^{\circ}\sigma_{YQ}$	σ^2_Q	
Copy independent covariances corresponding to Row 1	Row 6	X_∞	$\sigma^2_X\rho_{XX'}$	${}^{\circ}\sigma_{XY}$	${}^{\circ}\sigma_{XQ}$	${}^{\circ}\sigma_{XY}$	${}^{\circ}\sigma_{XQ}$	$\sigma^2_X\rho_{XX'}$
Copy independent covariances corresponding to Row 4	Row 7	Y_∞	${}^{\circ}\sigma_{XY}$	$\sigma^2_Y\rho_{YY'}$	${}^{\circ}\sigma_{YQ}$	$\sigma^2_Y\rho_{YY'}$	${}^{\circ}\sigma_{YQ}$	${}^{\circ}\sigma_{XY}$
Fill in by subtracting Row 1 from Row 2	Row 8	D_∞						
Fill in by subtracting Row 1 from Row 4	Row 9	D_b						
Fill in by subtracting Row 6 from Row 7	Row 10	D_∞						
Copy independent covariances from Row 5	Row 11	Q_e						

FIG. 1. Algorithm for constructing covariances and variances. (Covariances for linked observations are identified by the symbol ${}^{\bullet}\sigma$, and those for independent observations by ${}^{\circ}\sigma$. The broken line separates the original data from entries added later.)

A Better Estimate of True Change

Lord's formula uses only X and Y data, but we shall bring in two further categories of variables, W and Z. The W and X are Time-1 measures, but need not be simultaneous. Although we write W without vector notation, there are, in principle, any number of W variables that can be used singly or in combination. Our statements apply to any W or any weighted composite of the W.

A W might be any score describing the subject as he was prior to the treatment under study or W might be an index based on his life history. The scores Y and Z are posttreatment measures—again, not necessarily simultaneous. The Y might, for example, be a measure of performance at the end of training, and Z a retention test a month later. Where we are examining a difference score rather than a change score, no distinction between W and Z variables is needed.

The steps taken in going from Estimate 2 to Estimate 4 above can be extended to make use of W and Z information so as to reach an even better estimate of D_∞.

The complete estimator. If W is univariate and there is no Z information,

$$\hat{X}_\infty = \rho_{XX'}X + \frac{\sigma_{X_\infty(Y \cdot X)}}{\sigma^2_{(Y \cdot X)}}(Y \cdot X) + \frac{\sigma_{X_\infty(W \cdot X,Y)}}{\sigma^2_{(W \cdot X,Y)}}(W \cdot X,Y) + \text{constant} \quad [19]$$

Here $W \cdot X, Y$ is a partial variate, the deviation of W from the value predicted by the regression of W on X and Y.

If the W information is multivariate, a whole series of partial variates enters. The order of partialling is arbitrary; one might write terms

for X; $W \cdot X$; $Y \cdot W$; X; etc. Where there is Z information, one adds further terms, again employing partial variates; W, as well as X and Y, is partialled out. The estimation procedures must take into account any linkage between X, Y, W, and Z variables as in Equation 16.

A similar equation is written for $\hat{Y}_\infty$. Finally, one comes to an equation of the form

$$\hat{D}_\infty = \beta_1 X + \beta_2 Y + \beta_3 W + \beta_4 Z + \text{constant} \quad [20]\,[5]$$

Here $\beta_3 W$ stands for a string of several terms of the form $\beta_i W_i$ if there are many W (likewise for $\beta_4 Z$). This estimator is superior to Formula 4, provided that the sample size is large enough to justify assigning a large number of weights. Where sample size is insufficient, the number of predictors must be held down, most likely by employing the first one or more principal components of the W set as predictors (likewise for Z).

When the problem becomes complicated, it is better to use efficient computing routines than to write out elaborate formulas. The within-treatment covariance matrix for X, Y, W, Z is written. Additional rows and columns for X_∞ and Y_∞ are formed as in Figure 1, with care to enter independent or linked covariances as required. The X_∞ column is subtracted from the Y_∞ column to form the D_∞ column. When the symmetric matrix is complete, one applies a multiple-regression program, treating entries in the D_∞ column as "test-criterion covariances" and the appropriate X, Y, W, and Z as predictors. If the observed scores have X and Y linked, for example, then X_a and Y_a are used as predictors, and covariances for Y_b are ignored.

Demographic information and information about experience can and should be considered as W variables. With a variable such as sex, one has a choice of entering it directly as a variable coded 1 and 0, or of performing a separate regression analysis for each sex. Both procedures regress the person's score toward the mean for his own sex rather than toward the mean for all cases. The second procedure allows for the possibility that the regression surface for males differs from that for females. Separate within-subgroup regressions would seemingly be preferred when samples are truly large.

The difficulty is that the argument can be repeated for every other noncontinuous variable and for all combinations of them. Indeed, it applies to continuous variables also; for example, a regression surface for more anxious children may differ from that for the less anxious. These remarks amount to entertaining the possibility of nonlinear relationships. While this possibility is real enough, one can rarely get usable estimates of nonlinear functions from samples of practical size (Burket, 1964; Goldberg, 1969). Hence, after one has divided the sample into a few salient subgroups, each having a suitably large size, the dummy-variable technique seems to be the only feasible way to handle a variety of nonquantitative information.

Residualized Gains

Developments parallel to those above lead to so-called "residual gains" or "basefree measures of change."

Alternative Estimators

The raw residual-gain score is defined by

$$D \cdot X = Y - E(Y)\,|\,X = Y - \bar{Y} - \beta_{Y \cdot X}(X - \bar{X}) \quad [21]\,[1']$$

We designate this Formula 1′ to emphasize that it is comparable to Formula 1, the raw gain. If Y_a and X_a define the difference score, $\beta_{Y \cdot X}$ equals $^{\bullet}\rho_{XY}\sigma_Y/\sigma_X$. If Y_b rather than Y_a is used, $^{\circ}\rho_{XY}$ replaces $^{\bullet}\rho_{XY}$. The traditional definition given by Equation 21 is ambiguous; the residual $(Y_a - X_a) \cdot X_a$ is conceptually different from the residual $(Y_b - X_a) \cdot X_a$.

Residualizing removes from the posttest score, and hence from the gain, the portion that could have been predicted linearly from pretest status. One cannot argue that the residualized score is a "corrected" measure of gain, since in most studies the portion discarded includes some genuine and important change in the person. The residualized score is primarily a way of singling out individuals who changed more (or less) than expected.

True residual gain could be defined either as the expected value, over many observations on the same person, of $D \cdot X$ (defined in either of the two possible ways), or as the partial variate $D_\infty \cdot X_\infty$, the part of the true gain not predictable from true pretest status. $D_\infty \cdot X_\infty$ is more likely to be of interest. Certainly if we intend to pick out superior learners or persons whose self-concept falls far below the self-ideal, we would like to base the discrimination on a true-score disparity. Likewise, if we have correlational questions—for example, Does anxiety predict overachievement?—the variable seems better specified by $D_\infty \cdot X_\infty$. Hence we consider only the definition:

$$D_\infty \cdot X_\infty = D_\infty - \beta_{D_\infty X_\infty} X_\infty + \text{constant} = Y_\infty - \frac{\sigma_{Y_\infty} {}^{\circ}\rho_{X_\infty Y_\infty}}{\sigma_{X_\infty}} X_\infty + \text{constant} \quad [22]$$

No ambiguity arising from possible linkage of the observed X and Y enters this definition, though linkage must be considered in any estimation procedure.

Successive estimators of true residual change can be constructed on the same principles as before, but it will suffice to move directly to the estimator formed by the multiple-regression principle in the manner of Lord:

$$\widehat{D_\infty \cdot X_\infty} = \frac{\rho_{XX} \cdot \rho_{YY'} - {}^{\circ}\rho^2_{XY}}{(\rho_{XX} \cdot {}^{\circ}\rho_{XY})(1 - {}^{\bullet}\rho^2_{XY})} \times \left(Y_a - \frac{\sigma_Y {}^{\bullet}r_{XY}}{\sigma_X} X_a \right) + \text{constant} \quad [23]\,[4']$$

This and other constants in this section are defined to make the mean of estimated residual gain equal zero. This estimate turns out to be proportional to the raw residual gain. (If the estimate is made from independent Y and X, $^{\circ}\rho_{XY}$ replaces $^{\bullet}\rho_{XY}$).

If there is W or Z information, a still better estimator is

$$\widehat{D_\infty \cdot X_\infty} = \beta'_1 X + \beta'_2 Y + \beta'_3 W + \beta'_4 Z + \text{constant} \quad [24]\,[5']$$

Only in rare cases will this be proportional to the observed $D \cdot X$.

The computational algorithm used before can be extended to obtain the desired weights for Formula 4′ or 5′. Suppose we have filled out the square matrix for X, Y, X_∞, Y_∞, D_∞, ... Then we can form the covariance of any variable with $D_\infty \cdot X_\infty$ very simply. For example,

$$\sigma_{(D_\infty \cdot X_\infty)Q} = \sigma_{D_\infty Q} - \frac{\sigma_{D_\infty X_\infty}}{\sigma^2_{X_\infty}} \sigma_{X_\infty Q} \quad [25]$$

Hence we may multiply every entry in the X_∞ column of the matrix by $\sigma_{D_\infty X_\infty}/\sigma^2_{X_\infty}$, entering this in a column to one side. Subtracting the entry in this side column from the entry in the corresponding row of the D_∞ column gives a covariance to be entered in a $D_\infty \cdot X_\infty$ column. This column is now taken as a set of covariances of predictors with the criterion, and a multiple-regression program is applied.

The Tucker-Damarin-Messick Proposals

This analysis puts us in a position to review and clarify the rather puzzling paper entitled "A base-free measure of change" (Tucker, Damarin, & Messick, 1966; hereafter, referred to as TDM). They start much as we do by noting that the psychometrics of a change score applies to all kinds of difference scores. They suggest, as Lord did, that one should be most interested in the true difference score. They propose to divide this difference into two components, "one entirely dependent on the true score of the first or base-line test" and one "entirely independent of it." That is, they are interested in a true predicted gain and a true residual gain. As their abstract says, "equations for estimating both components are given." Since we shall recommend against use of their equations, we shall not go into details of their argument.

It might appear that TDM are concerned with estimating $E(D_\infty)\,|\,X_\infty$ and $D_\infty - E(D_\infty)\,|\,X_\infty (= D_\infty \cdot X_\infty)$. The former, of course, is a linear function of X_∞. TDM arrive at an equation (their Equation 26) that, in a form consistent with the present paper, is

$$\widehat{D_\infty \cdot X_\infty} = \left\{ \begin{matrix} Y_a - \dfrac{{}^{\bullet}\rho_{XY}\sigma_Y}{\rho_{XX}\cdot\sigma_X} X_a \\ \text{or} \\ Y_b - \dfrac{{}^{\circ}\rho_{XY}\sigma_Y}{\rho_{XX}\cdot\sigma_X} X_a \end{matrix} \right\} + \text{constant} \quad [26]$$

It is rather startling to find that this agrees with none of our formulas. It differs from Formula 1′ in that X is replaced by $X/\rho_{XX'}$. It differs by a further constant of proportionality from Formula 4′. The marked departure from Formula 4′ is made the more puzzling by the favorable references of TDM to the Lord and McNemar papers and by their recommendation of Formula 4 for the gain score itself.

Personal communication with the authors verified that a failure of communication had occurred[4]. As readers, we had given too little weight to one key phrase: The measures are "primarily intended for correlational work." That is to say, TDM have no intention of interpreting "basefree" scores for individuals. Such scores are intended only as an intermediate step toward correlations. TDM offer an estimator that does not give the best least-squares estimate of individual "basefree" scores because they seek instead estimates that correlate zero with X_∞.

Their intention is to determine correlations so as to learn what kinds of person show gains larger than would be predicted from the true pretest score. Correlating the estimated true residual gain from Equation 23 or 24 with another variable does not give the correlation TDM desire (unless $\rho_{XX'}$ is 1.00). To understand this, consider for a moment the simple correlation of Y with Q. If we do not know Y scores but do know ρ_{XY}, we might estimate Y from X scores by the usual regression equation. Then $\rho\hat{Y}Q$ will not be a good approximation to ρ_{YQ}; it will actually equal ρ_{XQ}. In general, one who wants to interpret correlations, covariances, or regression slopes ought not to work from estimated scores. TDM intended to recommend that in such a line of research one calculate special-purpose scores by Formula 26 and then determine correlations. This appears to be unsound. TDM desire to obtain correlations with various Q of γ and ζ, which in their notation are the residual and predictable portions of true gain, respectively. But the TDM formulas generate fallible values g and w; g equals γ plus an error. Obviously $\rho_{gQ} < \rho_{\gamma Q}$ and $\rho_{wQ} < \rho_{\xi Q}$. As this is not explained by TDM, their paper is likely to mislead the reader. The confusion is reflected, and to some degree intensified, when Traub (1967) and Glass (1968) comment on the TDM formula.

While one might adapt the TDM statements to get $\rho_{\gamma Q}$, $\rho_{\xi Q}$, etc., this is unnecessary. A straightforward manipulation of the matrix of observed covariances for X, Y, and Q (along with $\rho_{XX'}$ and $\rho_{YY'}$) yields the covariance of Q with $D_\infty \cdot X_\infty$ (i.e., with γ). The $\sigma^2_{D_\infty \cdot X_\infty} (= \sigma^2_\gamma)$ needed to reach a correlation is simply the covariance of D_∞ with $D_\infty \cdot X_\infty$ that we have already obtained. To get covariances for $D_\infty - D_\infty \cdot X_\infty$ (i.e., for ζ), one need only subtract column $D_\infty \cdot X_\infty$ of the covariance matrix from column D_∞. And $\sigma^2_{D_\infty} = \sigma^2_\gamma + \sigma^2_\zeta$.[5]

[4] L. R. Tucker, F. Damarin, and S. Messick, personal communication, September 1968 and April 1969.

[5] This matrix-extension procedure is entirely consistent with the TDM rationale. In fact, TDM tell us that their formula was arrived at by analytic treatment of this extended matrix.

A Multivariate Conception

The older statement of the problem as "the measurement of gain" or of "residual gain" implies a special affinity between X and Y—they are seen as "the same variable" in some sense. But change is multivariate in nature.

Even when X and Y are determined by the same operation, they often do not represent the same psychological processes (Lord, 1958). At different stages of practice or development different processes contribute to performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes. This does not rule out purely empirical studies of changes in the operationally defined variable. To assess such changes, even when one cannot describe them qualitatively, may be practically important. One must be careful not to fall into the trap of assuming that the changes are in a particular psychological attribute.

We may illustrate by referring to Fleishman's well-known studies of psychomotor scores at successive stages of practice (see Fleishman, 1966). On the first few trials, scores tend to correlate with cognitive measures; the usual speculation is that the high-cognitive subjects gain fastest because they most rapidly comprehend the instructions, display, strategy, etc. In a second stage, certain pretests of psychomotor ability correlate highest with scores, and the correlation of scores with cognitive pretests becomes rather small. One could easily conclude that cognitive ability is unimportant to learning in this second stage. But the drop in correlation (assuming no great rise in SD) demonstrates something more striking: that cognitive ability is *negatively correlated with change* from the first to second stage. And one can suggest a good reason. If bright persons

catch on fastest, by some trial t they have completed what can be done intellectually with the task. On trials $t + 1$ and later, their cognitive abilities produce no further gains. But by our hypothesis the dull have not comprehended fully by trial t, and therefore they have cognitive work left to do on trials $t + 1, t + 2, \ldots$. During this stage, any gain in performance that comes from improved comprehension will be a gain by those low in cognitive aptitudes, hence those aptitudes have a negative correlation with gains. The Fleishman studies have not been processed in terms of gain scores, but it has been commonly said that they indicate cognitive abilities to be "important during the early stages of practice on psychomotor tasks," or the like. It seems more accurate to say that cognitive abilities make their contribution at different times for different persons and that gains at any one time are due to different processes for different persons.

Something similar is to be said about the relation of mental age (MA) at one age to subsequent gains in MA or achievement. A positive relation would result insofar as the high scorers understand new material better, or are more efficient learners. But there would also be a negative relation, insofar as the high scorers are those who have already mastered some highly valuable technique (e.g., mediation) that the low scorers have yet to master. As they restructure their behavior, the persons with low scores at the start of the period may make large gains—gains the high scorers had previously made. Positive and negative elements are probably both present, which should make us much less surprised than we have been by reports of near-zero correlation of MA (in a constant-age group) with subsequent gain in MA (Anderson, 1939; Bloom, 1964, p. 26 ff., 62 ff.).

We reduce emphasis on the special role of X as precursor of Y, and regard the whole WX set as a vector describing the person's initial status. Then one may ask how Y varies as a function of the Time-1 data. To single out for intensive study persons who do better (or worse) than predicted, for example, it is wise to define expected outcome as the forecast of Y on the basis of all Time-1 information. Instead of $D_\infty \cdot X_\infty$ $(= Y_\infty \cdot X_\infty)$ one would estimate $D_\infty \cdot X_\infty, W_\infty$ $(= Y_\infty \cdot X_\infty, W_\infty)$. The machinery suggested above for partialling out X_∞ would be used, but extended by partialling the W_∞ also out of D_∞. The entire set of W and X measures constitute the "base." There are optional targets for investigation: D_∞; $D_\infty \cdot X_\infty$; $D_\infty \cdot X_\infty, W_\infty$; etc.

Learning or growth is multidimensional; many measures could be taken at each point in time. To select one particular Y as somehow integrating a variety of subcriteria is to sacrifice information and possible insight. A person's change is better described by a vector of true scores $W_\infty, X_\infty; Y_\infty, Z_\infty$. Each of these can be estimated by the methods used in Equations 19 and 20. One who wants to examine predicted and residual change will estimate $\hat{Y}_\infty$ from $\hat{W}_\infty$ and $\hat{X}_\infty$, and also from all variables together. He will obtain these scores:

$\hat{Y}_\infty \mid WXYZ$	(estimated true final status)
$\hat{Y}_\infty \mid \hat{W}_\infty \hat{X}_\infty$	(predicted true final status)
Difference	(estimate of unpredicted true residual)

The estimates of W_∞ and X_∞ come from W, X, Y, and Z. There would be estimates like those for Y for each Z or for orthogonal components of the Y_∞, Z_∞ space.

We can rearrange the vectors W, X and Y, Z in a great variety of ways. *Which target to choose can be decided only in the light of the purposes of the study.*

Purposes of Estimating Gains or Differences

Just why gains or differences are thought to be worth estimating can perhaps be inferred from the studies where estimates of some sort have been made in the past. The following aims may be noted:

1. To provide a dependent variable in an experiment on instruction, persuasion, or some other attempt to change behavior or beliefs.
2. To provide a measure of growth rate or learning rate that is to be predicted, as a way of answering the question, What kinds of persons grow (learn) fastest? Here, the change measure is a criterion variable in a correlational study.
3. To provide an indicator of deviant development, as a basis for identifying individuals to be given special treatment or to be studied clinically.
4. To provide an indicator of a construct that is thought to have significance in a certain

theoretical network. The indicator may be used as an independent variable, covariate, dependent variable, etc. An example is the interference score needed on the Stroop Color Word Test to represent the decline in reading rate when color names are printed in colors incongruous with the names.

Much of the confusion in the literature arises from a failure to distinguish these purposes and to match distinct methodological recommendations to them.

Gains as a Consequence of Treatments

There appears to be no need to use measures of change as dependent variables and no virtue in using them. If one is testing the null hypothesis that two treatments have the same effect, the essential question is whether posttest Y_∞ scores vary from group to group. Assuming that errors of measurement of Y are random, Y is an entirely suitable dependent variable.

The randomized experiment. Suppose that cases are assigned to treatments in a random or stratified-random manner. The X scores will vary within groups. An analysis of covariance to take this variation into account is advantageous so long as ρ_{XY} is large. (If $\rho < 0.4$, blocking is probably to be preferred, according to Elashoff, 1969.) The usual adjustment estimates the Y scores expected under the null hypothesis and then expresses each observed Y as a deviation from the estimate. Ordinarily it is desirable to base the adjustment, not on X, but on whatever linear combination of X and W best predicts Y within groups.

Where within-treatment regressions are linear but significantly different in slope, the difference between effects of treatments depends on the level of X. The "main effect" is not interpretable. The most meaningful report consists of regression functions for Y on the X_∞, W_∞ space, computed with the aid of the covariance matrix for true scores within each group in turn.

Nowhere in this section have we made use of a change score. We consider it likely that change will vary systematically with X_∞. Where this is the case, the essential result is a regression function, not a mean gain. The adjusted Y score of the significance test is a sort of residual gain, but the procedure does not involve calculating residual gain scores for individuals.

Comparison of treatment groups not formed at random. When treatments are applied to groups differentiated by a nonrandom process, the X_∞ distributions within the subpopulations represented by the groups are generally not the same. Consequently, the same observed X score implies a different level of true pretest ability, depending on the group.

If analysis of covariance is to be made, it is advisable to regress the covariate toward the mean of the treatment group before entering it in the analysis (Lord, 1960). If there is W information as well as X, it also contributes to the estimate. So does Y and Z information. Here is a paradox: A proposal to use the posttest score to estimate the pretest true score which will then be used to adjust posttest scores! The crucial point is that the estimator of the covariate is determined from within-group data. Since the estimate of any linear function of X_∞ and W_∞ has the same within-group mean as that function of X and W, the procedure does not introduce bias nor does it reduce any effect truly attributable to the treatment.

Application of analysis of covariance to studies where initial assignment was nonrandom, which was widely recommended 10 years ago, is now in bad repute. Even the elaborate technique just suggested is no more than a palliative. If the treatment groups differed systematically at the start of the experiment with respect to any relevant characteristic other than the covariate, even a perfect measure of the covariate cannot remove the confounding. To quote Lord (1967), "there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups [p. 305]." And Meehl (1970) calls such corrections "inherently fallacious [in press]."

The findings of the study can be usefully summarized by calculating within-group regression functions relating Y_∞ to X_∞, W_∞, using the covariance matrix for true scores. What cannot be done is to "compare treatment effects."

One-group designs. A third kind of experiment is the simple one-group study where one wishes to learn whether a treatment produces significant change, or to describe the magnitude of the effect. An estimate of true gain might appear to be pertinent. But it is not. For if one were to estimate D_∞ for each individual, and average, he would arrive back at the sample mean of *observed* gain. A significance test need only ask whether μ_Y is reliably different from μ_X. The difference in sample means for X and Y is the best available estimate of the mean D_∞.

Criteria in Correlational Studies.

Correlational studies are often intended to investigate a question such as this: Among persons with a given pretest score, what attributes distinguish those who profit most from the treatment? This may seem to ask about ρ_{DW} or $\rho_{D_\infty W_\infty}$, or perhaps about $\rho_{(D \cdot X)W}$ or $\rho_{(D_\infty \cdot X_\infty)W}$. It is more straightforward to ask about the regression of Y on X_∞ and W_∞, the corresponding correlation for Y or Y_∞, or related partial correlations. It appears that nothing is gained by referring to change measures in this context. The relationships of true scores can be investigated without estimating true scores for individuals.

Selecting Individuals on the Basis of Gain or Difference Scores.

Many who calculate difference scores are interested in making decisions about individuals—identifying underachievers for clinical attention or fast learners for special opportunities, for example. One can scarcely defend selecting such individuals on a raw-gain or raw-difference score, especially as these scores tend to show a spurious advantage for persons low on X. Selecting cases whose estimated Y_∞ is higher than that of others with similar $\hat{X}_\infty$ and $\hat{W}_\infty$ seems more sensible. To do this, regression equations should be called into play. That is, one selects persons for whom $\hat{Y}_\infty | WXYZ$ is much larger (or smaller) than $\hat{Y}_\infty | \hat{W}_\infty \hat{X}_\infty$.

The persons with positive deviations are those who did better than predicted. This means either that they started with some valuable attribute the W and X variables did not encompass, that their pretest true scores are underestimated or their posttest scores are overestimated, or that their success on Y was an accidental effect arising from some tactic casually adopted during learning or some sequence of lucky trials. It is very hard to dispose of the hypothesis that these unexpected gains were fortuitous.

Here, the focus of attention is on an estimated residual gain: not $D \cdot X$, not $D_\infty \cdot X_\infty$, but $\hat{Y}_\infty \cdot \hat{X}_\infty \hat{W}_\infty$ or, what is equivalent, $\hat{D}_\infty \cdot \hat{X}_\infty \hat{W}_\infty$. Where X alone is available as a predictor, the raw residual gain selects the same persons as Formula 23 does. But Formula 24 is to be preferred.

It is possible of course, given before-and-after scores on the same instrument, to estimate true gains of individuals and to identify those who did and did not gain. But to what purpose? This has no clear bearing on decisions about the future of these persons, and the decision rule for fresh cases is to be inferred from the regression surface.

Differences and Gain Scores as Constructs.

One of the most common uses of difference scores is to operationalize a concept: For example, self-satisfaction is sometimes defined as the difference between the rating of self and ideal-self on an esteem scale. One might likewise think of a gain score as reflecting "learning ability" on a certain task. Operational definitions will often take the form of linear combinations of operations.

But there is little a priori basis for pinning one's faith on $Y_\infty - X_\infty$ as distinct from the more general $Y_\infty - aX_\infty$. Just what weight to assign the "correcting" variable is an empirical question. To arbitrarily confine interest to D_∞ (which means that a is fixed at 1.00) is to rule out possible discoveries. This argues, then, for discovering what function of Y_∞ and X_∞ has the strongest relationships with variables that should connect with the construct.

The claim that an index has validity as a measure of some construct carries a considerable burden of proof. There is little reason to believe and much empirical reason to disbelieve the contention that some arbitrarily weighted function of two variables will properly define a construct. More often, the profitable strategy is to use the two variables separately in the analysis so as to allow for complex relationships.

One example of an "obvious" but questionable use of a subtractive correction is provided by a study in which skin conductance is a variable. At the start of the experiment a "base-

line" measure of the subject's galvanic skin response is taken. Then stress is applied and a second measure is taken. During a rest period the subject receives a drug or a placebo. Stress is again applied and a third measure taken. Call the measures, in order, W, X, and Y. Simple correction would use $Y - W$ as dependent variable and $X - W$ as covariate. We, however, would prefer to use X and W as separate covariates, with Y as dependent variable. This should give a more precise analysis when W is unreliable. (As suggested earlier, it would generally be still better to use $\hat{X}_w|WXY$ and $\hat{W}_w|WXY$ as covariates.)

Summary

Where true scores for individuals are desired, multiple regression procedures outlined herein make use of more information than do procedures hitherto advanced. There seems to be no occasion to estimate true gain scores. In the experiment where treatment groups are formed nonrandomly, estimates of true scores on the covariate can reduce the resulting bias.

Where individuals who have exceptionally high or low residual gains are to be identified, the raw residual gain serves as well as the alternate formulas hitherto advanced. To estimate the individual's true residual gain, however, a superior formula is available.

Where correlations and regression functions relating true gains or true residual gains to other variables are desired, a calculating routine is available that makes it unnecessary to estimate gain scores for individuals.

It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways.

REFERENCES

ANDERSON, J. E. The limitations of infant and preschool tests in the measurement of intelligence. *Journal of Psychology*, 1939, 8, 351–379.

BLOOM, B. S. *Stability and change in human characteristics*. New York: Wiley, 1964.

BURKET, G. R. A study of reduced rank models for multiple prediction. *Psychometric Monographs*, 1964, No. 12.

DUBOIS, P. H. *Multivariate correlational analysis*. New York: Harper, 1957.

ELASHOFF, J. D. Analysis of covariance: a delicate instrument. *American Educational Research Journal*, 1969, 6, 383–402.

FLEISHMAN, E. A. Human abilities and the acquisition of skill. In E. A. Bilodeau (Ed.), *Acquisition of skill*. New York: Academic Press, 1966.

GLASS, G. V. Response to Traub's "Note on the reliability of residual change scores." *Journal of Educational Measurement*, 1968, 5, 265–267.

GLESER, G. C., CRONBACH, L. J., & RAJARATNAM, N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 1965, 30, 395–418.

GOLDBERG, L. R. The search for configural relationships in personality assessment: the diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavorial Research*, 1969, 4, 523–536.

HÄRNQVIST, K. Relative changes in intelligence from 13 to 18. *Scandinavian Journal of Psychology*, 1968, 9, 50–82.

HARRIS, C. W. (Ed.) *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

LORD, F. M. The measurement of growth. *Educational and Psychological Measurement*, 1956, 16, 421–437. See also Errata, *ibid.*, 1957, 17, 452.

LORD, F. M. Further problems in the measurement of growth. *Educational and Psychological Measurement*, 1958, 18, 437–454.

LORD, F. M. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 1960, 55, 309–321.

LORD, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

LORD, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, 68, 304–305.

MCNEMAR, Q. On growth measurement. *Educational and Psychological Measurement*, 1958, 18, 47–55.

MEEHL, P. E. Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science*, IV. Minneapolis: University of Minnesota Press, 1970, in press.

STANLEY, J. C. General and special formulas for reliability of differences. *Journal of Educational Measurement*, 1967, 4, 249–252.

TRAUB, R. E. A note on the reliability of residual change scores. *Journal of Educational Measurement*, 1967, 4, 253–256.

TRIMBLE, H. C., & CRONBACH, L. J. A practical procedure for the rigorous interpretation of test-retest scores in terms of pupil growth. *Journal of Educational Research*, 1943, 35, 481–488.

TUCKER, L. R., DAMARIN, F., & MESSICK, S. A base-free measure of change. *Psychometrika*, 1966, 31, 457–473.

(Received June 24, 1969)