Are Simple Gain Scores Obsolete?

Richard H. Williams and Donald W. Zimmerman

Applied Psychological Measurement 1996; 20; 59
DOI: 10.1177/014662169602000106

Are Simple Gain Scores Obsolete?

Richard H. Williams, University of Miami
Donald W. Zimmerman, Carleton University

It is widely believed that measures of gain, growth, or change, expressed as simple differences between pretest and posttest scores, are inherently unreliable. It is also believed that gain scores lack predictive validity with respect to other criteria. However, these conclusions are based on misleading assumptions about the values of parameters in familiar equations in classical test theory. The present paper examines modified equations for the validity and reliability of difference scores that describe applied testing situations more realistically and reveal that simple gain scores can be more useful in research than commonly believed. *Index terms: change scores, difference scores, gain scores, measurement of growth, reliability, test theory, validity.*

Over a quarter century ago, Cronbach & Furby (1970) and many other authors (e.g., Gulliksen, 1950; Lord & Novick, 1968) concluded that simple differences between pretest and posttest scores have questionable value in behavioral and social science research. Yet this conclusion seems incompatible with the intuition of researchers in many disciplines who assume that measures of gains, changes, differences, growth, and the like are meaningful in experimentation, program evaluation, educational accountability studies, and the investigation of developmental growth and change.

During the past two decades, many researchers using difference (or gain) scores have had difficulty justifying the use of such measures, even when they appear to yield interesting and reproducible findings. However, recent research on this topic has provided results favorable to simple gain scores (e.g., Collins & Cliff, 1990; Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991; Rogosa & Willett, 1983; Willett, 1989; Williams, Zimmerman, Rich, & Steed, 1984a, 1984b; Wittman, 1988; Zimmerman & Williams, 1982a, 1982b). There is also a rather large and somewhat controversial literature on the relation between the reliability of difference scores and the power of inferential tests of significance based on difference scores (e.g., Humphreys, 1993; Zimmerman, Williams, & Zumbo, 1993a, 1993b).

This paper examines this topic anew and demonstrates that two equations for the reliability of differences in classical test theory (CTT) have been widely misinterpreted. Furthermore, modified equations are derived that reveal that gain scores are more reliable than formerly believed.

Are Gain Scores Inherently Unreliable?

It might be assumed that the assertion that gain/difference scores are unreliable would be based on empirical studies designed to estimate the reliability of measured gains; however, there is a paucity of data-based investigations. Instead, most arguments have been based on theoretical and methodological considerations (Cronbach & Furby, 1970; Gulliksen, 1950; Lord & Novick, 1968). Typically these involve somewhat arbitrary assumptions about the values of parameters in well-known CTT equations.

The reliability of a difference ($\rho_{DD'}$), like the reliability of a single score, is the ratio of true score variance to observed score variance, or, alternatively, one minus the ratio of error score variance to observed score variance:

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp. 59–69

0146-6216/96/010059-11$1.80

$$\rho_{DD'} = \mathrm{Var}(T_D)/\mathrm{Var}(D) = 1 - \mathrm{Var}(E_D)/\mathrm{Var}(D), \quad (1)$$

where $D = Y - X$ is the difference between pretest (X) and posttest (Y) scores, and T_D and E_D denote the true and error components of D. Various equations for $\rho_{DD'}$ are derived from Equation 1, such as (Williams & Zimmerman, 1977):

$$\rho_{DD'} = \left(\lambda\rho_{XX'} + \lambda^{-1}\rho_{YY'} - 2\rho_{XY}\right)/\left(\lambda + \lambda^{-1} - 2\rho_{XY}\right), \quad (2)$$

where

ρ_{XY} is the product-moment correlation between X and Y,
$\rho_{XX'}$ is the reliability of X,
$\rho_{YY'}$ is the reliability of Y, and
the parameter λ is the ratio of the pretest standard deviation (SD), σ_X, and the posttest SD, σ_Y (i.e., σ_X/σ_Y).

This ratio turns out to be important to understanding the psychometric properties of gain scores. The derivation of Equation 2, like almost all equations for the reliability of gains in the CTT literature, depends on the assumption that the correlation between the pretest and posttest error scores (E_X and E_Y, respectively) is 0 [i.e., $\rho(E_X,E_Y) = 0$]. This assumption, which is sometimes called "experimental independence," is doubted by some theorists and researchers (e.g., Guttman, 1953; Rozeboom, 1966; Williams & Zimmerman, 1977; Zimmerman & Williams, 1977). If E_X and E_Y are positively correlated, as they are likely to be when two testing occasions are in close temporal proximity, then Equation 2 underestimates the reliability of gain scores.

Most textbook authors and others who have contended that gain scores are unreliable have not based their arguments on Equation 2 (Cronbach & Furby, 1970; Gulliksen, 1950; Lord & Novick, 1968; Thorndike, Cunningham, Thorndike, & Hagen, 1991), but instead have drawn conclusions from the following special cases of Equation 2:

$$\rho_{DD'} = \left[\tfrac{1}{2}(\rho_{XX'} + \rho_{YY'}) - \rho_{XY}\right]/(1 - \rho_{XY}) \quad (3)$$

or

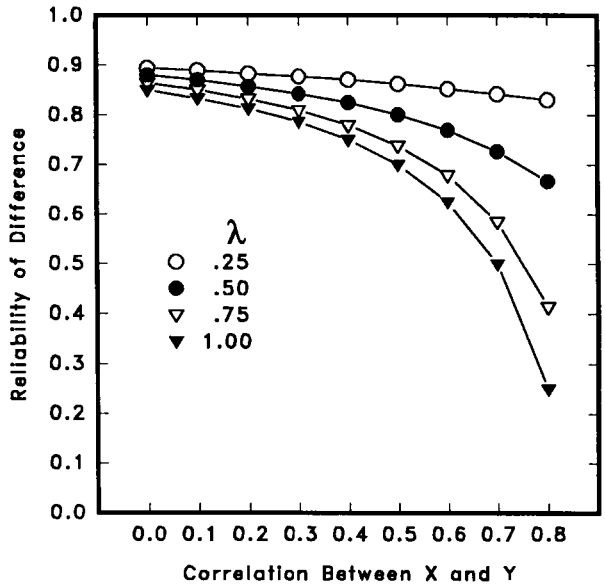
$$\rho_{DD'} = (\rho_{XX'} - \rho_{XY})/(1 - \rho_{XY}). \quad (4)$$

If $\sigma_X = \sigma_Y$, so that $\lambda = 1.00$, then Equation 2 reduces to Equation 3. If, in addition, $\rho_{XX'} = \rho_{YY'}$, then Equation 2 reduces to Equation 4.

These simplifying assumptions appear to be reasonable if the pretest and the posttest measures are assumed to be parallel in the usual CTT sense. However, this reasoning overlooks the fact that reliability and validity are not inherent characteristics of an educational or psychological test or measuring device, but are defined relative to the true score distribution of the population of examinees or experimental objects. That is, measuring instruments themselves do not inherently possess psychometric properties such as validity and reliability. Only the scores generated by administering the instruments to examinees possess these properties, and it is possible for scores on a particular test to be reliable for one population of examinees but unreliable for another population. These psychometric properties of scores may change as a consequence of the occurrence of gains or differences themselves.

Now assume that pretest and posttest measures are not parallel. That is, consider the reliability of a difference, given by Equation 2, under conditions in which the pretest and posttest measures possibly have unequal SDs and unequal reliabilities. Figure 1 shows the reliability of a difference, $\rho_{DD'}$, as a function of the correlation between pretest and posttest measures, ρ_{XY}, with the ratio of pretest and posttest SDs, λ, as a parameter. Figure 1 is based on calculations made from Equation 2 with $\rho_{XX'} = .80$ and $\rho_{YY'} = .90$. Figure 1 shows that reliability is lowest when $\lambda = 1.00$, and that it improves as λ decreases. It is well-known that the

Figure 1
The Reliability of a Difference, $\rho_{DD'}$, as a Function of the Correlation Between X and Y, ρ_{XY}, for Values of λ ($\rho_{XX'} = .80$ and $\rho_{YY'} = .90$)



reliability of a difference diminishes as ρ_{XY} increases, and Figure 1 is consistent with this notion.

Figure 1 also shows that the effect of ρ_{XY} on reliability is most potent when $\lambda = 1.00$ and becomes weaker as λ diminishes. Sharma & Gupta (1986) expressed this latter relation as follows: "When λ is not close to 1.00, $\rho_{DD'}$ becomes increasingly insensitive to variation in ρ_{XY}. In this situation, $\rho_{DD'}$ mainly depends on the values of λ, $\rho_{XX'}$, and $\rho_{YY'}$" (p. 108). Sharma and Gupta also used an optimization technique to prove that when $\rho_{XX'} = \rho_{YY'}$, the reliability of a difference is minimized when $\lambda = 1.00$. In other words, the assumptions necessary to develop Equation 4, found in many textbooks (Gulliksen, 1950; Thorndike et al., 1991), are precisely those that show the reliability of differences in the most unfavorable light. Note that a plot similar to Figure 1 can be constructed for $\lambda > 1.00$ simply by interchanging the values of the two reliability coefficients in Equation 2 and replacing λ by λ^{-1}. The same argument then shows that inequality of SDs, combined with inequality of reliability coefficients, is associated with high $\rho_{DD'}$.

Mechanisms of Change and Psychometric Properties of Difference Scores

Textbooks and journal articles on tests, measurement, and evaluation that discuss gain scores disparagingly often present tables of numerical values of $\rho_{DD'}$ constructed by substituting values of $\rho_{XX'}$ and ρ_{XY} in Equation 3 or Equation 4 (e.g., Gulliksen, 1950; Linn & Slinde, 1977; Thorndike et al., 1991). As noted, both Equations 3 and 4 are based on the assumption that σ_X and σ_Y are equal, but this assumption is misleading, particularly when gain scores rather than difference scores constructed from score profiles are considered.

True scores, as well as observed scores, can be expected to change as a result of an intervention. Moreover, the variability of scores frequently changes along with the magnitude of true scores and observed scores. If an intervention is effective and if the measuring device is capable of detecting change, then persons may be affected differentially, so that σ_X and σ_Y may differ.

For example, if the distribution of scores on a pretest is highly skewed, an experimental intervention might produce a posttest distribution that is more symmetrical and has a larger SD (and hence $\lambda < 1.00$).

Experiments conducted by educational psychologists designed to assess the impact of various instructional techniques on measured achievement are likely to produce such distributions, because most people participating in such studies usually have little knowledge of the subject matter prior to the experimental treatments (e.g., Williams, Zimmerman, Rich, & Steed, 1984a).

However, it is possible that pretest scores have a normal, or at least a unimodal symmetric distribution, so that an intervention produces negatively skewed posttest scores (in which case $\lambda > 1.00$). For this situation, it is also possible that the measuring instrument has a ceiling effect for the group of people tested, truncating the score distribution at the higher score levels.

For these reasons, it is quite restrictive to suppose that pretest and posttest scores are parallel according to the usual criteria for parallel measures. It is certainly doubtful that $\sigma_X = \sigma_Y$ in applied testing situations in which X and Y are scores before and after an experimental treatment or an intervention. Another pitfall to which measurement specialists and others succumb is converting from number-correct scores to standard scores (i.e., z scores) inappropriately. This type of transformation automatically produces a value of 1.00 for λ, even when pretest and posttest SDs are vastly different and, therefore, diminishes the reliability estimate (Zimmerman & Williams, 1982a).

Some authors have expressed doubt about the equality of SDs in the context of measurements of growth. For example, Feldt & Brennan (1989) noted:

> For expository purposes, it is useful to consider the special case of $[\sigma_X = \sigma_Y]$. This case is more realistic in the profile setting than in the growth setting. In the former, scales are rarely plotted on a profile unless they have been scaled to a common mean and standard deviation. In the growth context, however, variability typically changes.... Thus the presumption of equal standard deviations is often contradicted by the empirical data. (p. 118)

However, despite these arguments and empirical evidence to support them, many introductory textbooks in tests, measurement, and evaluation continue to criticize the reliability of gain scores (Gall, Borg, & Gall, 1996; Thorndike et al., 1991).

These considerations are consistent with observations made by Wittman (1988):

> Fortunately, many researchers protested against the condemnation of change scores. One of the earliest protests against the Cronbach and Furby (1970) verdict of difference scores was presented by Nesselroade and Cable (1974) and by Nesselroade and Bartsch (1977). Williams and Zimmerman (1977) concentrated on the ratio of the standard deviation before and after in their formulas for difference score reliability, thus showing how reliable difference scores can be in these situations.... The most recent comprehensive contribution to rehabilitate measurement of change with respect to difference score reliability was given by Rogosa, Brandt, and Zimowski (1982). (p. 554)

This formulation is also consistent with remarks made by Willett:

> ...the difference score was criticized for its (purported) inability to be both valid and reliable, for its negative correlation with initial status, and for its low reliability.... However, recent methodological work has shown these objections to be largely founded in misconception. The difference score has been demonstrated to be an intuitive, unbiased, and computationally simple measure of individual growth. (p. 588)

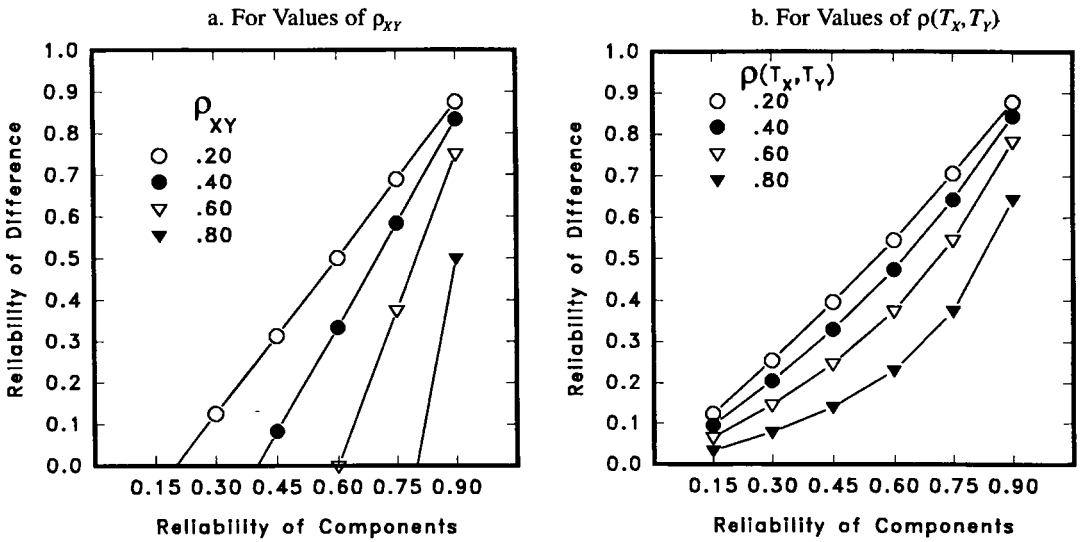
It is apparent, therefore, that a number of authors have recognized the inadequacy of the usual textbook approach. The equations presented here show their reservations to be well founded and reveal explicitly how misleading assumptions can influence calculations.

Gain Score Reliability As A Composite Function of Reliability of Components

Another subtlety regarding Equations 3 and 4 has been overlooked until recently (Zimmerman, 1994). Equation 3 appears to show $\rho_{DD'}$ as a function of $\rho_{XX'}$, $\rho_{YY'}$, and ρ_{XY}. The numerous tables in textbooks

mentioned above are based on this interpretation (Gulliksen, 1950; Thorndike et al., 1991). For example, in Figure 2a the reliability of a difference ($\rho_{DD'}$) is plotted as a function of the reliability of components ($\rho_{XX'}$) with the correlation between pretest and posttest scores (ρ_{XY}) as the parameter. The functions were obtained by substituting values into Equation 4.

Figure 2
The Reliability of a Difference, $\rho_{DD'}$, as a Function of the Reliability of Components, $\rho_{XX'}$



Functions such as these appear to substantiate the views of test theorists who maintain that gain scores are extremely unreliable. In this kind of analysis, however, it is important to be clear as to the question that is being asked and the assumptions that are made. Suppose the question is: "What changes occur in the reliability of a difference as the result of changes in the reliability of components?"

In the context of this question, it is natural to assume that an increase in the reliability of components is attributable to a decrease in error of measurement, and that the correlation between true scores on the separate measures, $\rho(T_X, T_Y)$, is held constant. Unfortunately these assumptions have not been explicitly stated. Because of the attenuation due to error of measurement, ρ_{XY} itself is a function of $\rho_{XX'}$ (Zimmerman, 1994). This fact is apparent from the well-known Spearman correction for attenuation (Nunnally & Bernstein, 1994, p. 241).

This means that the dependence of $\rho_{DD'}$ on $\rho_{XX'}$ is not of the form $Y = f(W,C)$, where W is a variable and C is a constant, but rather is a composite function of the form $Y = f[W, g(W)]$, where g is another function. Accordingly, it is more meaningful to express the reliability of a difference as a function of the reliability of components, with the correlation between true scores $[\rho(T_X, T_Y)]$ held constant as a parameter.

Figure 2b presents a family of functions of this type. These functions are obtained from

$$\rho_{DD'} = \left\{\rho_{XX'}\left[1-\rho(T_X, T_Y)\right]\right\} \Big/ \left[1-\rho_{XX'}\rho(T_X, T_Y)\right], \tag{5}$$

which is derived by substituting the value of ρ_{XY}, given by Spearman's correction for attenuation, for ρ_{XY} into Equation 4. Equation 5 now can be used to answer questions such as the following: "If the reliability of components is increased from .30 to .75, what increment in $\rho_{DD'}$ can be expected?" If $\rho(T_X, T_Y) = .20$, the

answer (as shown in Figure 2b) is that reliability increases from .25 to .70. Although Figures 2a and 2b are similar in that they are plotted on the same axes, Figure 2a cannot provide meaningful answers to questions regarding the effect of changes in the reliability of components.

Contrasting Figures 2a and 2b makes it apparent that the discrepancy between the reliability of a difference and the reliability of the components is less in Figure 2b than in Figure 2a. If textbooks and journal articles were to report values similar to those in Figure 2b, the negative view of the reliability of difference scores would be somewhat diminished.

The argument can be carried further by observing that Figure 1 does not take into consideration the fact that $\rho_{XX'}$ and $\rho_{YY'}$ influence ρ_{XY}, and that Figure 2b is based on the assumptions that $\lambda = 1.00$ and $\rho_{XX'} = \rho_{YY'}$. Hence, Figures 1 and 2b still show the reliability of difference scores in a somewhat unfavorable light. What is needed is an equation that expresses the reliability of a difference as a function of the reliability of components using both λ and $\rho(T_X, T_Y)$ as parameters. Such an equation can be obtained by substituting $\rho(T_X, T_Y)(\rho_{XX'}\rho_{YY'})^{1/2}$ in place of ρ_{XY} in Equation 2. The result is

$$\rho_{DD'} = \left[\lambda\rho_{XX'} + \lambda^{-1}\rho_{YY'} - 2\rho(T_X, T_Y)(\rho_{XX'}, \rho_{YY'})^{1/2}\right] \Big/ \left[\lambda + \lambda^{-1} - 2\rho(T_X, T_Y)(\rho_{XX'}, \rho_{YY'})^{1/2}\right]. \quad (6)$$

As mentioned above, all derivations have been based on the assumption that $\rho(E_X, E_Y) = 0$. A more general equation that does not involve this assumption is

$$\rho_{DD'} = \left[\lambda\rho_{XX'} + \lambda^{-1}\rho_{YY'} - 2\rho(T_X, T_Y)(\rho_{XX'}, \rho_{YY'})^{1/2}\right] \Big/ \left[\lambda + \lambda^{-1} - 2\rho(T_X, T_Y)(\rho_{XX'}, \rho_{YY'})^{1/2}\right] + \left\{\rho(E_X, E_Y)\left[(1 - \rho_{XX'})(1 - \rho_{YY'})\right]^{1/2}\right\} \Big/ \left[\lambda + \lambda^{-1} - 2\rho(T_X, T_Y)(\rho_{XX'}, \rho_{YY'})^{1/2}\right] \quad (7)$$

(Williams & Zimmerman, 1977; Zimmerman & Williams, 1982a). If $\rho(E_X, E_Y) = 0$, then Equation 7 reduces to Equation 6.

Figure 3 shows the reliability of a difference as a function of $\rho(T_X, T_Y)$, with λ as a parameter. In Figure 3, which is obtained from Equation 6, it is assumed that $\rho_{XX'} = .50$ and $\rho_{YY'} = .90$. Figure 3 again shows that the reliability of a difference is highest when $\lambda = .20$ and lowest when $\lambda = 1.00$. It is lowest when $\rho(T_X, T_Y)$ is high, but the effects of this correlation are less than the effects of ρ_{XY} (Figure 1), and are almost negligible for small values of λ. The striking reduction in the reliability of a difference that occurs in the right-hand part of Figure 1 is not present in Figure 3. For most data points in Figure 3, the values of $\rho_{DD'}$ are intermediate between $\rho_{XX'}$ and $\rho_{YY'}$, and for small values of λ they are quite close to .90, which was assumed for $\rho_{YY'}$.

Influence of Correlated Errors of Measurement on the Reliability of Differences

Figure 4 is based on Equation 7, which includes the correlation between error scores. The four functions represent values of $\rho(E_X, E_Y)$ of 0, .25, .50, and .75, respectively. In Figure 4, $\rho_{XX'} = .60$, $\rho_{YY'} = .80$, and $\lambda = .75$. It is apparent that reliability increases somewhat as $\rho(E_X, E_Y)$ increases. Although correlated errors have been generally neglected in CTT, there are strong reasons to believe that they exist in practical testing situations (e.g., Rozeboom, 1966; Williams & Zimmerman, 1977). This is another reason to suppose that the reliability of differences is higher than commonly believed.

This concept can be expressed in somewhat different terms, as follows. It is perhaps true that some of the random fluctuations comprising "error" occur independently on a pretest and posttest, and that error variances therefore are additive, as usually assumed. It is likely, however, that other random influences persist over time and modify pretest and posttest scores in a similar way. In other words, the couplet "pretest-posttest measurement," considered as a unit, may be subject to random error. If this is true, then the assumption of independence and additivity leads to an inflated value of the error variance associated with the difference score and a spurious underestimate of reliability.

Figure 3
The Reliability of a Difference, $\rho_{DD'}$, as a Function of the Correlation Between T_X and T_Y, $\rho(T_X,T_Y)$, for Values of λ ($\rho_{XX'}$ = .50 and $\rho_{YY'}$ = .90)

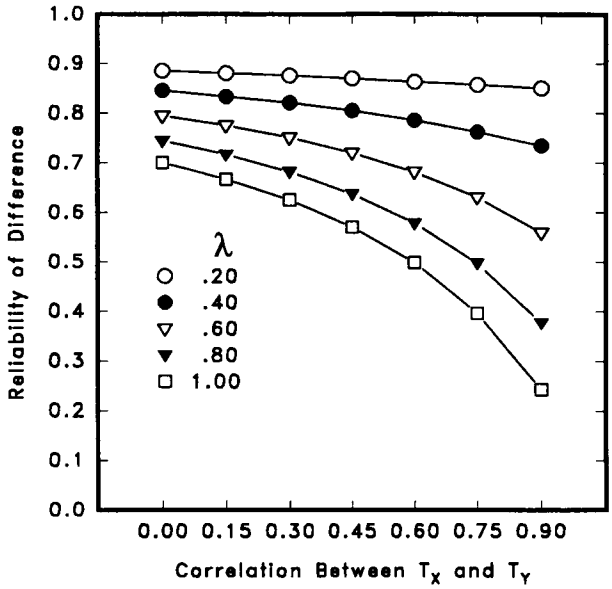
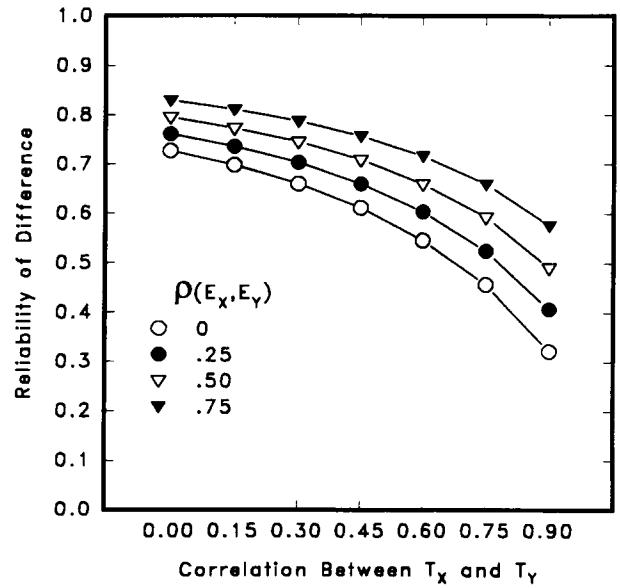


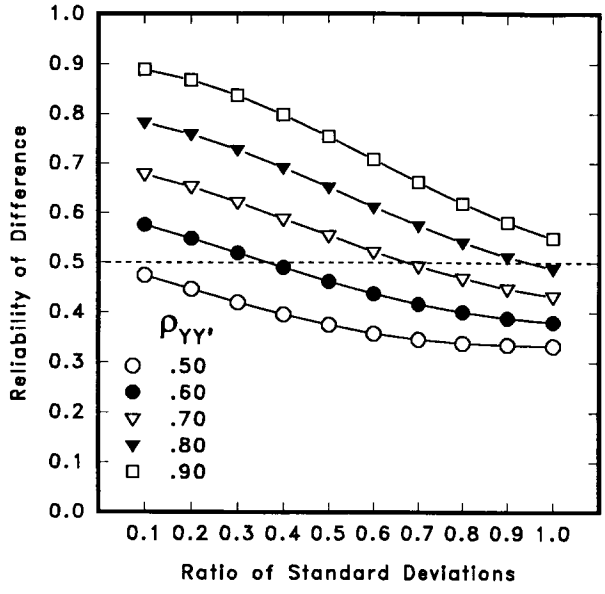
Figure 4
The Reliability of a Difference, $\rho_{DD'}$, as a Function of the Correlation Between T_X and T_Y, $\rho(T_X,T_Y)$, for Values of $\rho(E_X,E_Y)$



How Prevalent Are Reliable Differences?

The reliability of difference scores can be examined from still another perspective. Figure 5 is plotted from Equation 6; however, in Figure 5, λ is on the horizontal axis, $\rho_{XX'}$ is fixed at .50, $\rho(T_X, T_Y) = .50$, and $\rho_{YY'}$ ranges from .50 to .90 in increments of .10. These functions make it clear that values of $\lambda < 1.00$ combined with inequality of reliability coefficients are associated with high values of $\rho_{DD'}$. For example, when $\lambda = .1$ and $\rho_{YY'} = .9$, $\rho_{DD'} = .89$.

Figure 5
The Reliability of a Difference, $\rho_{DD'}$, as a Function of the Ratio of Standard Deviations, λ, for Values of $\rho_{YY'}$ [$\rho(T_X, T_Y) = .50$ and $\rho_{XX'} = .50$]



CTT equations, including Equations 3 and 4, have restricted attention only to the situation when $\lambda = 1.00$. When the entire "space" represented by the graph is examined, however, the reliability of differences appears more respectable. Only entries below the horizontal dashed line are cases in which $\rho_{DD'}$ is less than both $\rho_{XX'}$ and $\rho_{YY'}$. For all cases above the dashed line, the reliability of a difference is intermediate between the reliabilities of the components.

Furthermore, as the separation between $\rho_{XX'}$ and $\rho_{YY'}$ increases, the reliability of the difference score increases. In practice, if the treatment that intervenes between a pretest and a posttest is potent, the two cases most likely to occur are: (1) $\sigma_X < \sigma_Y$ and $\rho_{XX'} < \rho_{YY'}$; and (2) $\sigma_X > \sigma_Y$ and $\rho_{XX'} > \rho_{YY'}$. The first case is depicted in Figure 5.

Another way to examine the reliability of gain scores is

$$\rho_{DD'} = \left[\theta + \theta^{-1} - 2\rho(T_X, T_Y)\right] \Big/ \left[\left(\theta/\rho_{XX'}\right) + \left(\theta^{-1}/\rho_{YY'}\right) - 2\rho(T_X, T_Y)\right]. \quad (8)$$

Here θ, analogous to λ in Equations 2 and 6, is defined as $\sigma_{T(X)}/\sigma_{T(Y)}$. Table 1 displays values of $\rho_{DD'}$ as a function of the reliabilities of the components ($\rho_{XX'}$ and $\rho_{YY'}$), $\rho(T_X, T_Y)$, and θ. As the separation between $\rho_{XX'}$ and $\rho_{YY'}$ increases, $\rho_{DD'}$ improves. Just as with λ, $\rho_{DD'}$ is smallest when $\theta = 1.00$; also, it increases as θ increases.

Table 1
Reliability of a Difference, $\rho_{DD'}$, as a Function of Reliabilities of the Components, With θ and $\rho(T_X,T_Y)$ as Parameters

θ and $\rho_{XX'}$	$\rho(T_X,T_Y) = .20$				$\rho(T_X,T_Y) = .40$			
	$\rho_{YY'}$				$\rho_{YY'}$			
	.60	.70	.80	.90	.60	.70	.80	.90
$\theta = 1$								
.60	.55	.59	**.64**	**.67**	.47	.52	.57	**.61**
.70	.59	.65	**.70**	**.75**	.52	.58	.64	.69
.80	**.64**	**.70**	.76	**.82**	.57	.64	.71	.77
.90	**.67**	**.75**	**.82**	.88	**.61**	.69	.77	.84
$\theta = 2$								
.60	.56	.58	.59	**.60**	.51	.52	.54	.55
.70	**.64**	.66	.68	**.70**	.59	.61	.63	.65
.80	**.72**	**.75**	.77	.79	**.67**	**.70**	.73	.75
.90	**.80**	**.82**	**.86**	.88	**.75**	**.80**	**.83**	.86
$\theta = 4$								
.60	.58	.58	.58	.59	.55	.55	.56	.56
.70	**.67**	.68	.68	.69	**.65**	.65	.66	.66
.80	**.77**	**.78**	.78	.79	**.75**	**.76**	.77	.77
.90	**.86**	**.88**	**.88**	.89	**.85**	**.86**	**.87**	.88

Note. Boldface entries indicate that the reliability of the difference is intermediate between those of the components.

$\rho_{DD'}$ increases as the reliability of the components increases, but grows smaller as $\rho(T_X, T_Y)$ increases. The boldface entries in Table 1 represent cases in which $\rho_{DD'}$ is intermediate between the reliability coefficients of the components, in conformity with intuition. The other entries in Table 1 represent cases in which $\rho_{DD'}$ is less than that of both components, but in many cases not much less. This suggests that simple gain/difference scores are indeed useful.

The arguments presented here are not intended to suggest that simple difference scores are always, or even usually, reliable. Whether or not a test score is reliable in practice depends on the test construction procedure and the nature of the instrument, and in this respect a difference between scores is similar. It is undoubtedly true that many gain scores and difference scores are unreliable. The derivations presented here imply, however, that reliable differences cannot be ruled out solely by virtue of the statistical properties of measures of this type indicated by CTT equations. A function of two test scores is not unreliable just because that function is a difference.

The Validity of Difference Scores

Arguments similar to those presented above also reveal that the predictor-criterion validity of difference scores and gain scores can be higher than formerly believed. This conclusion is based on

$$\rho(Y-X,Z) = \left(\lambda^{-1/2}\rho_{YZ} - \lambda^{1/2}\rho_{XZ}\right) \Big/ \left(\lambda + \lambda^{-1} - 2\rho_{XY}\right)^{1/2}, \tag{9}$$

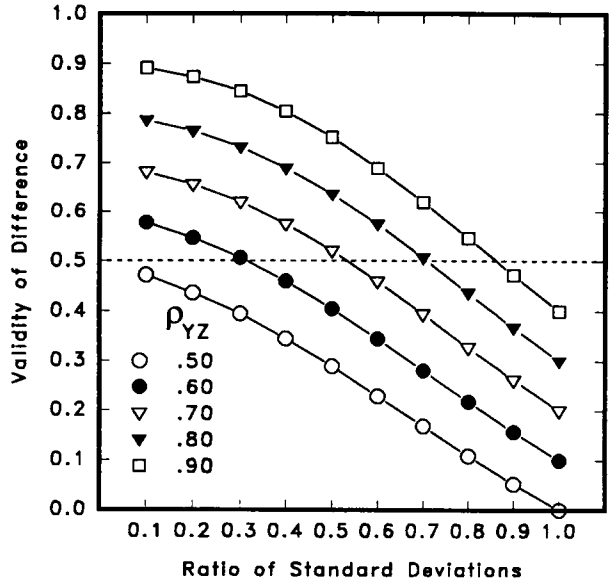
where $Y - X$ denotes a gain score and Z is an arbitrary criterion. λ is defined as above.

In the context of validity, like that of reliability, test theorists usually assume that pretest and posttest scores have equal variances. Furthermore, they believe that pretest and posttest scores have the same validity with respect to various criteria. That is, for a given Z, $\rho_{XZ} = \rho_{YZ}$. Under these conditions, values of $\rho(Y - X,Z)$ are typically rather low.

However, if $\rho_{YZ} > \rho_{XZ}$ and $\lambda < 1.00$, then the validity of $Y - X$ with respect to Z, $\rho(Y - X,Z)$, can be quite

high. A similar conclusion holds if $\rho_{YZ} < \rho_{XZ}$ and $\lambda > 1.00$ (Zimmerman & Williams, 1982b). Figure 6 displays the first case. In constructing Figure 6, it was assumed that $\rho_{XZ} = \rho_{XY} = .50$. These relationships have been investigated extensively by Gupta, Srivastava, & Sharma (1988) who derived conditions under which the validity coefficient has a maximum value. Once again, the results of these calculations contradict the assumptions usually selected by textbook authors for illustrations, which present the psychometric properties of differences in an unfavorable light. Figure 6 shows that when $\lambda = .2$ and $\rho_{YZ} = .7$, $\rho_{DZ} = .65$.

Figure 6
The Validity of a Difference, ρ_{DZ}, With Respect to a Criterion as a Function of the Ratio of Standard Deviations, λ, for Values of ρ_{YZ} ($\rho_{XZ} = \rho_{XY} = .50$ and $\rho_{XX'} = .50$)



Again, it should be emphasized that these arguments do not imply that gain scores in practice are highly correlated with various criteria. Historically, it has been difficult to discover measures that correlate highly with differences between test scores. As in the case of reliability, however, this situation characterizes instruments that are currently available, and the existence of valid difference scores cannot be ruled out by statistical arguments alone.

References

Collins, L. M., & Cliff, N. (1990). Using the longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bulletin, 108,* 128–134.

Cronbach, L. J., & Furby, L. (1970). How we should measure change—or should we? *Psychological Bulletin, 74,* 68–80.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 105–146). New York: Macmillan.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains NY: Longman.

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

Gupta, J. K., Srivastava, A. B. L., & Sharma, K. K. (1988). On the optimum predictive potential of change measures. *Journal of Experimental Education, 56,* 124–128.

Guttman, L. (1953). Reliability formulas that do not assume experimental independence. *Psychometrika, 18,* 123–130.

Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement, 17,* 11–14.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47,* 121–150.

Llabre, M. M., Spitzer, S. B., Saab, P. G., Ironson, G. H., & Schneiderman, N. (1991). The reliability and specificity of delta versus residualized change as measures of cardiovascular reactivity to behavioral challenges. *Psychophysiology, 28,* 701–712.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Nesselroade, J. R., & Bartsch, T. W. (1977). Multivariate perspectives on the construct validity of the trait-state distinction. In R. B. Cattell & R. M. Dreger (Eds.), *Handbook of modern personality theory* (pp. 221–238). Washington DC: Hemisphere.

Nesselroade, J. R., & Cable, D. G. (1974). "Sometimes it's okay to factor difference scores"—The separation of state and trait anxiety. *Multivariate Behavioral Research, 9,* 273–282.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Rogosa, D., & Willett, J. B. (1983). Demonstrating the reliability of the difference score. *Journal of Educational Measurement, 20,* 335–343.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood IL: Dorsey Press.

Sharma, K. K., & Gupta, J. K. (1986). Optimum reliability of gain scores. *Journal of Experimental Education, 54,* 105–108.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: Macmillan.

Willett, J. B. (1989). Some results on reliability for the longitudinal measure of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49,* 587–602.

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement, 77,* 679–689.

Williams, R. H., Zimmerman, D. W., Rich, J. M., & Steed, J. L. (1984a). An empirical study of the relative error magnitude of three measures of change. *Journal of Experimental Education, 53,* 55–57.

Williams, R. H., Zimmerman, D. W., Rich, J. M., & Steed, J. L. (1984b). Empirical estimates of the validity of four measures of change. *Perceptual and Motor Skills, 58,* 891–896.

Wittman, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.; pp. 505–560). New York: Plenum.

Zimmerman, D. W. (1994). A note on interpretation of formulas for the reliability of differences. *Journal of Educational Measurement, 31,* 143–147.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology, 16,* 135–152.

Zimmerman, D. W., & Williams, R. H. (1982a). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19,* 149–154.

Zimmerman, D. W., & Williams, R. H. (1982b). On the high predictive potential of change and growth measures. *Educational and Psychological Measurement, 42,* 961–968.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993a). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17,* 1–9.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993b). Reliability, power, functions, and relations: A reply to Humphreys. *Applied Psychological Measurement, 17,* 15–16.

Author's Address

Send requests for reprints or further information to Richard H. Williams, Department of Educational and Psychological Studies, University of Miami, P.O. Box 248065, Coral Gables FL 33124, U.S.A. E-mail: rwilliams@umiami.ir.miami.edu.