DOCUMENT RESUME

ED 050 175 24 TM 000 5

AUTHOR	O'Connor, Edward F., Jr.
TITLE	Extending Classical Test Theory to the Measurement of Change.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	Department of Health, Education, and Welfare, Washington, D.C. Office of the Commissioner of Education.
REPORT NO	CSE-60
BUREAU NO	BR-6-1646
PUB DATE	Oct 70
CONTRACT	OEC-4-6-061646-1909
NOTE	47p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Comparative Analysis, *Comparative Statistics, Correlation, Equated Scores, Evaluation Methods, *Measurement Techniques, *Reliability, Research Methodology, Research Problems, Research Tools, *Statistical Analysis, *True Scores
IDENTIFIERS	*Change Scores

ABSTRACT

The problem of the comparability of change scores is investigated. Change quotients and residual change scores are evaluated as alternative approaches and methods for estimating the true change and true score residual, the reliability of change scores and residuals, and procedures for constructing confidence intervals for residuals are explored. It is concluded that: (1) residuals can be used to compare the performance of individuals or groups while holding the initial status variables mathematically constant if the data meets the assumptions of a multivariate normal distribution; (2) group means should be used to compute residuals for comparing groups with the same sample size; (3) parallel-forms reliability of raw-score residuals and other estimated true scores are not necessarily equal to the index reliability; (4) for statistical analyses of the determinants of change, partial correlations or multiple regression analysis should be used with final status as the criterion and initial status as one of the covariates; and (5) errors in the predictor or initial status variables can change even the sign of partial correlation or multiple regression coefficients. It is recommended that test-retest estimates of reliability be used to correct coefficients for attenuation. A bibliography and statistical data are included. (AE)

CENTER FOR THE
STUDY OF
EVALUATION
UCLA
Graduate School
of Education
Los Angeles, California

EDUCATIONAL
CSE
RESEARCH
AND
DEVELOPMENT

EXTENDING CLASSICAL TEST THEORY
TO THE MEASUREMENT OF CHANGE

Edward F. O'Connor, Jr.

CSE Report No. 60
October 1970

CENTER FOR THE
STUDY OF
EVALUATION

Marvin C. Alkin, Director

[redacted]

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION is one of nine centers for educational research and development sponsored by the United States Department of Health, Education and Welfare, Office of Education. The research and development reported herein was performed pursuant to a contract with the U.S.O.E. under the provisions of the Cooperative Research Program.

Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique functions with an inter-disciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Theory and Methodology.

This publication is one of many produced by the Center toward its goals. Information on CSE and its publications may be obtained by writing:

Office of Dissemination
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

EXTENDING CLASSICAL TEST THEORY TO THE MEASUREMENT OF CHANGE

by

Edward F. O'Connor, Jr.

CSE Report No. 60
October 1970

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

2A

INTRODUCTION

The concept of a change score has considerable intuitive appeal. A person subtracts last week's weight from today's weight and talks of having gained or lost five pounds. Yet, change scores have more than their share of conceptual problems. Weights are comparable -- a two-hundred-pounder outweighs a one-hundred-pounder regardless of his other traits; but changes are not necessarily comparable -- a loss of twenty-five pounds may be a godsend for one individual but a disaster for another. Even in cases where changes in one direction are preferred, certain comparisons of changes appear inappropriate. For example, an instructor may grade physical education students on their improvement in running the mile. All of the students running an eight-minute mile at the beginning of the course may cut more than a minute out of their times; none of the four-minute milers are likely to improve by more than a few seconds. Clearly, the eight-minute milers "improved" their time by more seconds than did the four-minute milers. Yet no instructor would give A's to the slowest runners and F's to the fastest, regardless of his commitment to the concept of grading on improvement. Somehow, these "improvements" are not comparable for the purposes of evaluation. This inability to compare changes directly at different points of the scale, *even with ratio scales*, is the fundamental problem of the measurement of change.

The comparability problem is related to the fact that change scores are generally correlated with initial status. When change and initial status are negatively correlated, low-scores have an advantage in the sense they are likely to gain more. Similarly, in rarer instances when change and initial status are positively correlated, the initially high-scoring individuals have the advantage. The comparability problem can be alleviated by using either change quotients or residual change scores, both of which are independent of initial status. Change quotients and residuals are perfectly correlated with each other under certain circumstances. Residuals are to be preferred when the data meet certain assumptions which will be outlined in an ensuing section.

Methods for estimating the true change and true-score residual when the data are unreliable will be presented and the residual procedure will be extended to the comparison of groups, such as school systems. The reliability of change scores and residuals are discussed and procedures are suggested for constructing confidence intervals for residuals.

Change scores have also been used in statistical analyses of the determinants of change. A brief review of this use of change scores is provided which suggests that change scores are unnecessary and often even inappropriate for statistical studies. Alternative statistical procedures are suggested.

The Notational System

In general, capital letters refer to true scores or errorless scores, and small letters refer to the corresponding fallible scores. All scores are expressed in terms of deviation scores, i.e., their grand mean has been subtracted from them. This simplifies the computation because the mean of all deviation scores is zero. It does not affect the generality of any formula or proof since deviation scores can be converted back to the original scores whenever necessary.

X and x represent initial status,

Y and y represent final status, and

W and w represent a variable other than X or Y.

S_X^2 represent the variance of X; s_x^2 represents the variance of x.

R_{XY} represents the correlation between X and Y; r_{xy} between x and y.

Since the covariance of two true scores equals the expectation of the covariance between their corresponding fallible scores, both covariances will be represented by a capital C, such as C_{xy}.

Regression weights will be represented by A and B for true scores and a and b for fallible scores. Subscripts will be used unless the context indicates which regression weight is desired. $B_{YX.W}$ is the weight given X when both X and W are used to predict Y.

Other symbols will be defined as they appear.

THE RELATIONSHIP BETWEEN CHANGE AND OTHER VARIABLES

An early and continuous interest in psychology has been the relationship between change and other variables -- how can change be predicted? Thorndike (1924) cites six early studies on the relation of initial ability to gain. Other researchers like Woodrow (1946) correlated the "ability to learn" with other variables such as intelligence test scores. An examination of the weaknesses of the common statistical approaches suggest that change scores are unnecessary and often even inappropriate for statistical studies. Alternative statistical procedures are suggested.

The Correlation Between Change and Initial Status

Most correlations are reduced but not biased by errors in the data. A positive or negative correlation retains its sign but is smaller in absolute value. Thorndike (1924) demonstrated that the correlation between change and initial status is biased in a negative direction by errors in the pretest because the pretest error is also present in the change score but with the opposite sign.

$$x = X + e_x$$

$$g = y - x = G + e_y - e_x$$

where G = Y - X and g = y - x. Consequently, the covariance of the raw gain and raw initial status is not equal to the covariance of the corresponding true scores, as is generally the case:

$$\begin{aligned} C(g,x) &= C(G + e_y - e_x, X + e_x) \\ &= C(G, X) - s_{e_x}^2 \end{aligned}$$

Thomson (1924, 1925) and Zieve (1940) suggested analytic procedures which, in effect, added $s_{e_x}^2$ back to the raw score covariance before computing the correlation coefficient (Bereiter 1963, pp. 6-7).

Thorndike (1966) used parallel pretests to eliminate this bias. One pretest was used to compute the gain and the other was correlated with the gain. The average initial-gain correlation increased from -.20 to +.10. This concern with the initial-gain correlation appears to be a pseudo-problem, even for true scores. As Thorndike points out, correlation is positive only when the post-test variance is sufficently larger than the pre-test variance.

$$\begin{aligned} C(X,G) > 0 \text{ if and only if } C(X, Y-X) &> 0 \\ C_{XY} - S_X^2 &> 0 \\ R_{XY} S_Y - S_X &> 0 \\ R_{XY} &> \frac{S_X}{S_Y} \end{aligned}$$

Hence, the initial-gain correlation does not appear to add anything to our knowledge. In fact, if Thorndike's analysis is extended further, the initial-gain correlation issued can be parsimoniously subsumed under the heading "Predicting Y from X." If B_{YX} (or $R_{XY} \frac{S_Y}{S_X}$) is greater than one, equal to one, or less than one, the initial gain correlation will be correspondingly positive, zero, or negative (Garside, 1956).

Thorndike used mental age scores instead of I.Q.'s in his study. He points out that if he had used I.Q. scores with a standard variance at each age, the correlation between I.Q. at age 8 and the gain in I.Q. between age 9 and age 12 could be positive only if the age-8 test correlated more highly with the age-12 test than with the age-9 test--"a fairly improbable and unnatural event" (p. 126). He might have added that the correlation between I.Q. at age 8 and the I.Q. gain from age 8 to age 12 could not possibly be positive as long as the age-8 and the age-12 variances were equal since B_{YX} cannot exceed one unless S_Y is larger than S_X.

The difference between a positive and a negative initial-gain correlation seems more interesting than the difference between a B_{YX} of 1.05 and a B_{YX} of .95; yet both the initial-gain correlation and B_{YX} are determined by the same data, S_Y, S_X, and R_{XY}. The distinction between a positive and a negative initial-gain correlation appears to be artificial and misleading.

Change and Other Variables

Early studies correlated raw change with other variables and generally obtained near zero results (Woodrow, 1964). Lord (1963, p. 35) showed that such correlations may be quite misleading. If R_{GW} equals zero, then $R_{GW.X}$ will usually be positive. In other words, for every subgroup with the same initial status, W will be correlated positively with change. The question is whether the R_{GW} for the total group or the R_{GW} for each subgroup with the same initial status is more meaningful. Lord concludes, "In general, the more extraneous variables one can hold constant in a scientific study, the clearer the picture. For this reason, it is not the total group correlation R_{GW} but rather the partial correlation $R_{GW.X}(= R_{YW.X})$ that is usually of greater interest", (1963, p. 35). In other words, initial status is held mathematically constant so that the correlation between initial status and change does not influence our estimate of the relationship between change and a third variable, W.

When X is held constant, G is entirely dependent on the value of Y. Hence, $R_{GW.X}$ is mathematically equivalent to $R_{YW.X}$, but the interpretation of the two is slightly different. $R_{GW.X}$ is the correlation between change and W with X held constant while $R_{YW.X}$ is the correlation between Y and W with X held constant. The latter expression requires neither the computation nor even the concept of change scores. Similarly, Werts and Linn (1970, pp. 18-19) show that $B_{GW.X}$ equals $B_{YW.X}$. Just as the relationship between change and initial status can be more simply expressed in terms of B_{YX}, so the relationship between change and another variable W can be more simply expressed in terms of $B_{YW.X}$ or the equivalent partial correlation, $R_{YW.X}$. There is no need to compute change scores for correlational analysis.

Correcting Partial Correlation and Multiple Regression Coefficients for Unreliability

Unreliability in the data can reverse the sign of a partial correlation or multiple regression coefficient as well as affecting its sign. Consequently, zero-order correlations should be corrected for attenuation before entering them in partial correlation or multiple regression formulas.

Basefree Measures of Change

Thorndike, Bregman, Tilton, and Woodyard (1928) used a crude kind of partial correlation in their studies of Adult Learning, but logic behind the use of partials to study change was not clearly stated until DuBois (1957), Manning and DuBois (1958, 1962) and Lord (1958, 1963). Technically, Manning and DuBois used a kind of part correlation. They partialed initial status out of final status and then correlated the residuals with other residuals and variables. Their study (1962) showed that residual gains in learning studies were (a) more highly correlated with predictors such as aptitude tests than were raw gains, (b) more highly intercorrelated, and (c) could be accounted for by a single factor, which may be a general factor of psychomotor learning. In contrast, Woodrow (1946) had concluded from a review of studies using raw gains that intelligence was not related to the ability to learn and that there was no evidence for a general factor for learning ability. The difference between these sets of studies is that Manning and DuBois controlled for initial status through the use of part correlations. They concluded that, (a) "The correlations of residual gain are more consistent and more in line with what might be logically expected than are the correlations of crude gains...", (b) "Residual measures of learning seem to have more in common than do measures of crude gain in the same functions...", and (c) "The frequently low correlation between change in learned proficiency and aptitude measures should be re-interpreted in light of logical and empirical inadequacies of the crude difference criterion of change." (1962, pp. 318-19).

Manning and DuBois' residual approach does not take errors of measurement into consideration. Consequently, when x, y, and w are unreliable, the residual approach gives us $r_{w(y.x)}$ when $R_{W(Y.X)}$ is required. These coefficients, however, may even have opposite signs. Tucker, Damarin, and Messick (1966) attempt to correct this problem through the use of their "base-free measure of change". They partial true X rather than raw x out of y and correlate this "adjusted" residual with other variables. This procedure produces a part correlation which, at least, will always have the same sign as the corresponding true score part correlation but, nevertheless, would be a systematically biased estimate of $R_{W(Y.X)}$. For example, assume that Y and W are measured with perfect reliability but x is not. The estimate of $R_{W(Y.X)}$ would be:

$$R_{W(Y.X)} = \frac{r_{yw} - r_{wx} r_{yx}/r_{xx}}{\sqrt{1 + r_{xy}^{2}/r_{xx}}}$$

The correlation between the base-free measures of change and W would be:

$$r(w, y - B_{YX}x) = \frac{r_{wy} - r_{wx} r_{yx}/r_{xx}}{\sqrt{1 + r_{xy}^{2}/r_{xx}^{2} - 2r_{xy}^{2}/r_{xx}}}$$

which has the same sign as $R_{W(Y.W)}$ but a slightly different denominator. The point is not that the Tucker, et. al., approach is wrong; the above correlation could be adjusted to estimate $R_{W(Y.X)}$ or any other true score part or partial correlation that was required. However, it is far simpler mathematically to correct the appropriate partial correlation or multiple regression coefficient for attenuation without computing or conceptualizing in terms of change scores, residuals, or base-free measures of change.

CHANGE SCORES FOR COMPARING INDIVIDUALS

The introduction suggested that correlation between change and initial status made it inappropriate to use change score to evaluate individuals with different initial scores. An analogous problem occurred in the development of intelligence test scores. The first intelligence tests were scored in terms of "mental ages". A higher Mental Age (M.A.) meant the ability to answer more items correctly, but Mental Ages were not comparable in other ways for children with different chronological ages. For example, a Mental Age of seven is above average for a five-year-old, but below average for a nine-year-old. To make comparisons between children of various ages more meaningful, an Intelligence Quotient or I.Q. was defined as one hundred times the ratio of mental Age to Chronological Age (C.A.).

$$I.Q. = 100 \frac{M.A.}{C.A.}$$

This "ratio" I.Q. was still not completely comparable since it did not have the same standard deviation for all chronological ages. Hence, an I.Q. of 120 might mean the 95th percentile at one age and the 90th percentile at another age. More recent Intelligence Tests have used derivation I.Q.'s which have the same standard deviation for all ages (Mehrens & Lehman, 1969, p. 78).

It is important to consider very carefully what a deviation I.Q. score means and what it does not mean. Suppose that the deviation I.Q.'s have a standard deviation of 16 for all age groups. Then an I.Q. of 116 means that the person scored at the 84th percentile of the norm group for his age. If John has an I.Q. of 116 and Bill has an I.Q. of 84, John is above average for his age group and Bill is below average for his age group. However, one does not know whether John or Bill had a higher raw score unless he knows their chronological ages as well as their I.Q.'s. The I.Q.'s are simply a comparison of individuals while mathematically holding their age constant. Their ages are not "empirically" held constant because John's vocabulary at the age of five is not compared with Bill's vocabulary at the same age.

Similarly, Change Quotients (C.Q.) could be computed by holding initial status constant rather than age. Take, for example, the Physical Education instructor discussed in the introduction. He could grade his students on their improvement in running the mile by separating the students into groups according to their initial time and assigning C.Q.'s on the basis of the student's position within his own group. Runners who finished at 84th precentile of their group would be assigned at C.Q. of 116.

In this approach, the runner's C.Q. is derived by comparing him to other runners with the same initial time. Unless this group is very large, sampling error may seriously affect his C.Q. The sampling error becomes progressively more serious as the size of the comparison group decreases. Some grouping is possible, e.g., 8 minutes ± 15 seconds, but any attempt to group individuals with very different initial scores may defeat the purpose of computing change quotients. An approach is required which would decrease the sampling error by permitting the use of all of the data in assigning a C.Q. to a given individual. If the data meet the bivariate normal assumptions, then the Manning and DuBois' residuals (1962) provide such an approach by simply partialing initial status, X, out of final status, Y, and using the residual, Q.

$$Q = Y - B_{YX}X$$

If X and Y have a bivariate normal distribution, the Q's will be a normally distributed random variable with a zero mean and a constant variance at all levels of X, and will be independent of X.

Pretest and posttest times for running the mile will not meet the normal bivariate assumptions because the variance of Y (or Q) is not likely to be equal at all the levels of X. An appropriate non-linear transformation of the data is required. Fortunately, running speed is one such transformation (e.g., if John runs the mile in 6 minutes, his average speed is 10 miles per hour). Pretest and post-test speeds can plausibly be assumed to approximate a bivariate normal distribution. Consequently, speed will be used rather than time for computing the residuals.

Assume that there is an infinite population of individuals, and that their initial status, X, and their final status, Y, have a perfect bivariate normal distribution. It is easy to show that Change Quotients and residuals computed for this population would be perfectly correlated.

The Change Quotient for persons with an initial status X_k would equal

$$CQ_i = \frac{Y_i - \bar{Y}_k}{T} 16 + 100$$

where Y_i equals an individual's final status, $\bar{Y}_k$ equals the average final status of all persons starting with X_k, and T equals the standard deviation for Y given X, which is constant for all levels of X.

The residual, Q_i, would be equal

$$Q_i = Y_i - B_{YX}X_k \text{ or } Y_i - \dot{Y}_k$$

where $\dot{Y}$ equals $B_{YX}X_k$, the predicted Y for all individuals with initial status X_k.

But for an infinite population with a perfect bivariate normal distribution, $\hat{Y}_k$ equals $\overline{Y}_k$. Hence,

$$Q_i = Y_i - \overline{Y}_k$$

and residuals and Change Quotients are perfectly correlated.

The Relative Efficiency of Change Quotients and Residuals

To compare the relative efficiency of Change Quotients and residuals, consider a sample of 100 persons from the infinite population described above. The Change Quotients and residuals are not necessarily perfectly correlated nor are they necessarily equal to Change Quotients and residuals computed on the basis of the entire population.

To simplify the analysis, a simple linear transformation of the Change Quotients will be used:

$$CQ = Y_i - \overline{Y}_k$$

Now the Change Quotients and residuals computed on the basis of the entire population are identical, and can jointly be designated as $CQ/Q(pop)_i$, which represents the value of CQ/Q_i derived from the infinite population. It is not a population value since it represents only individual i.

A kind of standard error of measurement can be derived for either the sample-derived Change Quotients or the sample-derived residuals which will represent the extent to which sample-derived values differ from the population-derived CQ/Q(pop). The sample-derived residuals have a smaller standard error of measurement than do the sample-derived Change Quotients. The error of measurement for the Change

Quotients equals:

$$\begin{aligned} \text{Error (CQ)} &= CQ_i - CQ/Q(\text{pop})_i \\ &= Y_i - \overline{Y}_k(\text{sample}) - Y_i + \overline{Y}_k(\text{pop}) \\ &= \overline{Y}_k(\text{pop}) - \overline{Y}_k(\text{sample}) \end{aligned}$$

Hence, the standard error of the Change Quotients equals the standard error of $\overline{Y}_k$, or

$$\text{S.E.(CQ)} = \text{S.E.}(\overline{Y}_k) = T\sqrt{\frac{1}{n_k}}$$

where T equals the standard deviation of Y given X and n_k equals the number of persons in the sample with the same initial X_k.

Similarly, the standard error of the residual equals the standard error of Y_k, or, from Draper and Smith (1966, p. 22),

$$\text{S.E.}(Y_k) = T \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{(X-\bar{X})^2}\right]^{\frac{1}{2}} \qquad N = \text{total sample size}$$

Since a finite sample of one is being used rather than an infinite population, some grouping will be necessary to compute the Change Quotients. Assume interval size of one-half of a standard deviation, the comparison group centered around the mean of x ($Z_x = 0$) would then be expected to contain approximately twenty persons, and the comparison group centered around an x value of 2 standard deviations away from the mean ($Z_x = \pm 2$) would be expected to contain approximately three persons. With this grouping and a sample size of one hundred, the standard error of the residuals would be less than one-half the standard error of the Change Quotients. For example, when $Z_{x_k} = 0$, the standard errors are .10T and .22T for the residuals and Change Quotients, where T, once again, is the standard deviation of Y given X. For $Z_{x_k} = \pm 2$, the corresponding values are .22T and .58T. Since the grouping procedure outlined above introduces a small bias in the estimate of Change Quotients, the standard error for CQ's is actually a slight underestimate. Any attempt to increase the size of grouping interval would increase this bias.

The Reliability of the Residual when X and Y are Perfectly Reliable

The last section demonstrated that the residual was a more precise version of the Change Quotient when X and Y had a bivariate normal distribution. However, because of sampling errors, the sample-derived residual did not exactly equal the population-derived value even when X and Y contain no measurement errors. The difference between the population-derived value, Q_p, and the sample-derived value, Q_s, is illustrated below. In this example, X_u and Y_u represent the original uncorrected scores rather than deviation scores.

$$Q_p = (Y_u - \bar{Y}_p) - B_p (X_u - \bar{X}_p)$$

$$Q_s = (Y_u - \bar{Y}_s) - B_s (X_u - \bar{X}_s)$$

$$Q_p - Q_s = (\bar{Y}_s - \bar{Y}_p) - (B_s\bar{X}_s - B_p\bar{X}_p) + X_u(B_s - B_p)$$

where $B_p = B_{YX}$ for the population and $B_s = B_{YX}$ for the sample. The first two terms represent constants for a given sample, and consequently would affect all the sample residuals in the same way without altering their order. The third term, however, is a variable which would affect the order of the sample residuals. In short, John's residual might be larger than Mike's when the sample B_{YX} is used but smaller than Mike's when the more accurate population B_{YX} is used. Presumably, the researcher would be more interested in the population-derived order. Hence, there is a limited sense in which residuals based on true scores, X and Y, are less than perfectly reliable.

To derive an appropriate reliability coefficient, assume that two finite samples of size N are drawn from an infinite population with a bivariate normal distribution for X and Y. The sample estimates of B_p, B_1 and B_2, are then used to compute two sets of residuals, Q_1 and Q_2, for the entire population. R_{QQ} is derived as follows:

$$E(R_{QQ}) = \frac{E(Q_1Q_2)}{\left[E(Q_1^2)E(Q_2^2)\right]^{\frac{1}{2}}} \text{ where}$$

$$E(Q_1Q_2) = S_Y^2 + S_X^2B_p^2 - 2C_{XY}B_p = S_Y^2\ (1 - R_{XY}^2)$$

$$E(Q_1^2) = E(Q_2^2) = S_Y^2 + S_X^2\left[B_p^2 + E(e_b^2)\right] - 2C_{XY}B_p$$

and where $e_b = B_i - B_p$ and is normally distributed about a mean of zero with a variance of $E(e_b^2)$:

$$E\ (e_b^2) = \frac{S_Y^2\ (1 - R_{XY}^2)}{NS_X^2}$$ (Draper and Smith, 1966, p. 18-19)

$$E(Q_1^2) = S_Y^2\ (1 - R_{XY}^2) + \frac{1}{N} S_Y^2\ (1 - R_{XY}^2)$$

Hence: $$R_{QQ} = \frac{N}{N+1}$$

Since R_{QQ} equals .91 for a sample of only ten, this source of unreliability can be generally ignored with any reasonably sized sample. An exception to this general rule is the case of subjects at the extremes of the X distribution. R_{QQ} is the <u>average</u> reliability of the residuals, and residuals with extreme X's are less reliable than residuals for less extreme X's. For example, if B = 1.5 is used rather than the correct B = 1 in the formula Q = Y - BX, a residual with $X_u = \overline{X}_p$ would not be affected while a residual with an X_u four units away from $\overline{X}_p$ would contain an error of two units.

The Relationship of Residuals to "Real" Change

If a person gains or loses five pounds, "This is a definite fact, and not a result of an improper definition of growth" (Lord, 1963, p. 23). Similarly, Cronback and Furby comment, "One cannot argue that the residualized score is a 'corrected' measure of gain since in the 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". (1970, p. 74)

Using Multiple Pretest to Compute Residuals

Intelligence quotients compare mental ages while mathematically holding chronological age constant. Similarly, Change Quotients or residuals can be used to compare individuals on one variable while holding another variable constant. For example, the final examination for beginning Russian might produce only chance differences among most of the students on the first day of class. Although the students may begin with an equal (zero) knowledge of Russian, they do not have equal language ability. Consequently, a language aptitude test would seem to be a more accurate measure of a student's initial status.

Change Quotients or residuals can also be used to hold more than one variable constant. Psychologists, for example, often have test norms broken down according to sex, age, and socio-economic status.

Such a breakdown holds each of these variables constant while evaluating an individual's test score.[1] If Y and the X variables meet the assumptions of a multivariate normal distribution, the appropriate residual is the difference between Y and the predicted $\hat{Y}$, where $\hat{Y}$ is the multiple regression estimate of Y based on the X variables.

$$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3$$

$$Q = Y - \hat{Y} = Y - B_1X_1 - B_2X_2 - B_3X_3$$

Readers should consult a standard textbook on multiple regression analysis for the handling of dummy variables such as sex and other variables which do not meet the multivariate normal assumptions.

Any variable which is correlated with the posttest or criterion Y will be reflected in that criterion unless it is held constant (partialed out). Which variables should be held constant depends on purposes for collecting the data. A physical education teacher might want to hold age or weight constant while evaluating a pupil's performance--a coach selecting a track team would not.

ESTIMATING RESIDUALS AND CHANGE SCORES FOR FALLIBLE DATA

Estimating True Change

In many cases, it would be desirable to have the true gain,

[1] Here it is assumed that the test score has the same interpretation regardless of other variables. On a vocabulary test, for example, high scores are preferred, regardless of the person's age, sex, or socio-economic status. At the far end of the continuum are those tests where the interpretation is completely dependent on the person's status on another variable (e.g., a masculinity-femininity scale).

$G = Y - X$, rather than fallible values, x, y, and $g = y - x$. The raw gain, $g = y - x$, can be used to estimate the true gain, $G = Y - X$, but the reliability of the raw gain is notoriously low. Lord (1956, 1958, 1963) and McNemar (1958) suggested using the multiple regression analysis to estimate the true gain. In short, they used the estimator $G = b_{Gx.y}x + b_{Gy.x}y$, to estimate true gain where

$$b_{Gx.y} = \frac{r_{xG} - r_{yG}\, r_{xy}}{1 - r_{xy}^2} \quad \frac{S_G}{s_x}$$

$$b_{Gy.x} = \frac{r_{yG} - r_{xG}\, r_{xy}}{1 - r_{xy}^2} \quad \frac{S_G}{s_x}$$

The Lord-McNemar approach can be expressed in more general terms. To estimate the true gain, $G = Y - X$, with some estimator, $\hat{G} = ky + lx$, where k and l are weights or constants, t is defined as the error of estimate, i.e., difference between the true value, G, and our estimate, G:

$$t = G - \hat{G}$$

Calculus is used to find the weights k and l which will minimize the variance of these errors of estimates.

$$V = t^2 = \Sigma(G - \hat{G})^2 = \Sigma(G - ky - lx)^2$$

To minimize V, the partial derivatives of V with respect to k and l are set equal to zero and solved for k and l.

$$\frac{\partial V}{\partial k} = -2\Sigma(G - ky - lx)y = 0$$

$$\frac{\partial V}{\partial l} = -2\Sigma(G - ky - lx)x = 0$$

Lord and McNemar gave computational formulas for k and l.

The correlation between the estimator, $\hat{G}$, and the true gain will be as least as high as the correlation between the raw gain, $g = y - x$, and the true gain, and will usually be higher. This correlation is simply the multiple correlation of G given x and y and can be found using the normal formulas. The estimate, $\hat{G}$, is the least squares estimate of the true gain. It is an unbiased estimate if X, Y, x, and y have a multivariate normal distribution.

Estimating the True Residual

Similarly, the true score residual, $Q = Y - B_{YX}X$, can be estimated using the fallible values, the x and y:

$$\hat{Q} = ky + lx$$

$$V = \Sigma t^2 = \Sigma(Q - \hat{Q})^2 = \Sigma(Q - ky - lx)^2$$

$$\frac{\partial V}{\partial l} = -2\Sigma(Q - ky - lx)y = 0$$

$$\frac{\partial V}{\partial l} = -2\Sigma(Q - ky - lx)x = 0$$

The partial derivatives are set equal to zero and solved for k and l;

$$k = \frac{r_{xx}r_{yy} - r_{xy}^2}{r_{xx}(1 - r_{xy}^2)}$$

$$l = k\,b_{yx}$$

The least-squares estimator of $Q = Y - B_{YX}X$ is $k\,(y - b_{yx}X)$, or k times the raw score residual, $q = y - b_{yx}x$. This agrees with the formula presented by Cronbach and Furby (1970, Errata). (Recall that b_{yx} does not equal B_{YX} unless $r_{xx} = 1$.)

Similarly, for the multiple-X case, the least squares estimator of $Q = Y - B_fX_f - B_gX_g$ is $k(y - b_fx_f - b_gx_g)$ or k times the raw residual, $q = y - b_fx_f - b_gx_g$, where k equals:

$$\frac{(r_{yy}r_{ff}r_{gg} - r_{yy}r_{fg}^2 - r_{ff}r_{yg}^2 - r_{gg}r_{yf}^2 + 2r_{yf}r_{yg}r_{fg})\,(1 - r_{fg}^2)}{(1 - r_{fg}^2 - r_{yg}^2 - r_{yf}^2 + 2r_{yf}r_{yg}r_{fg})\,(r_{ff}r_{gg} - r_{fg}^2)}$$

Note that the k for both the single-X and multiple-X cases will equal one when x and y are perfectly reliable. This is to be expected since the estimated residual is exactly equal to the true residual whenever the data are error free.

The general rule is that the least-squares estimator of the true residual is more k times the raw residual. As will be shown later, k equals r^2_{qQ}, the square of the correlation between true residual and the raw residual. The k can generally be ignored since it does not affect the order of the residuals.

Using the Additional Variables to Estimate the True Gain and the True Residual

Cronback and Furby (1970) extended the Lord-McNemar approach by using additional variables, w and z, to make a more precise estimate of the true gain, G:[2]

$$\hat{G} = ky + lx + mw + nz$$

where w represents one or more supplementary measures of the individual's initial status and z represents one or more supplementary measures of the individual's final status. Again, the variance of errors of estimate is minimized.

$$V = \Sigma t^2 = \Sigma(G - \hat{G})^2 = \Sigma(G - ky - lx - mw - nz)^2$$

by setting the partial derivatives of k, l, m, and n, equal to zero

[2] Cronbach and Furby also distinguished between linked and unlinked errors, following a distinction made by Stanley (1967). Linked measurement errors may occur when the x and y observations are obtained at the same time or by the same observer. Here the Classical Test theory assumption of independent errors of measurement will be followed since the x and y observations for change scores will normally be collected at different times.

and solving for k, l, m, and n. This estimate of true gain will be designated as $\hat{G}/wxyz$. Similarly, true X and true Y could be estimated using $\hat{X}/wxyz$ and $\hat{Y}/wxyz$. Cronbach and Furby showed that

$$\hat{G}/wxyz = \hat{Y}/wxyz - \hat{X}/wxyz$$

Later they estimated true residuals by inconsistently alternating between a straightforward least-squares estimator of true residual, and an estimate based on $\hat{Y}/wxyz$ and $\hat{X}/wxyz$ which they may believe is equivalent to the least-squares estimate. For example, they estimated $Q = Y - B_{YX} = Y.X$ using the conventional multiple regression equation (1970, formula #24)

$$\widehat{Y.X} = b_1x + b_2y + b_3w + b_4z$$

where b_1, b_2, b_3, and b_4 equal the appropriate regression weights and Y.X equals Y with X partialed out. Then they estimated $Q = Y - BX - BW = Y.XW$ using $\hat{Y}.\hat{X}\hat{W}$ (1970, Errata), which is a strange combination of estimated true scores:

$$\hat{Y}.\hat{X}\hat{W} = \hat{Y}/wxyz - \hat{Y}/\hat{X}\hat{W}$$

where $\hat{Y}/\hat{X}\hat{W}$ is the least squares estimate of Y based on $\hat{X}/wxyz$ and $\hat{W}/wxyz$.

While $\hat{Y}/wxyz - \hat{X}/wxyz$ is equivalent to $\hat{G}/wxyz$, $\hat{Y}.\hat{X}\hat{W}$ is not equivalent to the appropriate $\widehat{Y.XW}/wxyz$ estimated by the following regression equation:

$$\widehat{Y.XW} = b_1x + b_2y + b_3w + b_4z$$

where b_1, b_2, b_3, and b_4 are the appropriate regression weights. That $\hat{Y}.\hat{X}\hat{W}$ does not necessarily equal $\widehat{Y.XW}$ can be shown by a simple example. Assume that $r_{yy} = 1$, $r_{xy} = 0$, $r_{xy} = r_{wy} = 0$, and r_{zy}, r_{zx}, and r_{zw} are any values other than zero. $\widehat{Y.XW}$ would equal y, while $\hat{Y}.\hat{X}\hat{W}$ would equal y minus some function of x, w, and z.

Bias and The Least Squares Estimate

Least squares estimators of the true gain and true residual are unbiased if Classical Test Theory and multivariate normal assumptions hold (Mood & Graybill, 1963, p. 329). For example, given x and y, the expectation of k times the raw residual, $kq = k(y - b_{yx}X)$, equals the true residual, $Q = Y - BX$, if the assumptions hold.

$$F[k(y - b_{yx}x)/xy] = Y - B_{YX}X$$

$$E(kq/xy) = Q$$

Curiously enough, the true residual $Y - B_{YX}X$ is not the least squares estimate of $y - b_{yx}X$, but is the least squares estimate of $q_c = y - B_{YX}X$, the corrected residual or "base-free measure of change".

$$E(Y - B_{YX}X/XY) = y - B_{YX}X$$

This may have led to Traub's (1967, 1968) and Glass's (1968) disagreement over the relative merits of the raw residual, $q = y - b_{yx}x$. Glass apparently accepts Traub's statement that the raw residual is not an unbiased estimator of the true residual: "Depending on whether scores on the pre-measure are above or below the mean and on whether the regression coefficient is positive or negative, [q], on the average, will systematically over- or under-estimate [Q]. Because of this bias, the error of measurement in [q], that is the quantity [q - Q], is not independent of [Q]". (1967, p. 255). Actually, both q-Q and q_c-Q are independent of Q, and the over-estimate/under-estimate problem is less serious than it appears. The raw x is, on the average, a systematic over-estimate or under-estimate of X depending on whether x is above or below the mean. Consequently, x is used when only the order of X is of interest, and $r_{xx}x$ is used when ar unbiased estimate of the magnitude of X is required. Similarly, the raw residual, $y - b_{yx}x$, should be used to estimate the order of the true residuals, and k times the raw residual should be used when it is necessary to estimate the magnitude of the true residual.

RESIDUAL CHANGE SCORES FOR COMPARING GROUPS

Residuals are also appropriate for comparing changes in groups of individuals. Dyer, Linn, and Patton (1967, 1969) have used residuals, which they call discrepancy measures, to evaluate the effectiveness of school systems. They deplore the common tendency "to compare the average achievement test score of students in a given system with some sort of national average or norm and to assume that the discrepancy between the two averages constitutes a measure of the educational effectiveness of the system". (1969, p. 591)

> For example, if the students from System A tend to score lower on reading tests than the students from System B, it is often assumed that the teaching of reading in System A is less effective than the teaching of reading in System B. Or if the incidence of juvenile delinquency in System X is greater than the incidence of juvenile delinquency in System Y, it is assumed that System Y is doing a better job character training and inculcating the attitudes of good citizenship than is System X. Such assumptions can be wholly unreasonable. Looking solely at what pupils are like as they emerge from any phase of an educational system tells nothing whatever about how the system is functioning. One has to know in addition what relationships may exist between the characteristics of youngsters as they come out of any phase of the system and the characteristics with which they entered that phase of the system. (1967, p. 8)

The 1969 study used grade-equivalent scale scores from the Iowa Tests of Basic Skills. Sixty-four school systems were compared, including the 9,972 students for whom there was complete data for both testings, one in the fifth grade and the other in the eighth grade. Dyer et al called their procedure "a matched-longitudinal sample", matched because the same pupils were included in the school means for both the pretests (or inputs) and the posttests (or outputs), and longitudinal because the pretests and posttests were three years apart

Two methods of computing the residuals were compared. The first method computed residuals for the individual pupils and then averaged these residuals over a school system to get the school system residual. The individual procedure shall be designated as Method-I. The second procedure averaged the pretests and posttests for each school system and used these means to compute residuals for the school system. The school-system mean procedure shall be designated as Method-M.[3]

The outputs were limited to five major subtests and a composite score, all taken in the eighth grade. The inputs at the fifth grade included the major and minor subtests and the composite, 15 input measures in all. The Method-I and Method-M analyses were performed separately for each of the six output measures

> A stepwise multiple regression procedure was used for these analyses. The input measure (i.e., fifth-grade test score) that had the highest correlation with the output measure was selected first. Given the first input measure, the input

[3] The study also compared two other methods for computing residuals. Since these residuals were largely uncorrelated with the Method-I and Method-M residuals, they are of less theoretical importance and will not be discussed here.

> measure that added the most to the multiple correlation was next selected for inclusion in the regression equation. This process was repeated by adding input measures to the regression equation one at a time until the squared multiple correlation increased by less than .01. The deviations of the school system means from the appropriate regression surface were then computed for each method. (1969, p. 595)

The Dyer et al data suggested that Method-I is preferable to Method-M. The mean correlation between the major subtests at the fifth grade and the same subtest three years later is slightly higher for individual scores than for school system means (81.2 vs. 79.5). Furthermore, Method-I produces a slightly higher mean multiple-correlation (.831 vs. .825). However, the Method-I and Method-M figures are not entirely comparable since the Method-I figures refer to correlations between individual scores and the Method-M figures to correlations between means.

Re-analysis[4] of the data showed that when both multiple-correlations referred to mean outputs,[5] the Method-M multiple-correlations were superior in every case, although the differences were small (mean multiple-correlation for Method-I was .803, for Method-M, .825).

More importantly, Dyer et al estimated reliability of the residuals produced by both methods and found that the school system residuals "were somewhat more stable" (p. 604) when Method-I was used.

[4] I am indebted to Dr. Robert L. Linn for graciously permitting the re-analysis of the Dyer, Linn, and Patton data for this study.

[5] The correlation between predicted school system mean output and actual mean output. Method-I should be judged on the precision with which it predicts mean outputs, since school systems, rather than individuals, are the unit of interest.

(This is as close as Dyer et al come to stating a preference for either method). To get these reliability estimates,

> The matched-longitudinal sample of students within each school system was divided into two random subsamples of equal size and Methods I and II were repeated with each subsample. Deviation scores for each school system measure were then computed for each subsample in the same manner as described above for the total sample. The deviations for the first subsample were correlated with the deviations for the second subsample. (p. 595)

The analyses were performed separately for each of the six outputs. The median r was .78 for Method-I and .72 for Method-M.

Despite the data presented by Dyer et al., Method-M is preferable when computing residuals for groups. While the Method-I residuals may be more reliable, they are also biased since they are correlated with both the inputs and the predicted outputs. They should be uncorrelated with both. In fact, the Method-I residuals are more reliable because they are systematically biased. This conclusion can be demonstrated both mathematically and empirically.

First, examine a regression equation, standard in every way except that school system means are predicted instead of individual scores.

$$\overline{Y} = B_1\overline{X}_1 + B_2\overline{X}_2 + B_3\overline{X}_3 + Q$$

or

$$Q = \overline{Y} - B_1\overline{X}_1 - B_2\overline{X}_2 - B_3\overline{X}_3$$

Here Q is the system mean residual, and B_1, B_2, and B_3 are the weights which minimize the variance of Q. Method-I first estimates the individual residuals (here designated P to distinguish their mean $\overline{P}$ from Q).

$$Y = A_1X_1+A_2X_2+A_3X_3+P$$

$$P = Y - A_1X_1-A_2X_2-A_3X_3$$

where A_1, A_2, and A_3 regressions weights with the school-system means.[6]

$$\overline{P} = \frac{1}{n} \Sigma P$$

$$= \frac{1}{n} \Sigma (y - A_1X_1-A_2X_2-A_3X_3)$$

$$= \overline{Y} - A_1\overline{X}_1-A_2\overline{X}_2- A_3\overline{X}_3$$

Hence, both methods use the same data (the school system means), but with potentially different sets of beta weights. The A regression weights are the least-squares solution for the individual scores, but not for the system means. Consequently, the $\overline{P}$ residuals will have a larger variance than the Q residuals and a smaller multiple-correlation.

Furthermore, the A regression weights insure that the individual residuals are uncorrelated with the individual inputs and individual predicted outputs. The B weights, on the other hand, insure that the school system residuals are uncorrelated with either the mean inputs or the predicted mean output. If the A weights and B weights are not identical, the Method-I system residuals (P) will be correlated with both the mean inputs and predicted mean outputs.

Re-analysis of the Dyer _et al_ data showed that Method-I residuals had a substantial correlation with their predicted outputs (the mean

[6] "Almost equivalent" is more precise. The grand mean of the individual data will not necessarily equal the grand mean of the school-system means and hence the $\overline{P}$ residuals will not necessarily have a mean of zero.

absolute value of the correlations was .147), while the Method-M residuals were essentially uncorrelated with their predicted outputs (the highest r was .0003).

The mean of correlations between each of the Method-I and Method-M residuals and all fifteen of the potential inputs was computed. Again, there was no overlap between the two distributions. The lowest mean correlation for Method-I was .065, the highest for Method-M was .040. The corresponding overall mean for Method-I and Method-M were .130 and .025. The absolute value of the correlations were used to compute these means in order to determine average size of the correlations. The direction of the bias was not considered.

Conclusion

The School System residuals based on the individual-data regression weights (Method-I) were biased because of their correlations with the predicted output and the various inputs and potential inputs. These correlations may seem small from a predictive perspective, but they represent a substantial bias. When residuals were computed for groups, the group should be the unit of analysis for all of the regression equations.

THE RELIABILITY OF CHANGE SCORES AND RESIDUALS

In Classical Test Theory, the following three definitions of reliability are equivalent: (1) parallel-forms reliability, the correlation between parallel tests, r_{xx}(p.f.), (2) proportional reliability, the proportion of true variance in the observed variance, r_{xx} (pro), and (3) the square of the index of reliability, the correlation between the true score and the raw score, r_{xx} (index). Although these definitions are equivalent for single raw scores when the Classical assumptions hold, they are conceptually different and can differ in many circumstances. Zimmerman and Williams (1965a, 1965b, and 1966), for example, demonstrated that proportional and parallel-forms reliability are not equal when the true score and error score are correlated, a realistic assumption for multiple-choice tests. In fact, under such circumstances, parallel forms reliability can be positive even though the error variance and observed variance are equal (1966).

Proportional reliability is defined here as the porportion of true variance in the observed variance.[7] Hence, if $\hat{X} = w + y$ is used to estimate X, the proportional reliability of this estimate is:

$$r_{\hat{X}\hat{X}} \text{ (pro)} = \frac{r_{yy}\, s_y^2 + r_{ww}\, s_w^2 + 2C_{wy}}{s_y^2 + s_x^2 + 2C_{wy}}$$

This definition of proportional reliability is equivalent to the parallel-forms definition of reliability when the Classical Test Theory assumptions hold.

This paper often uses estimates for true scores, such as estimates for the true gain and true residual. Parallel-forms reliability and index reliability are not necessarily equal for such estimates. For example, assume that true X is estimated using true Y and true W.

[7] This definition is comparable to Guilford's definition of reliability as "the proportion of true variance in obtained test scores" (1954, p. 350), incorrectly stated by Traub (1967) as "the ratio of true score variance to observed score variance". The ambiguity of the "ratio" definition led to Traub's (1967, 1968) and Glass's disagreement

Parallel-forms reliability would equal one for this estimate while index reliability would equal the squared multiple correlation. Index reliability cannot exceed parallel-forms reliability.

The general rule is that parallel-forms reliability and index reliability are equal only if the components of the raw score or estimate have the same relative weights as the corresponding true score components.

$$\text{If:} \quad Z = A_1X_1 + A_2X_2 + \ldots + A_jX_j + \ldots + A_nX_n$$

and

$$\hat{Z} = a_1x_1 + a_2x_2 + \ldots + a_jx_j + \ldots + a_nx_n = z$$

where the subscripts refer to different variables and all the x variables follow the Classical Test Theory assumptions regarding errors.

$$\text{Then:} \quad r_{zz}(\text{p.f.}) = \frac{\Sigma r_{jj} a_j^2 s_j^2 + \Sigma r_{jk} a_j a_k s_j s_k}{\Sigma a_j^2 s_j^2 + \Sigma r_{jk} a_j a_k s_j s_k} \qquad \text{where } j \neq k$$

(Similar to Guilford, 1954, p. 393)

and

$$r_{zZ} = \frac{\Sigma r_{jj} a_j A_j s_j^2 + \Sigma r_{jk} a_j A_k s_j s_k}{(\Sigma r_{jj} A_j^2 s_j^2 + \Sigma r_{jk} A_j A_k s_j s_k)^{1/2} (\Sigma a_j^2 s_j^2 + \Sigma r_{jk} a_j a_k s_j s_k)^{1/2}}$$

where $j \neq k$

Therefore: r_{zz} (p.f.) = r_{zz} (index) or r_{zZ}^2

if and only if: $a_j = cA_j$ for all j, where c is an arbitrary constant.

---

as to which true score should be used in reliability estimates for residuals.

Reliability of Change Scores and Other Composite Scores

The reliability of a change score y = z - x, then $s_x = s_y$ and $r_{xx} = r_{yy}$ is:[8]

$$r_{gg} = \frac{r_{xx} - r_{xy}}{1 - r_{xy}} \qquad \text{(Lord, 1963, p. 32)}$$

This formula highlights the effect of r_{xy} on r_{gg}. An increase in a positive r_{xy} will lower the reliability of the change score, because x and y will have more of their true variance in common and the difference between y and x will be primarily error. Note that r_{gg} cannot exceed r_{xx} unless r_{xy} is negative. Were r_{xy} negative, the use of change scores would be highly questionable, to say the least.

The more general formula for the reliability of change scores is:

$$r_{gg} = \frac{r_{yy}s_y^2 + r_{xx}s_x^2 - 2C_{xy}}{s_y^2 + s_x^2 - 2C_{xy}} \qquad \text{(Lord, 1963, p. 32)}$$

This formula is expressed in terms of the proportion of true variance to total variance. Reliability can also be expressed in terms of the proportion of error variance.

$$r_{gg} = 1 - \frac{s_{e_y}^2 + s_{e_x}^2}{s_g^2}$$

where $s_{e_y}^2 = (1 - r_{yy})\, s_y^2$, $s_{e_x}^2 = (1 - r_{xx})\, s_x^2$, and $s_g^2 = \Sigma(y-x)^2/n$

This formula can be generalized to give the parallel-forms/ proportional reliability of any composite score where there is independent information on the reliability of the various components.

$$r_{zz}\ (\text{p.f./pro}) = 1 - \frac{\Sigma a_i^2 s_{e_i}^2}{\Sigma a_j^2 s_j^2}$$

[8] Reliability coefficients are equal for all three definitions unless otherwise noted, e.g., r_{xx} (p.f.).

where $z = \Sigma a_j x_j$ and $s^2_{e_j} = (1-r_{jj})s^2_{x_j}$

Index Reliability of Estimated True Scores

Index reliability has been defined earlier as the square of the correlation between the true score and the raw score or estimate. For estimated true scores, index reliability equals the squared multiple correlation.

When k times the raw residual is used to estimate the true residual,

$$\hat{Q} = kq$$

where $Q = Y - \hat{Y}$, $q = y - \hat{y}$, and k is the value which minimizes the variance, $\Sigma(Q-kq)^2$, k equals the squared multiple correlation In other words, the square root of k is equal to the correlation between q and its true residual Q:

$$\sqrt{k} = r_{qQ} \text{ or } k = r^2_{qQ} = r_{qq} \text{ (index)}$$

k also equals the ratio of the variance of Q to the variance of q:

$$k = \frac{s^2_Q}{s^2_q}$$

The regression of Q on q can be represented as B_{Qq}. This beta weight also equals k.

$$B_{Qq} = r_{qQ} \frac{s_Q}{s_q} = \sqrt{k}\sqrt{k} = k$$

Hence k or r_{qq} (index) performs much the same function for residuals that r_{xx} provides for single raw scores. Multiplying the raw residuals by k (or B_{Qq}) reduces their variance to the point that minimizes the variance of the errors of estimates $(Q - \hat{Q})$. Just as $r_{xx}x$

is the least squares estimate of X, kq is the least squares estimate of Q.

For the specific case of $q = y - b_{yx}x$, the index and parallel-forms reliability coefficients are:

$$r_{qq} \text{ (index)} = \frac{r_{yy} - r_{xy}^2/r_{xx}}{1 - r_{xy}^2} = \frac{r_{xx}r_{yy} - r_{xy}^2}{r_{xx}\,(1 - r_{xy}^2)}$$

$$r_{qq} \text{ (p.f./pro)} = \frac{r_{yy} + r_{xy}^2/r_{xx} - 2r_{xy}^2}{1 - r_{xy}^2}$$

Which Reliability?

Parallel-forms reliability for residuals and other estimates are equivalent to proportional reliability and is the best estimate of the test-retest reliability of these estimates (assuming that the same weights are used for both samples). It is greater than or equal to index reliability which estimates the precision of the estimates (i.e., the extent to which they correlate with the true score). Consequently, parallel-forms reliability of estimates is generally an over-estimate of their precision. Both reliabilities should be reported.

The Reliability of the Difference Between Residuals

To compare residuals, either across individuals on the same variable or across variables for the same individual, one must construct confidence intervals based on the reliability of these residuals. Otherwise, decisions may be based on chance differences. For example, the 95% confidence interval for a given residual, q, is:

$$C.I. = q \pm 1.96\, s_q\, (1 - r_{qq})^{1/2}$$

and for the difference between the residuals for two different individuals on the same variable is

$$C.I. = (q_i - q_j) \pm 1.96\, s_q\, (2 - 2r_{qq})^{1/2}$$

To compare residuals for the same individual across variables is more complicated. First, the residuals must be standardized to the same variance; they already have the same mean, zero. The reliability of the difference between these standardized residuals equals the reliability of any difference or change score when the two variances are equal:

$$r_{dd} = \frac{r_{11} + r_{22} - 2r_{12}}{2(1 - r_{12})} \qquad \text{(Guilford, 1954, p. 394)}$$

where r_{11} and r_{22} are the reliabilities of the residuals and r_{12} is their intercorrelation. The 95% confidence interval for the intra-individual difference between two residuals is:

$$C.I. = (q_1 - q_2) \pm 1.96\, s_q\, (1 - r_{dd})^{1/2}.$$

The index reliability of the residuals should be used to compute these confidence intervals since confidence intervals are concerned with the value of the true scores. Parallel-forms or proportional reliability may be used when no estimate of index reliability is available.

Illustration Using the Dyer, Linn, and Patton Data

An earlier section of this paper discussed how Dyer, Linn, and Patton (19[illegible]) computed residuals separately for both halves of their school-system sample to obtain an estimate of the reliability of the

Method-I and Method-M residuals. This split-half correlation must be stepped up using the Spearman-Brown formula (Gulliksen), 1950, p. 63) to estimate the parallel-forms reliability of residuals based on the total samples. Table 1 in the appendix presents these stepped-up coefficients for the Method-M residuals. The parallel-form reliabilities are fairly high, ranging from .77 to .91 with a mean of .835. These estimates represent only very rough approximations since the school-system samples used to compute these residuals ranged from 10 to 1084 pupils. However, these parallel-form reliability estimates can be used to illustrate some of the problems of comparing residuals across variables.[8]

Table 1 also presents the intercorrelation between the Method-M residuals. These are also fairly high in many cases, suggesting that intra-school-system comparisons across some variables may be entirely unwarranted. Table 2 presents the reliability estimates for intra-school-system comparisons, based on the data in Table 1. These estimates vary from zero to .73, making any kind of generalization impossible. Obviously some comparisons, such as Language vs. Arithmetic, are reasonably reliable, while others are quite unreliable. This disparity illustrates the need for treating each research setting as a special case and computing the reliability of such differences before deciding whether to make intra-individual (or intra-school-system) comparisons of residuals.

[8] No estimate of index reliability is available. Of course, an index reliability estimate will also be only a rough approximation because of the range in the sample sizes used to compute the residuals.

Finally, Table 3 presents the intercorrelations between the predicted outputs for Method-M.[9] These correlations are quite high, ranging from .84 to .97, with a mean of .91. Corrected for attenuation, these coefficients would be very close to one. These data suggest that the predicted outputs are essentially linear transformations of each other.

RECOMMENDATIONS AND CONCLUSIONS

Residuals can be used to compare the performance of individuals or groups while holding the initial status variable(s) mathematically constant if the data meets the assumptions of a multivariate normal distribution. If the data contain errors, the raw-score residual can be used to estimate the true-score residual. It is not necessary to multiply the raw-score residuals by k, the square of the correlation between the raw-score residual and the true-score residual, unless the researcher is interested in estimating the magnitude as well as the order of the true-score residuals.

Group means should be used to compute residuals for comparing groups. It is preferable that all groups have the same sample size; otherwise the residuals may vary greatly in their reliability and variance.

The parallel-forms reliability of raw-score residuals and other estimated t:.e scores is not necessarily equal to the index reliability,

[9] Dr. Clarence H. Bradford suggested this analysis.

the square of its correlation with the true score it estimates. For proper interpretation, both reliabilities should be reported. In order to avoid decisions based on chance differences, confidence intervals should be constructed for comparing raw-score residuals.

For statistical analyses of the determinants of change, partial correlations or multiple regression analysis should be used with final status as the criterion and initial status as one of the covariates. Multiple regression analysis is preferable because of its greater flexibility. Change scores, residuals, or base-free measures of change should not be used in statistical analyses; they will either give the same results as the above approach or results which are more difficult to interpret.

Errors in the predictor or initial status variables can change even the sign of partial correlation or multiple regression coefficients. If these coefficients are to be interpreted, they should be corrected for attenuation, either by correcting the individual correlations before entering them into the analysis, or by employing short-cut computational formulas. *If the only interest is predicting the criterion*, multiple regression coefficients should not be corrected for attenuation; the corrected equation would have a lower multiple correlation (i.e., be a poor predictor). (Johnston, 1963, pp. 162-164).

It is strongly recommended that test-retest estimates of reliability be used to correct coefficients for attenuation. Since the initial and final status measures will normally be collected at different times, an estimate of short term stability of these measures is more appropriate than an estimate of their internal consistency. All reliability coefficients should be derived from the population under study.

REFERENCES

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

Cronbach, L. J., & Furby, L. How we should measure "change"--or should we? *Psychological Bulletin*, 1970, *71*, 68-80.

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

Dyer, H. S., Linn, R. L., & Patton, M. J. Feasibility study of educational performance indicators. Princeton: Education Testing Service, 1967.

Dyer, H. S., Linn, R. L., & Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*, 1969, *6*, 591-605.

Garside, R. F. The regression of gains upon initial scores. *Psychometrika*, 1956, *21*, 67-77.

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

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

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

Harris, G. W. (Ed.) *Problems in measuring change*. Madison: University of Wisconsin, 1963.

Johnston, J. *Econometric methods*. New York: McGraw-Hill, 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. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

Manning, W. H., & DuBois, P. H. Gain in proficiency as a criterion in test validation. *Journal of Applied Psychology*, 1958, *42*, 191-194.

Manning, W. H., & DuBois, P. H. Correlational methods in research on human learning. *Perceptual and Motor Skills*, 1962, *15*, 187-321.

McNemar, Q. On growth measurement. *Educational and Psychological Measurement*, 1958, *18*, 47-55.

Mehrens, W. A., & Lehmann, I. J. *Standardized tests in education.* New York: Holt, Rinehart, and Winston, 1969.

Mood, A. M., & Graybill, F. A. *Introduction to the theory of statistics.* (2nd ed.) New York: McGraw-Hill, 1963.

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

Thomson, G. H. A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. *Journal of Experimental Psychology*, 1924, *7*, 321-324.

Thomson, G. H. An alternative formula for the true correlation of initial values with gains. *Journal of Experimental Psychology*, 1925, *8*, 323-324.

Thorndike, E. L. The influence of chance imperfections of measures upon the relationship of initial score to gain or loss. *Journal of Experimental Psychology*, 1924, *7*, 225-232.

Thorndike, E. L., Bregman, E. O., Tilton, J. W., & Woodyard, E. *Adult learning.* New York: MacMillan, 1928.

Thorndike, R. L. Intellectual status and intellectual growth. *Journal of Educational Psychology*, 1966, *57*, 121-127.

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

Traub, R. E. Comment on Glass' response. *Journal of Educational Measurement*, 1968, *5*, 343-345.

Tucker, L. R., Damarin, F., & Messick, S. A base free measure of change. *Psychometrika*, 1966, *31*, 457-473.

Woodrow, H. The ability to learn. *Psychological Review*, 1946, *53*, 147-158.

Zieve, L. Note on the correlation of initial scores with gain. Journal of Educational Psychology, 1940, 31, 391-394.

Zimmerman, D. W., & Williams, R. H. Effect of chance success due to guessing on error of measurement in multiple-choice tests. Psychological Reports, 1965, 16, 1193-1196. (a)

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

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

APPENDIX

TABLE 1

THE PARALLEL-FORMS RELIABILITIES AND THE INTERCORRELATIONS OF METHOD-M RESIDUALS

	V	R	L	W	A	C
Vocabulary (V)	(.77)	.73	.53	.53	.44	.71
Reading (R)		(.80)	.64	.79	.64	.87
Language (L)			(.86)	.58	.40	.72
Work Study Skills (W)				(.81)	.71	.89
Arithmetic (A)					(.80)	.77
Composite (C)						(.87)

TABLE 2

THE PARALLEL-FORMS RELIABILITY OF THE INTRA-SCHOOL-SYSTEM DIFFERENCE BETWEEN RESIDUALS

	R	L	W	A	C
Vocabulary (V)	.30	.61	.66	.62	.38
Reading (R)		.53	.32	.44	0*
Language (L)			.73	.72	.52
Work Study Skills (W)				.50	0*
Arithmetic (A)					.28
Composite (C)					

*Because of sampling fluctuations, the obtained estimate was negative, a zero has been substituted since common practice permits only zero or positive reliabilities.

TABLE 3

THE INTER-CORRELATIONS OF THE

METHOD-M PREDICTED OUTPUTS

	R	L	W	A	C
Vocabulary (V)	.97	.86	.91	.86	.94
Reading (R)		.88	.96	.88	.96
Language (L)			.90	.84	.91
Work Study Skills (W)				.92	.97
Arithmetic (A)					.96
Composite (C)					

This publication is published pursuant to a contract with the U.S. Office of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official U.S.O.E. position or policy.