To Contact the author see

Email: bruno.zumbo@ubc.ca

Webpage: http://educ.ubc.ca/faculty/zumbo/zumbo.html

Best wishes,
Bruno.

THE SIMPLE DIFFERENCE SCORE AS AN INHERENTLY POOR MEASURE OF CHANGE

SOME REALITY, MUCH MYTHOLOGY

Bruno D. Zumbo

INTRODUCTION

Change is a fundamental concept in the social and behavioral sciences. For some areas, such as developmental psychology or areas involving program evaluation, change is a central aspect of study. In other areas, change may not be the central aspect of study, but it can still be of concern. Research in educational, cognitive, biological, clinical, or social psychology examines change whenever treatments are compared to a control group or to a base rate.

Historically, the life and physical sciences have not found the measurement of change as contentious as have the behavioral sciences, perhaps due to the assumed greater reliability of the dependent measures used. There exists, however, a vast technical literature on system dynamics. dynamic modeling, and time series analysis in the quantitative applied sciences, especially statistics, physical

Advances in Social Science Methodology, Volume 5, pages 269-304.

ISBN: 0-7623-0262-3

engineering, physics, and economics (cf. T.W. Anderson, 1971; O.D. Anderson. 1975; Brillinger, 1975; Hannan, 1970; Holtzman, 1970; Naslin, 1965; Winderknecht, 1971). However, the cogency of this material for the behavioral and social sciences seems limited to theoretical heuristics, for so far as this author is aware it requires essentially error-free observations on all relevant variables. A recent exception, however, is the use of survival analysis (Singer & Willett, 1992; Willett & Singer, 1988, 1991a, 1991b).

But in behavioral research, unlike the life and physical sciences, we seldom know in advance just what factors govern the phenomenon under study, nor how to accurately measure even those that we do suspect. Therefore, the measurement and analysis of change has been a persistent problem for social/behavioral science statisticians, psychometricians, and empirical researchers for more than 60 years. Within the behavioral sciences proper, three distinct, albeit interrelated, traditions in the analysis of temporally ordered data can be discerned: the psychometric, the structural or causal modeling, and the factorial approaches.

The *psychometric* tradition, which is by far the oldest of the three approaches, is well represented in the literature by Bereiter (1963), Cronbach and Furby (1970), Lord (1963), O'Connor (1972), Rogosa and Willett (1983), Werts and Linn (1970), Williams and Zimmerman (1996a, 1996b), Zimmerman (1997) and Zimmerman and Williams (1982a, 1982b, 1998). Its main concern, modest but basic, has been to estimate true change on a given variable through observed change, with technically proper regard for the unreliability of measurement.

The *structural or causal modeling* tradition is my proposed label for the spectrum of process models presenting patterns of change within latent variable systems (Loehlin, 1987; McDonald, 1985). Its main concern is how selected hypotheses about patterns of change can be investigated using concepts derived from linear structural equation modeling. Fundamentally, the concern is with the concept of structuring correlations with a strongly hypothetico-deductive outlook that conjectures causal systems. These are then fit to data and tested by structural equation methods.

Rogosa and his colleagues (e.g., Rogosa, 1980, 1985, 1987; Rogosa & Willett, 1985) have introduced a modicum of cold caution in using structural equation models in the measurement of change. They have demonstrated that the between-wave correlation matrix is not informative for the modeling of change. Keeping in mind Rogosa's warnings (particularly concerning "simplex" correlation matrices) some very promising methods and program of research has emerged from (a) an interaction between classical test theory and the structural equation modeling approach (Raykov, 1991, 1992a, 1992b, 1992c, 1993a, 1993b, 1994) (b) using covariance structure analysis to detect correlates of individual change (Willett & Sayer, 1994), and (c) the combination of latent curve and multilevel models (MacCallun, Kim, Malarkey, & Kiecolt-Galser, 1997).

Finally, multivariate specialists of the *factor analytic* flavor have weighed in with a handful of approaches to process multivariate temporal data, all of which

(Cattell, 1966; Corballis & Traub, 1970; Harris, 1963; Rozeboom, 1978a, 1978b; Tucker, 1963) are now subsumed under the title of longitudinal factor analysis (LFA). Recent advances in LFA can be found in Eid (1996), Millsap and Meredith (1988), Molenaar (1985), Swaminathan (1984), and Tisak and Meredith (1989). The LFA tradition differs from the structural modeling tradition, in part, because LFA emphasizes an inductive (*a posteriori*) disclosure of source structure (i.e., an exploratory) rather than a confirmatory approach.

The psychometric, structural and causal modeling, and factor analytic traditions summarize the approaches to the analysis of change in the social and behavioral sciences. However, there is such a vast amount of published literature on the analysis of change (even ignoring the research in the physical and life sciences), and so little concensus among authors, that a comprehensive treatment of the topic would be virtually impossible. As a result, this review has been restricted in scope to that of the oldest and most commonly encountered approach: the psychometric tradition.

Although there are many ways to measure change, the most common technique within the psychometric tradition is to calculate a gain or change score. It is the analysis and interpretation of this simple measure that has caused many of the debates and so much lack of consensus among methodologists and empirical researchers on how to proceed with the measurement of change. Cronbach and Furby (1970) in their widely cited paper on change, concluded "...that gain scores are rarely useful, no matter how they may be adjusted or refined," and, "It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways" (p. 80). As Bryk and Raudenbush (1987) asserted, we still are not able to accurately measure the true relationship between change and initial status (a fundamental issue in the measurement of change) and that the correct methodology for the measurement of change in two-wave designs is still elusive.

In the nearly 30 years since Cronbach and Furby's paper, the change score has become the most maligned index of change. Cronbach and Furby created such doubts about the difference score that granting agencies, journal editors, and even the committees of graduate students' theses have been known to deplore the use of change scores (Cattell, 1982). What Cattell perceived over a decade ago is true even now. As Malgady and Colon-Malgady (1991) stated, the measurement of change still remains an enigma and little has changed since Cronbach and Furby's indictment of the change score.

As a note, this frequent avoidance of the difference score is somewhat puzzling given that commonly accepted statistical tests (such as the test for interaction in the test split-plot design) involve difference scores. Maxwell and Howard (1981) respond to the recommendation to use a split-plot ANOVA rather than an analysis of change scores by demonstrating that the two statistical tests are equivalent. That is, they proved that the *F* statistic for the difference score approach equals the *F* for the interaction in the split-plot ANOVA (also see, Huck & McLean, 1975).

The recent developments in the measurement of change via difference scores have been widely discussed in the specific domains of psychometrics. However, it is this author's opinion that the modern view of change scores has not reached a broader audience of social scientists. A didactic exposition on this topic is therefore needed. To this end, the measurement model for the analysis of change will be introduced, followed by a discussion of the supposed unfairness, unreliability, and invalidity of the simple difference score. Next a brief introduction to some alternatives to the difference score will be presented. Finally, because this is primarily a didactic review for the general social and behavioral scientist, specific recommendations for analyzing change for several common designs are presented. These recommendations are meant to reflect the current state of affairs regarding the use of difference scores.

Before we begin discussing the simple difference score, it is important to note that substantive theory holds precedent when selecting a change model. In data analysis, there is an overarching concern with model validity. That is, the most statistically sophisticated, optimal, and mathematically elegant solution is pointless if the solution incorrectly describes the phenomenon under study. This is highlighted in Cook and Campbell's (1979) recommendations for the analysis of fan-spread and change models. Furthermore, the importance of substantive theory can be clearly seen in Haertel and Wiley's (1990) paper, which foreshadowed the future of the measurement of change.

THE SIMPLE DIFFERENCE SCORE

The simple difference score can be obtained by subtracting the initial measure from the final measure for each individual. Such a measure is known by various authors as the simple difference, change, or gain score.

It should be noted that the indices of change discussed in this section were developed at a period in time when two-wave designs were very common in evaluation and psychological research. Presently, in the methodological literature multiple-wave designs are being favored. An early recommendation of multi-wave designs and a discussion of how the two-wave methodology is restrictive can be found in Nesselroade, Stigler, and Baltes (1980). Nonetheless (due to the cost of multiple-wave designs) the two-wave approach is still quite common.

The primary purpose of this section is to examine the unfairness, unreliability, and invalidity attributed to the simple difference score. This will be followed by a brief discussion of alternatives to the change score.

Discussions of indices of change have stayed almost exclusively within the classical test theory model (Lord & Novick, 1968) wherein an observed score, *X*, is the simple composite of a true score, *T*, and an error score, *e*. In what follows, "classical test theory" refers to the modern version based on the expected-value concepts of true and error score, as developed by Novick (1966). Although presented as an assumption in this paper, more precisely, the independence of T and e is a tautological result of this classical test model and formally not an assumption. This does not alter our conclusions. A recent discussion of the assessment of change within an item response theory context can be found in Embretson (1994) and within a generalizability theory context by Cardinet (1994). A more general characterization of test theory can be found in Steyer (1988, 1989) and Zimmerman (1975, 1976).

The index of reliability when T and e are assumed to be independent and the errors themselves are independent is

$$\rho(X) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} \tag{1}$$

where $\rho(X)$ denotes the reliability, σ_T^2 denotes the variance of the true score, and σ_e^2 denotes the variance of the error score. Still another manner in which to envision the index of reliability is the squared correlation between the true and observed scores.

The classical model for any given individual is then

$$X_{ip} = T_{ip} + e_{ip} \tag{2}$$

where the subscript i denotes the occasion of measurement and the p indexes individuals. It is important to note that the true score is indexed by occasion and individual. This highlights that not only could the true score differ for a population of individuals, but more importantly, that it can change with time. That is, the change in true score reflects the fact that the individual has changed from time 1 to time 2.

A common confounding problem when discussing change scores is that the assumptions are not explicitly stated. As Rogosa, Brandt, and Zimowski (1982) have noted, this has caused many misunderstandings about change scores. In this chapter, it is assumed that two-wave data are available from a sample of n individuals. We will furthermore assume that the errors arise independently from identical normal distributions with a mean of zero and a fixed variance, σ_e^2, for all i occasions, and p individuals. That is, the e_{ip} are independent, identically distributed, normal (Gaussian) random variables (i.e., NIID(0, σ_e^2)). Although the assumption of normally distributed errors is not necessary in the derivation of a reliability coefficient, it was included herein because as Zumbo (1994) and Lind and Zumbo (1993) asserted, normality is an often overlooked and lurking assumption in the estimation of classical test theory statistics. As Zumbo (1994) stated, formally the assumption of normality is not needed in the derivation of the reliability coefficient, but due to the fact that sample estimates of the reliability are based on variance and/or covariances normality comes into play.

The assumption of independent errors is made herein because (a) independence is traditionally assumed in papers that argue for the simple difference score as an inherently poor measure of change, and (b) this paper is targetted for the general social scientist. As has been stated by several authors (Rozeboom, 1966; Williams & Zimmerman, 1982; Zimmerman & Williams, 1982a, 1982b, 1982d), the assumption of independent errors is somewhat dubious in most settings and particularly so in the measurement of change. However, the correlation among the errors is, as of yet, generally unquantifiable, and so despite the *existence* of formulas involving correlated errors, these formulas cannot be implemented by empirical researchers. The exception to this is the use of covariance structure analysis (structural equation modeling) allowing for meaningful correlations among the errors (Higgins, Zumbo, & Hay, 1999).

It should also be noted that given Eq. (1) low reliability can result two ways. The first cause of unreliability may be poor *measurement precision*. That is, if, for illustrative purposes, one was to hold constant the true score variance, then one could see that a low reliability results from an increase in the error variance.

The second source of unreliability may be a *homogeneous sample*. That is, if one held constant the error variance, one would see that the low reliability could also result from decreased true score variance. This may also be explained as the effect that a restricted range of true scores would have on the squared correlation between true scores and observed scores. These sources of unreliability are important to note because they play an essential role in understanding the reliability of difference scores.

Consistent with this section's notation, the difference score is denoted as,

$$\Delta X_p = X_{2p} - X_{1p} \tag{3}$$

where ΔX_p is the observed change (difference or gain) score for individual p, and X_{2p} and X_{1p} are the observed scores as defined in Eq. (2). Furthermore, Eq. (2) can be expressed as the true score change model

$$\begin{aligned} \Delta X_p &= \Delta T_p + \Delta e_p, \text{ where} \\ \Delta T_p &= T_{2p} - T_{1p}, \text{ and} \\ \Delta e_p &= e_{2p} - e_{1p}. \end{aligned} \tag{4}$$

As Rogosa, Brandt, and Zimowski (1982) stated, given that we have made the assumption of NIID(0, σ_e^2), the error of the difference score, as defined above, is normally distributed with a mean of zero and a variance of $2\sigma_e^2$. This will be particularly relevant when deriving the reliability of the difference score.

Unfairness

Most researchers are interested in the true change, ΔT_p, rather than the observed change, ΔX_p. Following in this vein, Linn and Slinde (1977) and others have stated that the simple difference score is an unfair measure of change because it gives an advantage to individuals who have certain values at time 1 (usually those with high initial scores). However, the expected value of ΔX_p over its propensity distribution shows that it is an unbiased estimate of ΔT_p regardless of the value of Δe_p or the true score at time 1 (Rogosa et al., 1982). This indicates that under no circumstances is the difference score an unfair measure. Nonetheless, the difference score is still seen as inherently unfair.

How can an unbiased estimate be an unfair measure? As Rogosa and Willett (1983) and Rogosa et al. (1982) demonstrated, this confusion is bound up with the misunderstandings about the correlation between change and the initial value. At different times, various authors (e.g., Furby, 1973; Lord 1963) have argued that the correlation between the change and the value at time 1 *must* be negative. The negative correlation between change and initial status is the condition most commonly discussed in the very influential papers by Linn (1977), Cronbach and Furby (1970), and Lord (1963), therefore it will be treated in detail herein. A positive correlation is called the "law of initial values" in the neurological and physiological literature (Wilder, 1957). However, a positive correlation between change and initial status corresponds to a setting where the variances increase over time, called the fanspread. The fanspread is often discussed in the program evaluation literature. A zero correlation between change and initial status is the least often considered setting, however, see Bloom (1964) for discussion.

Irrespective of the unbiasedness property of the observed change, two reasons for the negative correlation are often given: (a) regression toward the mean, and (b) that a glance at Eq. (4) shows that Δe and e_1 have different signs. First, those authors who argued for the presence of regression toward the mean were in fact arguing for its ubiquitousness (Furby, 1973; Lord 1963). Unfortunately, regression toward the mean is an ill-defined and an often reified concept (e.g., Furby, 1973) that is not necessarily ubiquitous. Rogosa et al. (1982) have shown that the correlation between ΔX and X_1 is a poor estimate of the correlation between ΔT and T_1. In fact, over and above the problem of attenuation due to measurement error, the estimate of $\rho(\Delta T, T_1)$ by $\rho(\Delta X, X_1)$ is negatively biased. Therefore, it is because of this bias that negative correlations between observed change and the observed initial status are often obtained when the true score correlation is, in fact, zero or positive.

Because of the negative bias and the substantive interest in the true score, regression toward the mean is often discussed in terms of population values. In an attempt to formally define the regression effect, Healy and Goldstein (1978), Rogosa et al. (1982), and Rogosa and Willett (1985) defined regression toward the mean as

$$\frac{E[T_2|T_1 = C] - \mu_{T_2}}{\sigma_{T_2}} < \frac{C - \mu_{T_1}}{\sigma_{T_1}} \tag{5}$$

In the above expression C denotes a constant, E denotes the expectation operator, T_1 and T_2 are the true score at time 1 and 2, and μ_{T1}, μ_{T2}, σ_{T1} and σ_{T2} are the population true score values and standard deviations at time 1 and 2, respectively. In words, Eq. (5) simply states that given a time 1 true score value of C, the time 2 true score will be less standard units away from its mean than the time 1 score (C) is from its mean. This inequality is satisfied whenever the correlation between the true scores at times 1 and 2, $\rho(T_1, T_2)$, is less than one. Therefore, Healy and Goldstein argued that given that this correlation is expected to be less than one, the regression toward the mean is considered ubiquitous. However, Rogosa et al. (1982) recommended that Eq. (5) is best considered a harmless mathematical tautology and one that provides little insight for the study of change.

Furthermore, Rogosa et al., as well as Healy and Goldstein argued that a more realistic definition of regression toward the mean expressed in terms of the actual metric of the true score (rather than in units of the standard deviations, as in Eq. (5)) is

$$E[T_2 \mid T_1 = C] - \mu_{T2} < C - \mu_{T1} \tag{6}$$

Therefore, only if $\sigma_{T1} = \sigma_{T2}$, as is done in Lord (1963) and Furby (1973) is Eq. (5) equivalent to Eq. (6). However, more importantly, Eq. (6) is satisfied only when the correlation of change and initial status is less than zero. Therefore, regression toward the mean is not ubiquitous but rather exists when the correlation between change and initial status is negative, and this exists when the standard deviations (i.e., the metric) of the scores on the two occasions are set to equality (Rogosa et al., 1982).

Some authors have stated that the correlation between change and initial status tends to be negative because an error component appears with opposite signs in X_1 and ΔX (Lord & Novick, 1968; Thomson, 1924; Thorndike, 1924). Given the assumptions in this paper, Zimmerman and Williams (1982d) mathematically proved the following to be true:

1. $\rho(\Delta X, X_1) < 0$ if and only if $\sigma_{X1}/\sigma_{X2} > \rho(X_1, X_2)$,
2. $\rho(\Delta X, X_1) > 0$ if and only if $\sigma_{X2} > \sigma_{X1}$, and
3. $\rho(\Delta X, X_1) = 0$ if and only if $\sigma_{X1}/\sigma_{X2} = \rho(X_1, X_2)$.

In the above expressions $\rho(\Delta X, X_1)$, $\rho(X_1, X_2)$, σ_{X1}, and σ_{X2} denote the correlation between the change and initial status, the correlation between the time 1 and time 2 scores, and the standard deviations of the time 1 and time 2 observations, respectively. These results allow one to determine the correlation between gains and initial status from the statistics of pretest and posttest observed scores without any knowledge of the true scores, error scores, reliability coefficients and standard errors of measurement. Zimmerman and Williams' (1982d) findings can be translated into the following decision rules:

1. If researchers would like to know if they are in a context of negative correlation between change and initial values, they simply need to compare the ratio of the standard deviations to the pretest-posttest correlation. If $\sigma_{X1}/\sigma_{X2} > \rho(X_1, X_2)$, then there is negative correlation. If not, then the negative correlation is not true.
2. If researchers would like to know if change and initial value are positively correlated, then they need to compare the standard devaitions. If the time 2 standard deviation is greater than the time 1 standard deviation, then the positive correlation is evident. If not, then the positive correlation is not true.
3. If researchers would like to know if the change and initial values are not correlated, then they compare the ratio of the pretest-posttest standard deviations to the pretest-posttest correlations. If $\sigma_{X1}/\sigma_{X2} = \rho(X_1, X_2)$, then the correlation between change and intial status is zero.

These questions can be answered with the aid of the correlation coefficient's sampling theory (i.e., hypothesis testing). Therefore, by using the above three questions in a careful process of elimination researchers can determine whether they are in a regression toward the mean, fanspread, or zero correlation context. These findings are in agreement with Rogosa and his colleagues (1982) that the correlation of gain and initial status is dependent upon the ratio of the standard deviations of the pretest and posttest scores.

Finally, arguing from a continuous growth curve perspective, Rogosa et al. (1982) have demonstrated that for the two-wave design (i.e., straight-line growth), the correlation between change and time 1 scores depends crucially on the time at which initial status is measured, $t(i)$. Theoretically, for any straight-line growth the correlation between change and time 1 score is monotonically increasing, having a lower asymptote of -1.0 for $t(i) = -\infty$, passing through zero for a single $t(i)$, and increasing to an upper asymptote of 1.0 for $t(i) = \infty$. Therefore, for any family of linear growth curves a very different correlation between true change and initial status will be obtained, depending on whether the initial status to be chosen is later, earlier, or in between.

A final note on the issue of unfairness. The evidence presented thus far is derived from theoretical variables (or measures) that do not have an artificial ceiling or floor. This is important to note because the deleterious results of ceiling and/or floor effects are often confused with the issue of the difference score being inherently unfair. The results presented by Rogosa et al. (1982) and Zimmerman and Williams (1982d) are valid given that the variances of the measures are not artificially restricted. Ceiling and floor effects are characteristic of poor measures or measures that are being used inappropriately for the measurement of change. The simple difference score, like most statistical procedures, does not work well in the context where the measures are being used inappropriately. Thus, problems due to ceiling or floor effects should not be confused with arguments regarding the inherent unfairness of the simple difference score.

In summary, the ubiquitousness of the phenomenon of regression toward the mean is often used as a basis to argue the unfairness of the simple difference score. Upon formal scrutiny, the ubiquitousness of regression toward the mean does not stand. Rather, it was demonstrated that regression toward the mean is an artifact of standardizing (in fact, equalizing) the variances of the two measures. Furthermore, because Δe and e_1 have different signs the correlation was expected to be negative. The effect of the opposite signs was also shown to be dependent on the ratio of the standard deviations of the measures at time 1 and time 2. Rogosa et al. (1982), from a continuous growth perspective, showed that the correlation between change and initial status is dependent upon the time 1 value selected. Therefore, contrary to current view, the simple difference score *is* a fair measure of change, because it is an unbiased estimate of true change.

Unreliability of the Simple Change Score

Despite the obvious and optimal statistical property of unbiasedness, several authors have criticized the difference score for its supposed unreliability so thoroughly and continuously (e.g., Bereiter, 1963; Cohen & Cohen, 1983; Cronbach & Furby, 1970; Linn & Slinde, 1977) that empirical researchers have become hesitant to use it. However, recent developments have shown that the claimed deficiencies are not necessarily true (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983, 1985; Zimmerman, Brotohusodo, & Williams, 1981; Zimmerman & Williams, 1982a, 1982b; Zimmerman, Williams, & Zumbo, 1993a, 1993b).

Given Eq. (4) and the definition of reliability in Eq. (1), the reliability of the simple difference score is

$$\rho(\Delta X) = \frac{\sigma^2_{\Delta T}}{\sigma^2_{\Delta T} + \sigma^2_{\Delta e}} \qquad (7)$$

First, following the same argument as for Eq. (1), a low reliability of the simple difference score may be due to a small variance in true change rather than only a large error variance. Figure 1 helps motivate the discussion of variance in true change. The graph on the left side of Figure 1 represents homogeneous change where all the individuals changed the same amount from time 1 to time 2. It is clear that $\sigma^2_{\Delta T} = 0$ for the homogeneous case. The graph on the right represents the heterogeneous case where all the individuals did not change the same amount.

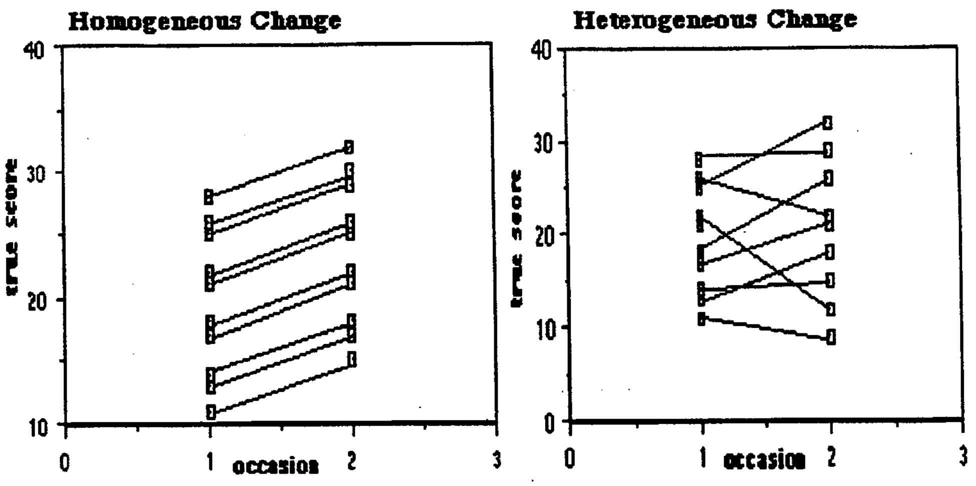


Figure 1. Scatterplots of homogeneous and heterogeneous change.

In this case $\sigma^2_{\Delta T}$ would be quite large. It is simple to see from Eq. (7) that the reliability of the simple change score would be zero for the case of homogeneous change. Furthermore, as was noted earlier, this is not an artifact of the simple change score, but rather is a property of reliability even with raw scores (see Eq. (1) and its discussion).

The essential point here is that the difference score, just like any other measure, is reliable when individual differences exist. The low reliability when individuals change at the same rate correctly indicates that one cannot detect individual differences that are not there. Quite frankly, if a variables were to be defined such that there is no true variability, then the measure would be of very little interest to empirical social and behavioral scientists.

Second, Eq. (7) is expressed in unobservable quantities, $\sigma^2_{\Delta T}$ and $\sigma^2_{\Delta e}$. Given that we have assumed NIID(0, σ^2_e) for the measures at time 1 and 2, Eq. (7) can be reexpressed as

$$\begin{aligned}\rho(\Delta X) &= \frac{\sigma^2_{\Delta T}}{\sigma^2_{\Delta X}} \\ &= \frac{\sigma^2_{X_1}\rho(X_1) + \sigma^2_{X_2}\rho(X_2) - 2\sigma_{X_1}\sigma_{X_2}\rho(X_1, X_2)}{\sigma^2_{X_1} + \sigma^2_{X_2} - 2\sigma_{X_1}\sigma_{X_2}\rho(X_1, X_2)},\end{aligned} \tag{8}$$

where σ^2_{X1}, σ_{X1}, σ^2_{X2}, σ_{X2} are the variance and standard deviation of the scores on occasions 1 and 2, respectively, $\rho(X_1)$ and $\rho(X_2)$ are the reliabilities of the measures on times 1 and 2, and $\rho(X_1, X_2)$ is the correlation between the measures on time 1 and 2. This expression can be found in various forms throughout the psychometric literature (e.g., Rogosa et al., 1982). [Reexpressions of Eq. (8) which relax our strict assumptions can be found in Williams and Zimmerman (1977), and Zimmerman and Williams (1982b).]

It has been demonstrated by Rogosa et al. (1982) and Rogosa and Willett (1982) that Eq. (8) or some reexpression of it has been used repeatedly in the literature to show that the difference score is much more unreliable than the scores themselves (e.g., Lord, 1956; Linn & Slinde, 1977). Traditional demonstrations of the unreliability (e.g., Lord, 1963) consider the values of $\rho(\Delta X)$ under the conditions where $\rho(X_1) = \rho(X_2) = \rho(X)$, $\sigma^2_{X1} = \sigma^2_{X2}$, and usually a large positive value of $\rho(X_1, X_2)$. These constraints imply that $\sigma^2_{T1} = \sigma^2_{T2}$ and $\rho(T_1, \Delta T) < 0$. The results of this have been discussed here previously with regards to regression toward the mean.

Furthermore, traditional considerations consider parameters of $\rho(X) = .70$, .80, or .90, and $\rho(X_1, X_2) = .50$, .60, .70, .80, or .90. In fact, in the case where both $\rho(X)$ and $\rho(X_1, X_2)$ equal .90 the reliability of the change score, $\rho(\Delta X)$, equals zero. Given this, it is not difficult to envision why the difference score was being strongly criticized. However, applying the standard disattenuation formula—$\rho(T_1, T_2) = \rho(X_1, X_2) / \rho(X)$—to the parameters listed above, one finds that the correlation between the true scores ranges from .56 to 1.0 where a majority of the correlations are well over .70. This crude analysis suggests that the parameter space investigated by traditional expositions (which show that difference score is unreliable) was too restricted. Furthermore, because most of the disattenuated correlations are quite large, the parameter space represents a homogeneous change setting.

Zimmerman, Williams, and Zumbo (1993a) presented the formula

$$\rho(\Delta X) = \frac{\rho(X) - \rho(X)\rho(T_1, T_2)}{1 - \rho(X)\rho(T_1, T_2)}, \tag{9}$$

as an expression for determining the reliability of the difference score with the correlation between the true scores at time 1 and 2 treated as a parameter. A table similar to Table 1 is reported in Zimmerman et al. (1993a). The entries in Table 1, determined by Eq. (9), are the reliabilities of the difference score as a function of the common reliability, $\rho(X)$, and the correlation between the true scores at time 1 and 2, $\rho(T_1, T_2)$. First, a cursory scanning of the entries down a column for any given value of $\rho(T_1, T_2)$ shows that the reliability of the difference score increases with increasing $\rho(X)$. Of course, as $\rho(X)$ changes, the $\rho(X_1, X_2)$ will change, although $\rho(X_1, X_2)$ is not a parameter in the equation. Second, and more importantly at this juncture, for any given value of $\rho(X)$ the reliability of the difference score decreases as $\rho(T_1, T_2)$ increases. Again, when $\rho(T_1, T_2)$ is a large positive

Table 1. Reliability of Differences with the Correlation Between True Scores at Time 1 and Time 2 as Parameters

	$\rho(T_1, T_2)$			
$\rho(X)$	.3	.5	.7	.9
.1	.07	.05	.03	.01
.2	.15	.11	.07	.02
.3	.23	.18	.11	.04
.4	.32	.25	.17	.06
.5	.41	.33	.23	.09
.6	.51	.43	.31	.13
.7	.62	.54	.41	.19
.8	.74	.67	.55	.29
.9	.86	.82	.73	.47

magnitude this represents a strong increasing linear relationship between the true scores at time 1 and 2. This large positive value of $\rho(T_1, T_2)$ represents a homogeneous change setting. As discussed previously, a reliability measure cannot detect individual differences that are not there.

Few authors have examined the reliability of the difference score if the true scores are negatively correlated. Therefore, Table 2 is a further exploration of the findings in Table 1 given that the correlation between the true scores is negative. As can be seen from Eq. (9), a negative correlation will result in an increase in the reliability of the difference score. This is noted in Table 2 where the entries of the table are all larger than the original reliabilities of the time 1 and 2 measures. Table 2 was created so as to allow comparison with the entries in Table 1. If one focuses on any given row of Table 2, it is evident that as the correlation becomes more negative the reliability increases. Thus, it is clearly advantageous to use the difference score when there is heterogeneous change, such as that evident when the $\rho(T_1, T_2)$ is negative and moderate in magnitude.

Table 2. Reliability of Differences with a Negative Correlation Between True Scores at Time 1 and Time 2

	$\rho(T_1, T_2)$			
$\rho(X)$	−.3	−.5	−.7	−.9
.1	.13	.14	.16	.17
.2	.25	.27	.30	.32
.3	.36	.39	.42	.45
.4	.46	.50	.53	.56
.5	.57	.60	.63	.66
.6	.66	.69	.72	.74
.7	.75	.78	.80	.82
.8	.84	.86	.87	.88
.9	.92	.93	.94	.94

It is interesting to note that when $\rho(T_1, T_2)$ is negative, this commonly represents decline rather than growth or gain. The results are quite different than when a positive correlation is present. The results in Table 2 suggest that when decline is present the simple difference score has more than adequate reliability. In fact, in all parameter values considered in Table 2, the reliability of the differences is greater than the reliability of the pretest or posttest measures. This is strikingly different than the findings in Table 1. In fact, *even* homogeneous decline has more than adequate reliability.

As a note of clarification, the above observations regarding the necessity of heterogeneous change when investigating the reliability of indices of change does not hold when developing assessment instruments sensitive to change. First, an assessment instrument sensitive to change is quite different than an index of change (such as the simple difference score). An assessment instrument sensitive to change is a test, questionnaire, or other measure that utilizes a scaling and measurement model wherein the scores reflect a change in the attribute of interest without the aid of change indices such as the simple change score (e.g., Collins & Cliff, 1990). In developing instruments sensitive to change a high test-retest reliability is often recommended (e.g., Rushton, Brainerd, & Pressley, 1983). However, Collins and Cliff stated that "Given the same degree of measurement precision, a population whose members all develop at the same rate will produce larger test-retest correlations than a population whose members develop at different rates" (pp. 129-130). Paradoxically, then, the test-retest correlation is not a good indicator of the precision (i.e., reliability) with which an instrument reflects change unless there is homogeneous change.

In summary, many authors have shown in a more thorough exploration of the parameter space than was first reported (e.g., Cronbach & Furby, 1970) and under less restrictive assumptions that the reliability of the difference score can be quite adequate (Rogosa et al., 1982; Rogosa & Willett, 1983; Zimmerman & Williams, 1982a, 1998, Zimmerman, Williams, & Zumbo, 1993a, 1993b).

Invalidity of the Simple Change Score

Bereiter (1963) and Linn and Slinde (1977) have criticized the simple difference score by arguing that it could not be both a reliable and valid measure of change at the same time. This is what Bereiter called the invalidity/unreliability dilemma. Rogosa et al. (1982) have demonstrated that there is no such dilemma; high reliability does not necessarily imply low validity, and low reliability does not necessarily mean lack of measurement precision (see the previous comments on homogeneous change).

Given the assumptions in traditional demonstrations of the unreliability (i.e., $\sigma^2_{X1} = \sigma^2_{X2}$, and $\rho(X_1) = \rho(X_2) = \rho(X)$), Eq. (8) can be easily simplified to

$$\rho(\Delta X) = \frac{\rho(X) - \rho(X_1, X_2)}{1 - \rho(X_1, X_2)}, \tag{10}$$

where $\rho(X)$ is the common reliability for measures on time 1 and 2, and $\rho(X_1, X_2)$ is the correlation across the two times. First, it can now be seen that if $\rho(X_1, X_2) > \rho(X)$, the $\rho(\Delta X)$ is a negative number. This is considered undefined because, by definition, the index of reliability is bounded by zero and one. As Zimmerman, Williams, and Zumbo (1993a) stated, these impossible results are accountable by virtue of the Cauchy-Schwartz inequality or likewise the correlation inequality (Bickel & Doksum, 1977). Both of these inequalities are equivalent to the statement that the absolute value of the correlation must be less than or equal to one, $|\rho(X_1, X_2)| \leq 1$. Second, if the correlation between the two occasions is a large positive number and/or close in magnitude to the common reliability, then $\rho(\Delta X)$ will necessarily be low. Hence, it appears that high validity (as measured by the correlation between measures on time 1 and 2) necessitates unreliability.

Beyond the fact that Eq. (10) is derived under very restricted conditions, Rogosa et al. (1982) and Rogosa, Floden, and Willett (1984) have stated clearly that the use of time 1, time 2 correlations as measures of construct validity is incorrect. Quite simply, a large positive value of $\rho(X_1, X_2)$ implies that the rank ordering of individuals is the same on the two occasions. It would not be surprising to find the rank ordering of individuals change drastically (i.e., a low value of $\rho(X_1, X_2)$) when measuring the same construct on two occasions (particularly when growth or change is occurring). This merely indicates that some individuals changed more than others on the construct being measured, as portrayed in Figure 1.

A large positive value of $\rho(X_1, X_2)$ does not necessarily imply that the same construct is being measured on the two occasions. For example, a group of individuals may be rank ordered the same way if memory capacity is measured on time 1 and age on time 2. The individuals happen to rank in the same order for both variables. This obviously does not imply that the same construct is being measured on both occasions. Therefore, the invalidity/unreliability dilemma is scarcely a dilemma at all (Rogosa et al., 1982) because it is found under very restricted conditions and it misrepresents $\rho(X_1, X_2)$ as an index of construct validity. Importantly, the whole process of the measurement of change assumes a quantitative change and not a qualitative change in the construct. Thus, if the same construct is not measured on occasion 1 and 2, we cannot address the question of change (Bereiter, 1963; Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1963).

Furthermore, the notion of validity has changed in recent years. Validity, as a psychometric resolution to observational indeterminacy, is currently defined as a unified concept focusing on construct validity with various types of evidential support. See Hubley and Zumbo (1996) for a review of the current views of validity theory and Zumbo (1998) for recent results. Messick (1989) has offered a reconceptualization of validity. He suggested reporting test validity in terms of two interconnected facets linking the source of the justification of the measurement instrument (either evidential or consequential) to the function of the testing (either interpretation or use). This articulation avoids reference to different types of validity (e.g., criterion, construct, and content) and the notion that any one type is sufficient evidence in support of a validity claim.

That is, the evidential basis of test interpretation requires evidence of construct validity. For test use, one must provide evidence of test relevance or utility as well as of construct validity. The consequential bases involve a consideration of the value implications for test interpretation, and the social consequences of test use. Hubley and Zumbo (1996) supported Messick's model and argued that many of the traditional aspects of validity and psychometric models (e.g., reliability, item and test bias, dimensionality) should be considered data collection and analysis strategies in support of the unified validity evaluation. Validity is no longer simply the correlation between time 1 and time 2. Thus, scores on a measure can be valid, yet lack this correlation.

In establishing the validity of change scores, Andrews (1983) introduced functional validity as an essential form of validity evidence. Briefly, functional validity is the unique increment in criterion-related validity (or using a more recent term, criterion-related evidence) attributed to the use of a change score (for more details, see Andrews (1983) and Andrews, Bonta, & Hoge (1990)). Andrews introduced functional validity in the context of wanting to predict behavior subsequent to an intervention or treatment (in his case, incarceration). He argued that there is an improved sampling of the predictor domain through the introduction of assessments of change (for a similar statement regarding emotional states and learning, see, Boyle, 1987). In this framework, the conceptual and practical significance of a change index is that it is theoretically capable of detecting real change not predictable from initial testing and it has the potential of establishing evidence for incremental predictive criterion validity. In using change scores a researcher should investigate the functional validity in contexts where retesting is not commonly utilized. Following Hubley and Zumbo (1996), the evidence for functional validity can be used to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from change scores.

In summary, then, many researchers have used Eqs. (8), (10), or their kin to argue that the simple difference score should not be used. As has been noted, this is true only when the following assumptions are true:

1. $\rho(X_1, X_2)$ has a large positive magnitude,
2. $\sigma^2_{X1} = \sigma^2_{X2}$,
3. $\sigma^2_{T1} = \sigma^2_{T2}$,
4. $\rho(X_1) = \rho(X_2)$, and
5. $\rho(T_1, \Delta T) < 0$.

The question that begs to be asked is, "Why are these assumptions made in the derivations?" The first assumption appears reasonable as long as one assumes that scores on a measure are valid only if the test-retest correlation is large in magnitude.

The second assumption is often made to simplify the derivations. As Lord (1963) stated, it appears reasonable to assume this dynamic equilibrium (i.e., that the initial and final variances, over a group of individuals, are equal) in many test-retest situations where the time period between testing is small. This, of course, is not necessarily so and depends on the construct being measured. As well, the time 1 and time 2 measures often are considered to be parallel, which is empirically defined by the realization of assumption two (Lord & Novick, 1968). Assumption two works in concert with the often unstated and unrecognized assumption of the independence of errors to impose assumptions three, four, and five on the model.

It must be emphasized that if independence does not hold, the formulas presented in this review may lead to incorrect reliability estimates (Williams & Zimmerman, 1983; Zimmerman, Brotohusodo, & Williams, 1981). In particular, Labouvie (1980) and Williams and Zimmerman (1977) have argued that the mere repetitious use of a measuring instrument will result in correlated errors; therefore, the reliability of the simple change score may, in fact, be larger than the reliability of the time 1 or time 2 measures alone. As shown previously, the independent errors along with the equality of observed score variances implies that the true score variances will be equal on the two occasions, and further that the population reliabilities are equal across the two occasions. This implies that the correlation between the true score at time 1 and the change in true score is negative (i.e., regression toward the mean).

When the above five conditions are relaxed, the reliability of the difference score is adequate. Therefore, the widely claimed poor performance of the simple difference score is due to the psychometric constraints on the model (Rogosa et al., 1982). It now clear from the literature surveyed that the inequality of parameters that have been found to yield valid change scores also yield reliable change scores. Provided that the observed scores are reliable, the reliability and validity of the change scores indicates that the simple difference scores often are highly correlated with the true differences (Richards, 1975).

It is important to note at this point that the practice of transforming the time 1 and time 2 scores to z-scores or T scores (or any other transformation which equalizes the variances) prior to calculating difference scores is ill-advised. This sort of transformation will obviously result in conditions wherein the change score performs poorly. If standardized scores are desired, then one should calculate the raw difference for each individual so that difference can be transformed correctly to z-scores, T-scores, or any other desired transformation.

What is presently conspicuously missing in the voluminous literature on the methodology for the measurement of change is a survey (or meta-analysis) of existing empirical studies with the objective of investigating the actual degree of variance inequality, correlation and reliabilities of the pre- and post-test scores. This survey should be conducted for various areas of research. One may find, for example, that the conditions wherein the simple difference score performs poorly are commonly found in applied research settings. Furthermore, the values of these parameters could then be used in a series of Monte Carlo studies where the correlation between observed and true change could be calculated. A series of studies of this nature would help in settling the true state of affairs for empirical researchers.

Furthermore, it should be emphasized, at this point, that the presentation in this review is not intended to suggest that simple change (gain or difference) scores should be ubiquitously used in research settings. The reliability of differences scores, like that of all test scores, depends on the experimental procedures and on the proper use of the instruments by the investigator. This presentation, like that of Zimmerman and Williams (1982a, 1982b), indicates that change scores *can* be reliable and valid and that it would be premature to discard such measures in research and evaluation. As Zimmerman and Williams (1982a) stated:

> It is very likely that there are many situations in which simple pretest-posttest differences are quite useful. These measures need not be avoided because they are inherently unreliable from a statistical point of view, although the caution urged by previous authors is justified. (p. 153)

Clearly, Zimmerman and Williams, in their advocacy for the simple change score, do not encourage its use indiscriminantly.

Nevertheless, more recent literature suggests that, to a large extent, the intuition of researchers in many fields who believe that the simple change score can be a substantively meaningful index is well-founded. The questions that then beg to be asked are, "When should the simple difference score be used?" and "What does one do in those settings where the simple difference score should not be used?" These questions will be dealt with in the remainder of this paper.

Alternatives to the Simple Change Score

In the last decade, a series of research projects spearheaded by Rogosa and his colleagues, as well as Zimmerman and Williams, have demonstrated that the simple difference score is an acceptable index of change, at least under certain conditions. However, it is in the alternatives to the simple difference score where the least concensus exists. These alternatives are categorized as residualized change scores and multi-wave analyses.

Residualized Change Scores

As was mentioned in the first subsection of this chapter, it was believed that the simple change score was unfair because it was base-dependent (i.e., negatively correlated with initial status). This base-dependence has led to the development of the residualized change score as an alternative to the simple change score. In addition, the residualized change score is used to correct for ceiling effects.

Rogosa et al. (1982) showed that the true residual change score for an individual, p, is

$$E[T_{2p} \mid T_{1p}] = \mu_{T2} + \beta_{T2T1}[T_{1\mathrm{p}} - \mu_{T1}].$$

In words, this expression shows the expected value of the true score at time 2, given that the true score at time 1 is taken into account, is equal to the average true score at time 2 plus a weighted deviation of the individual's true score at time 1 and the average true score at time 1. The weighting coefficient for the deviation is the population regression slope from predicting T_2 from T_1. In this sense, the residualized true score is the residual obtained from the population regression of T_2 on T_1 (Cronbach & Furby, 1970; Rogosa et al., 1982). Therefore, as Rogosa et al. have shown, the true residual score for a given individual p is given by:

$$\Delta T(\text{resid})_p = [T_{2p} - \mu_{T2}] - \beta_{T2T1}[T_{1p} - \mu_{T1}], \tag{11}$$

where the notation is the same as the above except for ΔT(resid), which denotes the true residual change. Rogosa et al. defined the residual change as the amount an individual would have changed if all persons had the same true initial status. Obviously, in most settings the residualized change, ΔT(resid), will be quite different than the true change, ΔT. It is this difference that has led some authors (e.g., Cronbach & Furby, 1970; Rogosa et al., 1982) to recommend that researchers avoid the residualized difference score.

Despite the lack of equivalence of ΔT and ΔT(resid), many empirical researchers have adopted the residual change score. It is this author's opinion that this has happened for three reasons. First, if one follows the recommendations of Cronbach and Furby (1970) or Rogosa et al. (1982) to abandon the residualized change score, then one is left with rephrasing the research question to one of nonchange, or, to use growth curves with multi-wave data. Clearly, due to the abundance of two-wave data being reported, the empirical researcher has found neither of these options viable in all settings. Note that Zumbo (1994) warned against partitioning multi-wave data into multiple two-occasion analyses; the results from the partitioning can be quite biased.

The second reason for staying with the residual change score, despite arguments otherwise, is that researchers are quite comfortable with the concept of regression which entails a view that one variable can be, in some sense, equivalent to another variable or variables. This sense of substitutability is engrained in the use of regression analysis where the predicted dependent variable (commonly denoted as $\hat{Y}$) is considered equivalent to a composite of the independent variable(s) (commonly denoted as X). Importantly, the residualized difference score was being developed at about the same time that researchers in social and behavioral sciences were moving toward a common use of regression analysis in research settings (Cohen & Cohen, 1983; Kerlinger & Pedhazur, 1973).

Third, researchers appear to be using the residualized change score as a marker (i.e., an indirect measure) of true change. It is clear from Eq. (11) that ΔT(resid) is a linear function of ΔT, therefore, one could conjecture that statements based on ΔT(resid) apply to ΔT. This is particularly true of of statistical hypothesis testing using ΔT(resid).

Given that Eq. (11) is expressed in terms of true scores, Dubois (1957) and Manning and Dubois (1962) proposed estimating the residual change from the regression analysis of the observed time 2 measure on the observed time 1 measure. It is important to note that this is an ordinary least squares regression predicting time 2 from time 1. Furthermore, since the independent variable (X_1) has measurement error, the residual is an inefficient and inconsistent estimator of the true residuals. This inefficiency and inconsistency has led several authors to propose alternative measures of the residual change score (Bond, 1979; Messick, 1981; Tucker, 1979; Tucker, Damarin, & Messick; 1966) called base-free measures of change. Most empirical researchers, however, use the simple Manning and Dubois estimator (see Cohen & Cohen, 1983; Zimmerman & Williams, 1982c). Therefore, the remainder of this section will focus on the Manning and Dubois estimator (to be referred to as the residual change score or residualized change score).

Two pragmatic notes on the residual change score are fitting at this juncture. First, in applied settings the residual change score is computed using ordinary least squares (OLS) regression. OLS regression makes the assumption of normally (i.e., Gaussian) distributed pretest and posttest measures for parameter estimation as well as hypothesis testing (Lind & Zumbo, 1993). Outliers are always of concern using OLS; however, the presence of outliers is of *paramount* concern for the use of residualized change scores.

In regression terms, the residual change score is the time 2 deviation from the statistically expected regression for predicting time 2 from time 1. Therefore, an outlier or outliers will *drastically* bias the estimation of residual change scores. To treat the problem of outliers, Lind and Zumbo (1993) do not recommend data cleaning strategies (i.e., removal of outliers). Rather, they recommend the use of robust estimation methods, particularly M-estimators. These robust estimation techniques should be used to obtain b-weights and residuals with the reduced influence of outlying points.

The second pragmatic note on residual change scores involves the often cited recommendations by Hummel-Rossi and Weinberg (1975) to use regression analysis rather than change scores. That is, the time 2 measure is used as the dependent variable, with the time 1 measure considered as an independent variable. They recommended that the pretest measure be used as a covariate (i.e., forced into the regression equation first) followed by the remaining independent variables (observed measures and/or categorical variables representing group membership). The augmented model can be tested to determine whether the incremental explained variance is statistically significant. Although often considered methodology different than the residualized change score, the Hummel-Rossi and Weinberg approach is, under certain conditions, statistically equivalent to conducting the same analysis without the time 1 measure as an independent variable and using the Manning and Dubois residualized change score as the dependent variable. This is not intended to undermine the Hummel-Rossi and Weinberg recommendations, but rather to clarify that this is simply an alternative manner in which to conduct an analysis using the residualized change score.

Reliability and Validity of the Residualized Change Score

Williams and Zimmerman and their colleagues (Williams & Rich, 1990; Williams, Zimmerman, & Mazzagatti, 1987; Williams, Zimmerman, Rich, & Steed, 1984; Williams & Zimmerman, 1982a, 1983; Zimmerman & Williams, 1982) have investigated the comparative validity and reliability of the simple and residualized change scores, and the base-free measures of change. Their work was developed with an eye toward being able to compare the reliabilities and validities from sample information.

Regarding the reliability of the residualized change score, Williams and Zimmerman (1983) derived three sets of empirical conditions from which researchers can compare the reliabilities of the simple and residualized change scores, based on statistics such as the standard deviations, correlation, and reliabilities of the initial and final measures (see Zumbo (1992) for a correction to one of the algebraic results in Williams and Zimmerman (1983)). First, if the measures at time 1 and 2 are parallel (i.e., $\sigma_{X1} = \sigma_{X2}$), then $\rho(\Delta X\text{resid}) > \rho(\Delta X)$. Second, if $\rho(X_1, X_2) > 0$ and $\rho(X_2) > \rho(X_1)$, and either:

$$\text{(a) } \frac{\sigma_{X_1}}{\sigma_{X_2}} < 1.0 \text{ and } \frac{\sigma_{X_1}}{\sigma_{X_2}} < \rho(X_1, X_2) \text{ then } \rho(\Delta X) > \rho(\Delta X\text{resid}), \text{ or if}$$

$$\text{(b) } \frac{\sigma_{X_1}}{\sigma_{X_2}} \geq 1.0 \text{ and } \rho(X_1, X_2) > \frac{-\frac{\sigma_{X_2}}{\sigma_{X_1}}[\rho(X_1) + \rho(X_2) - 2] + \Phi}{4[1 - \rho(X_1)]}, \text{ where}$$

$$\Phi = \frac{\sigma_{X_2}}{\sigma_{X_1}}\sqrt{[\rho(X_1) + \rho(X_2) - 2]^2 - 8[1 - \rho(X_1)][\rho(X_1) - \rho(X_2)]},$$

then $\rho(\Delta X) > \rho(\Delta X\text{resid})$. Finally, if $\rho(X_1, X_2) > 0$, $\rho(X_2) < \rho(X_1)$, and either:

$$\text{(a) } \frac{\sigma_{X_1}}{\sigma_{X_2}} < 1.0,\ \rho(X_1, X_2) < \frac{\sigma_{X_1}}{\sigma_{X_2}},\ \text{and } \rho(X_1, X_2) > \frac{-\frac{\sigma_{X_2}}{\sigma_{X_1}}[\rho(X_1) + \rho(X_2) - 2] - \Phi}{4[1 - \rho(X_1)]},$$

where Φ is defined above, then $\rho(\Delta X) < \rho(\Delta X\text{resid})$, or if (b) $\sigma_{X1}/\sigma_{X2} \geq 1.0$ then $\rho(\Delta X\text{resid}) > \rho(\Delta X)$. Empirical evidence in support of these conditions have been given by Zimmerman, Andrews, Robinson, and Williams (1985) and Williams, Zimmerman, and Mazzagatti (1987).

Regarding the comparative validity of the residual and simple change scores, Zimmerman and Williams (1982b), and Williams and Zimmerman (1982a) derived formulas that would allow one to decide in a very simple manner which type of difference score possesses greater validity (see Gupta, Srivastava, & Sharma (1988) for an extension of the conditions presented by Zimmerman and Williams). In this context, validity has been defined by the authors as correlational evidence (see Messick, 1989) investigating the relationship between a measure and a criterion, *c*. The correlation between a difference score and a criterion is denoted $\rho(\Delta X, C)$, and the correlation between a residualized difference score and a criterion is denoted $\rho(\Delta X\text{resid}, C)$. The guideline for determining the relative magnitude of the validity coefficients is:

$$\rho(\Delta X, C) > \rho(\Delta X\text{resid}, C) \text{ if and only if } \rho(X_1, X_2) > \sigma_{X1}/\sigma_{X2}. \qquad (12)$$

Again, this can be investigated by testing a the statistical hypothesis involving the correlation coefficient. This guideline has been empirically verified by Zimmerman, Williams, Rich, and Steed (1984) and Rich and Williams (1990) and states that the residualized difference score is a more valid measure of change if, and only if, the ratio of time 1 and time 2 standard deviations is greater than or equal to the correlation between the measures at times 1 and 2. Otherwise, if the ratio of the time 1 and time 2 standard deviations is less than the correlation between the measures at times 1 and 2, then the simple difference score is more valid.

In summary, when questions of reliability or validity are of concern, a researcher can come to a quick decision about the relative merits of the simple and residualized change scores from the sample data at hand with the aid of even a hand held calculator. The guidelines presented by Zimmerman and Williams suggest that the residual change score be used in the cases where the simple change score is unreliable and unfair (particularly when the standard deviations are equal). It is particularly noteworthy, and comforting, that the guidelines given by Zimmerman and Williams regarding the comparative reliability and validity of the simple and residualized change scores are in accord with findings where the simple difference score performs poorly. Put simply, in situations where the time

1 and time 2 variances are equal, both the theoretical investigations and the Zimmerman and Williams guidelines suggest not to use the simple difference score.

Multi-wave Designs

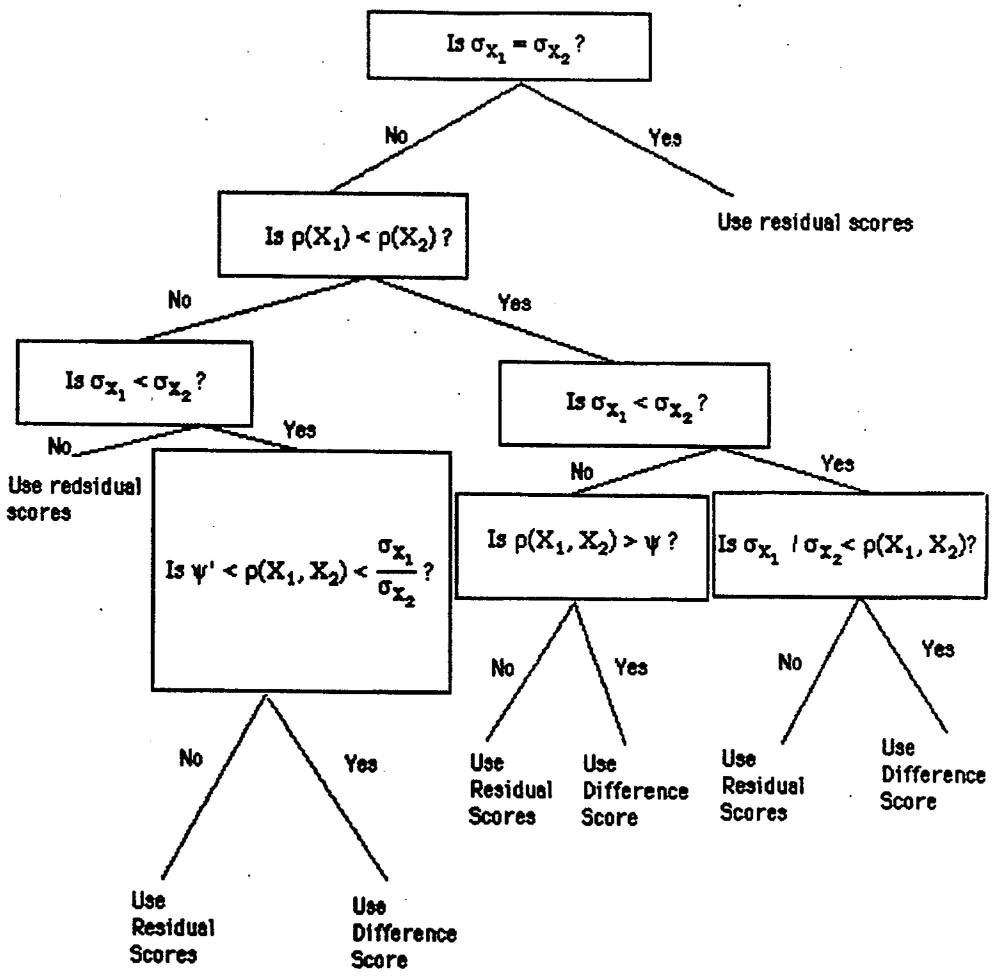
A radically different alternative to the residualized change score has been offered by Rogosa and his colleagues. In a series of papers, Rogosa et al. (1982) and Rogosa and Willett (1983, 1985) argued that most of the problems that have plagued the difference score are due to the fact that the scores are used within a two-wave design. Therefore, they examined the alternatives to the difference score with an eye toward advocating multi-wave individual growth curve models.

Traditionally, change has been viewed within a two-wave design. Until recently, it has not been viewed as a continuous growth over a long period of time. Change, and hence the measurement of change, is viewed as a discrepancy from time 1 to time 2. Alternatives to this two-wave design currently envision the measurement of change as a multi-wave process motivated by the statistical tradition of modeling individual growth over time (e.g., Rogosa, Brandt, & Zimowski, 1982). Rogosa and Willett (1985) stated that "the importance of obtaining observations on each individual at more than two points cannot be overstated" (p. 225).

They recommended modeling individual change with linear and nonlinear functions and using the parameter estimates as the indicators of the nature and degree of change. Rogosa and Willett further presented evidence that shows how the time at which the measures are taken can have major effects on the correlation of change with other variables. These are important points that should not be dismissed for they emphasize that researchers should not, where possible, rely exclusively on two-wave designs. The new methods utilizing hierarchical linear models by Bryk and Raudenbush (1987), and methods for the development of tests sensitive to monotonic change based on dynamic principles by Collins and Cliff (1990), should be strongly considered in the multi-wave setting.

However, due to financial and contextual factors (e.g., attrition) multi-wave designs are not always feasible. Furthermore, missing data can often compromise causal statements. Then, what is one to do? Is one to take a pessimistic approach, much like Cronbach and Furby, and recommend that questions of change should be rephrased into a nonchange question? There is far from a concensus in the literature on how to answer this question. The answer, in my opinion, is that one should still conduct the two-wave study; however, the data should be analyzed and the results should be interpreted with considerable caution.

As a historical note, the emphasis on individual growth curves is reminiscent of methods once used in learning theory (e.g., Anderson, 1963; Estes, 1956; Grant, 1956). The current debates over individual versus group growth curves, and multi-wave versus two-wave designs seems to be *redressing* (much like old wine in new bottles) many of the issues in the early learning literature.


Is $\sigma_{X_1} = \sigma_{X_2}$?
No
Yes
Use residual scores
Is $\rho(X_1) < \rho(X_2)$?
No
Yes
Is $\sigma_{X_1} < \sigma_{X_2}$?
No
Yes
Use redsidual scores
Is $\sigma_{X_1} < \sigma_{X_2}$?
No
Yes
Is $\psi' < \rho(X_1, X_2) < \frac{\sigma_{X_1}}{\sigma_{X_2}}$?
Is $\rho(X_1, X_2) > \psi$?
Is $\sigma_{X_1} / \sigma_{X_2} < \rho(X_1, X_2)$?
No
Yes
Use Residual Scores
Use Difference Score
No
Yes
Use Residual Scores
Use Difference Score
No
Yes
Use Residual Scores
Use Difference Score

Note: In this figure two new symbols are introduced. They ar defined as,

$$\psi = \frac{-\frac{\hat{\sigma}_{X_2}}{\hat{\sigma}_{X_1}}[\hat{\rho}(X_1) + \hat{\rho}(X_2) - 2] + \Phi}{4[1 - \hat{\rho}(X_1)]}, \text{ and } \psi' = \frac{-\frac{\hat{\sigma}_{X_2}}{\hat{\sigma}_{X_1}}[\hat{\rho}(X_1) + \hat{\rho}(X_2) - 2] - \Phi}{4[1 - \hat{\rho}(X_1)]},$$

$$\text{where } \Phi = \frac{\hat{\sigma}_{X_2}}{\hat{\sigma}_{X_1}}\sqrt{[\hat{\rho}(X_1) + \hat{\rho}(X_2) - 2]^2 - 8[1 - \hat{\rho}(X_1)][\hat{\rho}(X_1) - \hat{\rho}(X_2)]}.$$

Any symbols with a hat-notation (^) denote sample estimates of the population parameter.

Figure 2. A flowchart to help decide whether to use difference (simple difference) or residualized scores.

Zumbo, B. D
(1999)
The Simple
Difference Score
as an inherently
poor measure of
Changes. pp. 293 – 304

RECOMMENDATIONS FOR THE ANALYSIS OF TWO-WAVE DATA

This section recommends statistical procedures for dealing with two-wave data. The research question should always dictate the statistical methods; therefore, alternatives will be discussed according to the research question. Also recall our earlier comments on the precedence of substantive theory. In the two-wave setting, the most common research questions involve change scores as descriptive measures, group comparisons, or correlates of change.

Figure 2 will play an important role in these recommendations. This figure is a flowchart to help decide whether to use simple difference or residualized scores. The decision rules in this flowchart are based on the derivations by Williams and Zimmerman (1983) and Zumbo (1992). Therefore, the decisions made from Figure 2 are based on maximizing the reliability of the index of change. I will demonstrate how to use the flowchart with an example. In this example the standard deviations at times 1 and 2 are 3.4 and 6.2, respectively. The correlation between the measures at times 1 and 2 is .70, and the reliabilities are .60 and .70, respectively. The first observation is that the standard deviations are not equal. This leads to the query of whether the reliability at time 1 is less than that at time 2. Establishing that the standard deviation is smaller at time 1 than at time 2, and that a ratio of these standard deviations is less than the correlation between the time 1 and 2 correlation leads to recommending the difference score. The example demonstrates that with the aid of the sample statistics a researcher can decide whether to use the residualized or simple difference score.

Two pragmatic issues are important to note at this point. First, in experimental settings or other situations where the reliability of the observations is not readily available, Williams and Zimmerman (1983) presented a simple decision rule that can be utilized instead of the steps in Figure 2. That is, on the basis of maximizing reliability, one should utilize the simple difference score instead of the residualized difference if and only if $\rho(X_1, X_2) > \sigma_{X1}/\sigma_{X2}$. This decision rule was derived on the basis of restricting the set of conditions in Figure 2 to those that are *expected* to apply to the majority of cases found in practice. When the information is available, however, Figure 2 should be utilized.

The second pragmatic point is that, given that the conditions in Figure 2 are specified in population values, statistical hypothesis tests can be utilized at each node of the decision tree. These hypothesis testing strategies are of particular relevance in small sample settings. The sampling theory for variance ratios and correlation coefficients can be found in most introductory statistics textbooks. However, the sampling theory for reliability for two dependent samples, estimated by coefficient alpha, can be found in Feldt (1980) or for smaller samples in Kristof (1964).

Change Scores as Descriptive Measures

In some research settings change measures are used for descriptive purposes. That is, a pre-post study is conducted with the objective of using a descriptive index to identify individuals who change a very large or small amount. This may be done in educational or learning studies where the objective is to identify exceptional learners. In this case one should use an index which maximizes reliability (i.e., maximizes the precision of measurement) and is just. This can be achieved by using the flowchart in Figure 2. Additionally, in cases where ceiling or floor effects are present, the residual change score is appropriate.

A note of caution should be heeded in interpreting residualized change scores. Theoretically, these scores reflect the deviation from the expected pretest score, a value which may not always be substantively appealing (e.g., a negative predicted posttest score).

Group Comparisons

One-group

The most common questions asked with the one-group design are "Did the subjects significantly change?" or "Did the treatment of intervention cause a change from the pretest to posttest?" The recommended analysis to answer these questions is the one-sample *t*-test using a change index as the dependent variable. However, two issues must be considered when conducting this analysis: (a) desired causal statements, and (b) which change index to use.

With regard to making causal statements from the one-group pretest-posttest design, it should be noted that this design is not appropriate for causal statements without due consideration to possible threats to internal validity, such as history or maturation (see Cook & Campbell, 1979).

Regarding choosing a change index, an important issue is the power of the statistical significance test. The power of a statistical significance test is determined by sample size, population variance, significance level, the magnitude of the alternative hypothesis, and directionality of the test. If all of these variables have fixed values, then the power of a statistical significance test is completely determined and is independent of the reliability of measurement (Zimmerman, Williams, & Zumbo, 1993a, 1993b). Zimmerman et al. suggest that this means that the statistical power is related to the reliability of measurement, but is not a function of the reliability of measurement unless either true variance or error variance is constant. Quite simply, then, reliability is determined by the relative magnitude of true and error variance, while power of a statistical significance test is determined by the absolute magnitude of the total variance. See Williams, Zimmerman, and Zumbo (1995) for a review of the reliability of measurement and statistical power controversy.

Although many authors of statistics textbooks have correctly expressed the basic idea that statistical power increases when measurements become more reliable they are often remiss to explain that high reliability does not guarantee high statistical power, and furthermore that low reliability does not always preclude high statistical power. Put simply, the difference between two highly variable groups is difficult to detect even with perfect reliability, while a small difference between homogeneous groups can sometimes be detected by a measure yielding scores with considerable unreliability.

What this means for the researcher using one-group designs is that it is reasonable to select the index of change that maximizes reliability (using Figure 2), however, this should be coupled with consideration to the other determinants of statistical power as well as measures of effect size. Additionally, in cases where ceiling or floor effects are present, the residual change score is appropriate.

Two or More Groups

Given that the one-group design is a special case of the two-or-more-groups design, the research questions commonly asked with these designs is similar to that of the one-group design. In fact, the recommendations for the one-group design should be carried forth to the two-or-more-groups design.

However, it should be noted that in the two-or-more-groups design the question of interest usually is "Does one group change more than another and/or what is the cause of the greater change?" Therefore, a central issue is whether one has random or nonrandom assignment. Given proper experimental controls and random assignment traditional linear model procedures (e.g., mixed-model ANOVA, ANCOVA) are suitable methods of analysis. However, the whole enterprise of comparing groups without proper random assignment has been the source of a voluminous literature (e.g., Weisberg, 1979). The contentious issue has been the causal statements made from comparisons without random assignment when the pretests differ for the groups involved. This has lead to various preliminary methods to investigate pretest equality (Overall & Ashby, 1991).

Three methods for adjusting pretest differences are commonly considered:

1. calculation of simple difference scores,
2. ANCOVA with the pretest measure entered as a covariate,
3. a posttest/pretest ratio.

It is noteworthy that the ANCOVA method, like the Hummel-Rossi and Weinberg (1975) method, is seldom recognized as an analysis of residualized change scores (see, for example, Malgady & Colon-Malgady, 1991; Thompson, 1992). Furthermore, each of these methods have been criticized for their inadequacies (Campbell & Boruch, 1975; Campbell & Stanley, 1966; Lord, 1960; Rubin, 1973). In fact, it has been suggested by methodologists that there is no adequate and defensible solution to this problem (Cronbach & Furby, 1970; Lord, 1969). This recommendation has, in fact, been one of the major sources of discontent in using pre-post designs.

However, recent developments in mathematical statistics (for a summary see, Holland & Rubin, 1983) have provided a model to help choose among the simple difference, residualized difference, and percentage change scores. When causal inferences are of interest, Wainer (1991) has presented a readable summary of the Holland and Rubin methodology. Wainer has shown that if one is careful and explicit about the goals of the study as well as the assumptions that one is willing to make, then the Holland and Rubin model allows one to decide which of the three adjustments is most appropriate for causal inferences.

For the purpose of this paper I will consider the three suggestions of Holland and Rubin's model for deciding between simple and residualized difference scores (in this case ANCOVA). A detailed discussion of Holland and Rubin's model is beyond the scope of this paper. The reader interested in the mathematical model that drives these recommendations should see Holland (1986) and Rubin (1978).

First, when considering descriptive questions such as, "Who gained more from treatment *X*?" the index of gain (change) can be selected on the basis of Figure 2. This should not hark back to the days when claims as to the supposed effectiveness of interventions/treatments were made from questions where there was not a measure of treatment effect. It is important to note that this is not a causal question, but a purely descriptive one which can be answered without making the necessary assumptions for causal inference. Questions such as, "Who gained more from treatment *X*?" are still sometimes mistakenly considered as causal questions; however, Holland and Rubin's model clearly shows otherwise. The distinguishing feature between descriptive and causal questions is that for descriptive questions their is no implied comparison under different levels of a control condition (Holland & Rubin, 1983).

Second, when considering causal questions, such as "Is the treatment more effective for group 1 or group 2?" proper randomization is a sufficient condition to make causal statements.

Third, when considering causal questions, such as "Is the treatment more effective for group 1 or group 2?" and proper randomization is *not* present, then Holland and Rubin's model suggests that the biggest stumbling block to causal inference is that there is no control condition, denoted *c*. In their model, the influence of the treatment is always relative to some condition, *c*. The data analyst's problem, then, is to make assumptions that allow specification of a *c*. Holland and Rubin's model suggests that if the researcher is willing to assume that *c* is the pretest score, then the simple difference score will allow causal inferences. However, if the researcher is willing to assume that *c* is a linear function of the value at pretest and that the same linear function applies irrespective of group membership, then the residualized change score will allow causal inferences.

The fundamental problem for the causal inference is estimating the missing data, *c*, by the pretest or the estimated posttest (this estimation being based on the pretest). As Holland and Rubin (1983) suggested, the decision of whether to use the simple or residualized change score is based on two untestable assumptions (i.e., c = pretest or c = estimated posttest). The reseacher's role is to decide which assumption is more viable based on prior information, the substantive literature, and intuition.

Correlates of Change

The most common questions asked with the correlates of change are "What type of individuals change most/least from the treatment or intervention?" or "Is variable *X* correlated with change and can we later predict the amount of change from variable *X*?" In these questions one is fundamentally interested in investigating the relationship between the amount of change and a number of other variables.

The results previously reviewed, on the comparative validity of the simple and residual change scores, are useful in this correlates of change setting. In fact, Eq. (12) can be used directly to help decide whether to use simple or residualized change scores. Alternatively, as Rich and Williams (1990) stated in the absence of attenuation paradoxes (Williams & Zimmerman, 1982b) large validity coefficients imply large reliability coefficients. Then, Figure 2 can be used in choosing between the simple and residualized change scores. The recommended methodology is to use a regression analysis with the simple change score as the dependent variable or conducting a residualized change score analysis by forcing the pretest measure into the equation first (treating it as a covariate) and investigating the unique contribution of the variables of interest (see, Gardner & Neufeld, 1987 for interpretation of simple differences in this context).

In summary, for the previously described studies a researcher can choose between the simple and residualized change score based on their comparative reliability and/or validity. However, the described Group Comparison studies are not as straight-forward because reliability and validity need to be considered in light of the desire to make causal statements. If the result of the decisions based on reliability and validity are antithetical to the recommendations considering causality, then the researcher should use multi-wave designs.

CONCLUSIONS

The paper by Cronbach and Furby (1970) has had an enormous impact on the practice of behavioral and social research. Cronbach and Furby developed and reviewed many measures of change; however, researchers have used this paper to (a) suggest that questions of change either be rephrased as nonchange questions or not be asked at all, or (b) support the statement that simple and/or residualized change scores are inherently poor measures of change. Considering the often cited reference to Cronbach and Furby, the present chapter is a long overdue response to that very influential work. Note that the paper by Rogosa, Brandt, and Zimowski (1982) more correctly is considered the response to Cronbach and Furby's dismissal of change scores; however, the chapter goes beyond the Rogosa et al. results with an eye towards accessibility by the general behavioral or social scientist—the potential consumers of change scores. The overly prescriptive effect of Cronbach and Furby's recommendations needs to be redressed.

The present review of the voluminous literature on the measurement of change points to the fact that there certainly is some reality to the statement that simple change scores are poor measures of change. However, the statement that it is an inherently poor measure and hence has no redeeming qualities is pure mythology. The evidence reviewed herein supports the claim that under certain realistic conditions the simple difference score can be quite appropriately used. In fact, Eq. (12) and Figure 2 demonstrate clearly some conditions wherein the simple difference score is recommended.

Motivated by the work of Cronbach and Furby (1970) and Rogosa et al. (1982), recent papers on the analysis of change have discounted the two-wave design entirely (e.g., Bryk & Raudenbush, 1987; Collins & Cliff, 1990). Most of these recent studies are interested in longitudinal growth or decline of abilities or attributes—for a discussion, see, Nesselroade, Stigler, and Baltes, (1980). In this context, multi-wave designs are appropriate. However, two-wave designs are still necessary in certain settings where multi-wave designs are not possible or are not of interest. For example, researchers in the social and behavioral sciences are frequently interested in the short-term effects of a treatment or intervention. Is one then to discourage these studies and editors refuse to publish them? The primary point of this chapter is that these studies should be conducted and published, and that the simple difference score can be used in some settings while the residualized difference score is appropriate in others.

I would like to respond to Cronbach and Furby's (1970) question, "How should we measure change—or should we?" How we should measure change depends, of course, on the research question. If one is interested in change over a long period of time, then there are many exciting alternative methodologies. If, however, the research question involves a two-wave design, recent evidence indicates that either the simple difference score or the residualized difference score is the preferred measure. And should we measure change? Of course, we should. Change is a fundamental and pervasive psychological concept that we now have the methodological machinery to address.

ACKNOWLEDGMENT

I thank Jo Tombaugh, Don Zimmerman, Don Andrews, Bob Hoffmann, and Anita Hubley for reading and commenting on earlier versions of this paper. Special thanks to Jo Tombaugh who clarified and simplified my notational system and a note of debt to Arnold Binder (1984) whose subtitle clearly expressed the message in this paper. I would like to thank my former professors D.W. Zimmerman, W.W. Rozeboom, K.V. Wilson, and E. Howarth, for it was through watching and learning from them that I gained my inspiration.

REFERENCES

Anderson, N.H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Review, 70,* 162-179.

Anderson, T.W. (1971). *The statistical analysis of time-series.* New York: Wiley.

Anderson, O.D. (1975). *Time-series analysis and forecasting: The Box-Jenkins approach.* London: Butterworths.

Andrews, D.A. (1983). Assessment of outcome in correctional samples. In M.J. Lambert, E.R. Christensen, & S.S. DeJulio (Eds.), *The assessment of psychotherapy outcome.* New York: John Wiley.

Andrews, D.A., Bonta, J., & Hoge, R.D. (1990). Classification for effective rehabilitation: Rediscovering Psychology. *Criminal Justice and Behavior, 17,* 19-52.

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

Bickel, P.J., & Doksum, K.A. (1977). *Mathematical statistics.* Oakland, CA: Holden-Day, Inc.

Binder, A. (1984). Restrictions on statistics imposed by method of measurement: Some reality, much mythology. *Journal of Criminal Justice, 12,* 467-481.

Bloom, B.S. (1964). *Stability and change in human characteristics.* New York: John Wiley.

Bond, L. (1979). On the base-free measure of change proposed by Tuck, Damarin, and Messick. *Psychometrika, 44,* 351-355.

Boyle, G.J. (1987). Commentary: The role of intrapersonal psychological variables in academic school learning. *Journal of School Psychology, 25,* 389-392.

Brillinger, D.R. (1975). *Time-series: Data analysis and theory.* New York: Holt, Rinehart and Winston.

Bryk, A.S., & Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147-158.

Campbell, D.T., & Boruch, R.F. (1975). Making a case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C.A. Bennett & A.A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs.* New York: Academic Press.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cardinet, J. (1994). Comparing psychometric, edumetric and profile analysis procedures for the measurement of student learning. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems & issues.* Ottawa: University of Ottawa.

Cattell, R.B. (1966). Patterns of change. In R.B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally.

Cattell, R.B. (1982). The clinical use of difference scores: Some psychometric problems. *Multivariate Experimental Clinical Research, 6,* 87-98.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/ orrelation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation: Design and analysis issues in field settings*. Boston: Houghton Mifflin Company.

Corballis, M.C., & Traub, R.E. (1970). Longitudinal factor analysis. *Psychometrika, 35*, 79-93.

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

Dubois, P.H. (1957). *Multivariate correlational analysis*. New York: Harper.

Eid, M. (1996). Longitudinal factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online* (Vol. 1, No. 4.). Internet: http://www.pabst-publishers.de/mpr/

Embretson, S.E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems & issues*. Ottawa: University of Ottawa.

Estes, W.K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134-140.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99-105.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8*, 172-179.

Gardner, R.C., & Neufeld, R.W.J. (1987). Use of the simple change score in correlational analyses. *Educational and Psychological Measurment, 47*, 849-862.

Grant, D.A. (1956). Analysis-of-variance tests in the analysis and comparison of curves. *Psychological Bulletin, 53*, 141-154.

Gupta, J.K., Srivastava, A.B., & Sharma, K.K. (1988). On the optimum predictive potential of change measures. *Journal of Experimental Education, 55*, 116-118.

Haertel, E.H., & Wiley, D.E. (1990, April). *Poset and lattice representations of ability structures: Implications for test theory*. Paper presented at the symposium entitled "Test theory for a new generation of tests" at the annual meeting of the American Educational Research Association, Boston, MA.

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

Hannan, E.J. (1970). *Multiple time series*. New York: Wiley.

Healy, M.J.R., & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology, 5*, 277-280.

Higgins, N.C., Zumbo, B.D., & Hay, J.L. (1999). Construct validity of attributional style: Modeling context-dependent item sets in the Attributional Style Questionnaire. *Educational and Psychological Measurement, 59*, 804-820.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.

Holland, P.W., & Rubin, D.B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Hillsdale, NJ: Erlbaum.

Holtzman, J.M. (1970). *Nonlinear System Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Hubley, A.M., & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*, 207-215.

Huck, S.W., & McLean, R.A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*, 511-518.

Hummel-Rossi, B., & Weinberg, S.L. (1975). Practical guidelines in applying current theories to the measurement of change. *Journal Supplement Abstract Service-Catelog of Selected Documents in Psychology, 5*, 226 (MS. No. 916).

Kerlinger, F.N., & Pedhazur, E.J. (1973). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Kristof, W. (1964). Testing differences between reliability coefficients. *British Journal of Mathematical and Statistical Psychology, 17,* 105-111.

Labouvie, E.W. (1980). Measurement of individual differences in intraindividual changes. *Psychological Bulletin, 88,* 54-59.

Lind, J.C., & Zumbo, B.D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology, 34,* 407-414.

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.

Loehlin, J.C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16,* 421-437.

Lord, F.M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307-321.

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

Lord, F.M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72,* 336-337.

MacCallum, R.C., Kim, C., Malarkey, W., & Kiecolt-Galser, J. (1997). Studying multivariate change using multilevels models and latent curve models. *Multivariate Behavioral Research, 32,* 215-253.

Malgady, R.G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement, 51,* 803-807.

Manning, W.H., & Dubois, P.H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills, 15,* 287-321.

Maxwell, S.E., & Howard, G.S. (1981). Change scores—Necessarily anathema? *Educational and Psychological Measurement, 41,* 747-756.

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale, NJ: Erlbaum.

Messick, S. (1981). Denoting the base-free measure of change. *Psychometrika, 46,* 315-217.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Millsap, R.E., & Meredith, W. (1988). Component analysis in cross-sectional and longitudinal data. *Psychometrika, 53,* 123-134.

Molenaar, P.C.M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50,* 181-202.

Naslin, P. (1965). *The dynamics of linear and non-linear systems.* London: Blackie.

Nesselroade, J.R., Stigler, S.M., & Baltes, P.B. (1980). Regression toward the mean and the study of changes. *Psychological Bulletin, 88,* 622-637.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

O'Connor, E.F., Jr. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42,* 73-97.

Overall, J.E., & Ashby, B. (1991). Baseline corrections in experimental and quasi-experimental clinical trials. *Neuropsychopharmacology, 4,* 273-281.

Raykov, T. (1991). Measurement of change in longitudinal data: A classical test theory approach within the structural equation modeling methodology. *Studia Psychologia, 33,* 44-50.

Raykov, T. (1992a). Base-free measurement of change: A structural equation modeling approach. *Zeitschrift-fuer-Psychologie, 200,* 79-86.

Raykov, T. (1992b). On structural models for analyzing change. *Scandinavian Journal of Psychology, 33,* 247-265.

Raykov, T. (1992c). Structural models for studying correlates and predictors of change. *Australian Journal of Psychology, 44,* 101-112.

Raykov, T. (1993a). A structural equation model for measuring residualized change and discerning patterns of growth or decline. *Applied Psychological Measurement, 17,* 53-71.

Raykov, T. (1993b). Classical test theory based and factor analytic structural models for analyzing change: A note on differences and similarities. *Scandinavian Journal of Psychology, 34,* 94-96.

Raykov, T. (1994). Studying correlates and predictors of longitudinal change using structural equation modeling. *Applied Psychological Measurement, 18,* 63-77.

Rich, J.C., & Williams, R.H. (1990). Predicting the relative magnitudes of validity coefficients for four types of gain scores. *Perceptual and Motor Skills, 71,* 1011-1014.

Richards, J.M., Jr. (1975). A simulation study of the use of change measures to compare educational programs. *American Educational Research Journal, 12,* 299-311.

Rogosa, D.R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245-258.

Rogosa, D.R. (1985). Analysis of reciprocal effects. In T. Husen & N. Postlethwaite (Eds.), *International encyclopedia of education.* London: Pergamon Press.

Rogosa, D.R. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12,* 185-195.

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

Rogosa, D.R., Floden, R., & Willett, J.B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology, 76,* 1000-1027.

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

Rogosa, D.R., & Willett, J.B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203-228.

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

Rozeboom, W.W. (1978a). *General linear dynamic analysis (GLDA) models for studying change.* Unpublished manuscript, Center for Advanced Studies in Theoretical Psychology & Department of Psychology, University of Alberta.

Rozeboom, W.W. (1978b). *Dynamic analysis of multivariate process data.* Unpublished monograph, Center for Advanced Studies in Theoretical Psychology & Department of Psychology, University of Alberta.

Rubin, D.B. (1973). The use of matched sampling and regression adjustments to remove bias in observational studies. *Biometrics, 29,* 185-203.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 7,* 34-58.

Rushton, J.P., Brainerd, C.J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18-38.

Singer, J.D., & Willett, J.B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110,* 268-298.

Steyer, R. (1988). Conditional expectations: An introduction to the concept and its applications in the empirical sciences. *Methodika, 2,* 53-78.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3,* 25-60.

Swaminathan, H. (1984). Factor analysis of longitudinal data. In H.G. Law, C.W. Snyder, Jr., J.A. Hattie, & R.P. McDonald (Eds.), *Research methods for multimode data analysis*. New York: Praeger.

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

Thompson, B. (1992). Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology, 13,* iii-xviii.

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

Tisak, J., & Meredith, W. (1989). Exploratory longitudinal factor analysis in multiple populations. *Psychometrika, 54* 261-281.

Tucker, L.R. (1963). Implications of factor analysis of three-way matrices for measurement of change. In C.W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

Tucker, L.R. (1979). Comment on a note on a base-free measure of change. *Psychometrika, 44,* 357.

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

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin, 109,* 147-151.

Weisberg, H.I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin, 86,* 1149-1164.

Werts, C.E., & Linn, R.L. (1970). A general linear model for studying growth. *Psychological Bulletin, 73,* 17-22.

Wilder, J. (1957). The law of initial values in neurology and psychiatry. *Journal of Nervous and Mental Disease, 125,* 73-86.

Willett, J.B., & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363-381.

Willett, J.B., & Singer, J.D. (1988, April). *Doing data analysis with proportional hazards models: Model building, interpretation and diagnosis*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Willett, J.B., & Singer, J.D. (1991a). How long did it take? Using survival analysis in educational and psychological research. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association.

Willett, J.B., & Singer, J.D. (1991b). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research, 61,* 407-450.

Williams, R.H., & Zimmerman, D.W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement, 37,* 679-689.

Williams, R.H., & Zimmerman, D.W. (1982a). Comparative validity of simple and residualized difference scores. *Psychological Reports, 50,* 91-94.

Williams, R.H., & Zimmerman, D.W. (1982b). Reconsideration of the "attenuation paradox"—and some other paradoxes in test validity. *Jouranl of Experimental Education, 50,* 164-171.

Williams, R.H., & Zimmerman, D.W. (1983). The comparative reliability of simple and residualized difference scores. *Journal of Experimental Education, 51,* 94-97.

Williams, R.H., & Zimmerman, D.W. (1984). A critique of Knapp's "The (un)reliability of change scores in counseling research." *Measurement and Evaluation in Guidance, 16,* 179-182.

Williams, R.H., Zimmerman, D.W., & Mazzagatti, R.D. (1987). Large sample estimates of the reliability of the simple, residualized, and base-free measures gain scores. *Journal of Experimental Education, 55,* 116-118.

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

Williams, R.H., Zimmerman, D.W., & Zumbo, B.D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *Journal of Experimental Education, 63*, 363-370.

Williams, R.H., & Zimmerman, D.W. (1996a). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 59-69.

Williams, R.H., & Zimmerman, D.W. (1996b). Commentary on the commentaries of Collins and Humphreys. *Applied Psychological Measurement, 20*, 295-297.

Winderknecht, T.G. (1971). *General dynamic processes*. New York: Academic Press.

Zimmerman, D.W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40*, 395-412.

Zimmerman, D.W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement, 36*, 85-96.

Zimmerman, D.W. (1997). A geometric interpretation of the validity and reliability of difference scores. *British Journal of Mathematical and Statistical Psychology, 50*, 73-80.

Zimmerman, D.W., Andrews, D.A., Robinson, D., & Williams, R.H. (1985). A note on the non-parallelism of pretest and posttest measures in assessing change. *Journal of Experimental Education, 53*, 234-236.

Zimmerman, D.W., Brotohusodo, T.L., & Williams, R.H. (1981). The reliability of sums and differences of test scores: Some new results and anomalies. *Journal of Experimental Education, 49*, 177-186.

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). The relative error magnitude in three measures of change. *Psychometrika, 47*, 141-147.

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

Zimmerman, D.W., & Williams, R.H. (1982d). A note on the correlation of gains and initial status. *Journal of General Psychology, 107*, 203-207.

Zimmerman, D.W., & Williams, R.H. (1998). Reliability of gain scores under realistic assumptions about properties of pretest and posttest scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343-351.

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.

Zumbo, B.D. (1994). The lurking assumptions in using generalizability theory to monitor an individual's progress. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems & issues* (pp. 261-278). Ottawa: University of Ottawa.

Zumbo, B.D. (1992). The comparative reliability of simple and residualized difference scores: A corrigendum. *Journal of Experimental Education, 61*, 81-83.

Zumbo, B.D. (Ed.). (1998). *Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences* (Special volume, 45, of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*). Amsterdam: Kluwer Academic Press.