Is Reliability Obsolete? A Commentary on "Are Simple Gain Scores Obsolete?"

Linda M. Collins

The Pennsylvania State University

Williams & Zimmerman (1996) provided much-needed clarification on the reliability of gain scores. This commentary translates these ideas into recognizable patterns of change that tend to produce reliable or unreliable gain scores. It also questions the relevance of the traditional idea of reliability to the measurement of change. *Index terms: change scores, classical test theory, difference scores, gain scores, intraindividual differences, measurement of growth, reliability, test theory, validity.*

There are few topics in social science methodology that have elicited as much confusion, misunderstanding, and anxiety as the topic of the reliability of gain scores. The controversy began with the influential Cronbach & Furby (1970) article, which not only called gain scores into question but even asked whether we should measure change at all. Like a fire spreading out of control in a drought-ridden forest, the ideas presented in that article seemed to consume everything in their path. The inherent unreliability of gain scores eventually became taken for granted, and many journal reviewers and editors began insisting that published articles use alternative approaches, such as residualized gain scores [see Rogosa, Brandt, & Zimowski (1982) for a discussion of whether this represents a distinct alternative]. This reaction to the Cronbach and Furby article has persisted for over 25 years.

However, the situation is complex and deserves more than a knee-jerk reaction. Gain scores are not inherently unreliable; instead, their reliability depends on a number of factors, which are explained in a wonderfully clear and well thought out manner by Williams & Zimmerman (1996). These authors noted that most of the literature averring that gain scores are unreliable has made the implicit assumption that the pretest and posttest standard deviations are equal. They showed that the reliability of gain scores ($\rho_{DD'}$) is a function of the reliability of the pretest X ($\rho_{XX'}$), the reliability of the posttest Y ($\rho_{YY'}$), the correlation between X and Y (ρ_{XY}), and a parameter λ that represents the ratio of the pretest standard deviation σ_X to the posttest standard deviation σ_Y:

$$\lambda = \frac{\sigma_X}{\sigma_Y}. \tag{1}$$

The key to understanding the reliability of gain scores lies in λ and ρ_{XY}. The effect of λ is as follows: Given fixed $\rho_{XX'}$, $\rho_{YY'}$, and ρ_{XY}, $\lambda = 1$ results in the smallest reliability of gain scores. Williams & Zimmerman (1996) showed that most of the arguments that gain scores are inherently unreliable have made the assumption that $\lambda = 1$. But as λ deviates from 1 in either direction, $\rho_{DD'}$ increases. The effect of ρ_{XY} is as follows: Given fixed $\rho_{XX'}$, $\rho_{YY'}$, and λ, as ρ_{XY} decreases, $\rho_{DD'}$ increases. In both cases, the increase in reliability is up to a maximum $\rho_{DD'}$ determined by $\rho_{XX'}$ and $\rho_{YY'}$. Thus, given fixed $\rho_{XX'}$ and $\rho_{YY'}$, gain scores are least reliable

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 289–292

0146-6216/96/030289-04$1.45

when ρ_{XY} is large and positive and $\lambda = 1$, and most reliable when ρ_{XY} is small and λ deviates considerably from 1.

Characteristics of Growth and Their Implications for Reliability

The circumstances under which gain scores are reliable and those under which they are not can be discussed in terms of distinct patterns of growth that can be recognized, and possibly anticipated, in empirical data. In Figure 1, pretest (Variable X) and posttest (Variable Y) data collected on six persons are shown, with the pretest and posttest data for a person connected by a line. The slopes of these lines indicate the rate of growth between pretest and posttest for each person.

Figure 1a is an example of data in which ρ_{XY} is large and positive and $\lambda = 1$; in other words, the rank order of persons and the standard deviations of X and Y are identical at each time. This includes situations in which no growth occurs between pretest and posttest. As Figure 1a shows, if there is growth it is of about the same direction and magnitude for the entire group. In Figure 1b, ρ_{XY} is large and positive and λ is small. Under these circumstances growth is some variation of a fan spread, in which persons change in different amounts, and even in different directions, but the rank order of persons is not disturbed. In Figure 1c, ρ_{XY} is small and $\lambda = 1$. In general, whenever ρ_{XY} is small the gain scores will have a relatively large variance, and λ has little effect.

These illustrations show conceptually why the gain scores in Figure 1a have the lowest reliability. The reliability of gain scores is

$$\rho_{DD'} = \frac{\sigma^2_{T_D}}{\sigma^2_{T_D} + \sigma^2_{E_D}}, \tag{2}$$

where $\sigma^2_{T_D}$ represents true gain score variance and $\sigma^2_{E_D}$ represents error gain score variance. Equation 2 shows that given a fixed error variance, reliability cannot be large when there is little true score variability. Figure 1a shows that when ρ_{XY} is large and $\lambda = 1$, persons change at the same rate, and thus there can be little true score variability in change over time. Figures 1b and 1c illustrate data in which there can be more true score variability in change, and thus there is the possibility of a more reliable measure of change.

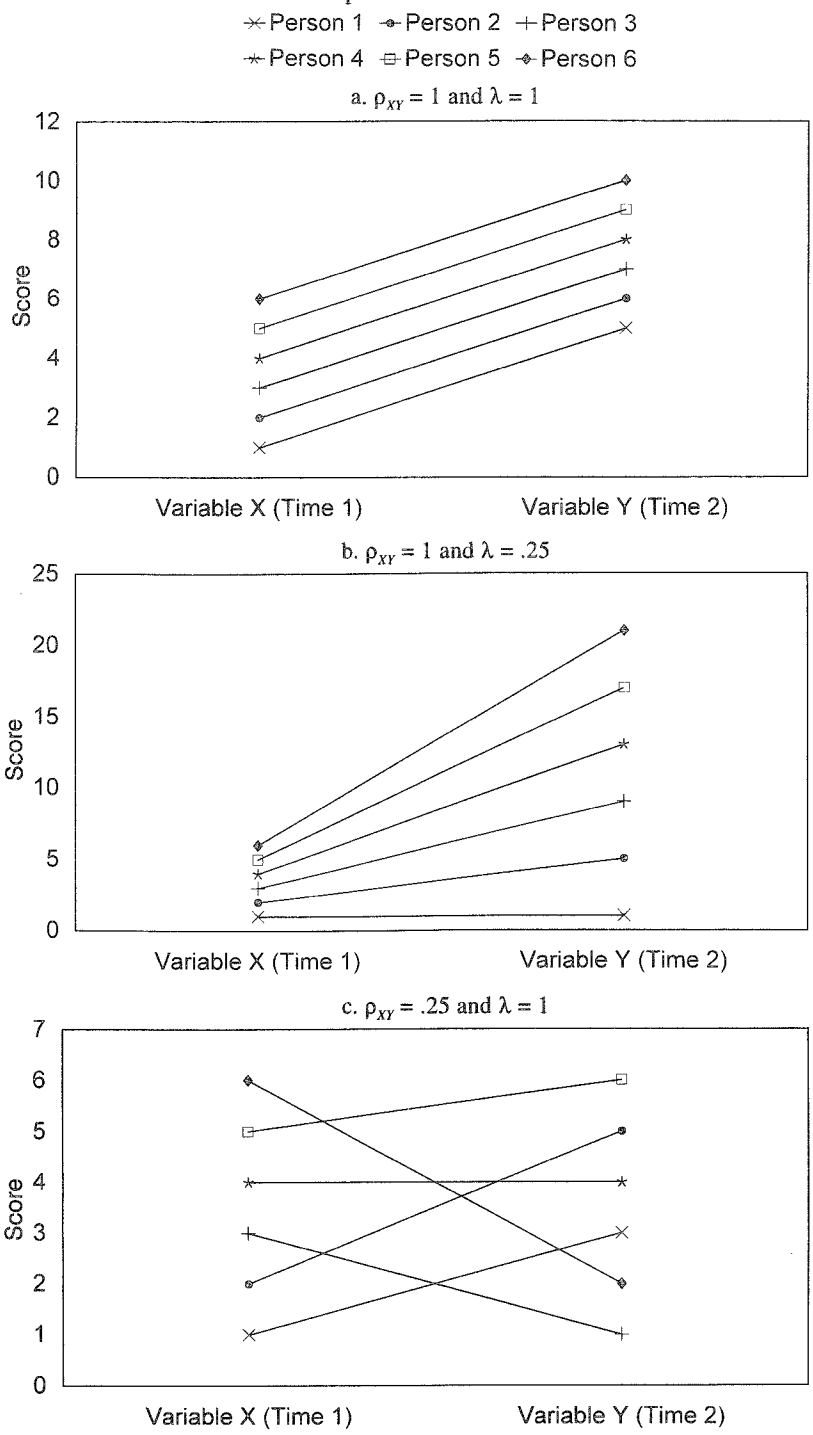
Is Reliability Obsolete?

Reliability is a concept intended to reflect the precision of a measure. When we use gain scores, we are attempting to measure change between pretest and posttest. The measure illustrated in Figure 1a is unreliable. Does it therefore follow that it is an imprecise measure of change?

I would argue that it does not. When change over time is of interest, the emphasis is usually on tracing growth or decline within persons. This change is reflected in intraindividual variability. Yet there is nothing in the definition of reliability in Equation 2 that includes intraindividual variability. Suppose there is no error in the data illustrated in Figure 1a, that is, the instrument in question perfectly measures intraindividual change over time. According to Equation 2, this instrument has a reliability of 0. Its properties as a measure of change are not reflected in the reliability. Thus, it is possible for a measure to show poor reliability, even to have a reliability of 0, and yet to be a highly precise measure of change (Collins, 1996). This simple fact makes the reliability of gain scores moot in many situations. I do not make this point to detract in any way from the elegant and highly useful presentation of Williams & Zimmerman (1996). But I would like to pose the following question: Who cares whether gain scores are reliable, if this says nothing about whether they are precise measures of change?

Instead of relying on classical test theory ideas about reliability, it is time to develop new measurement theories and practical instrument evaluation procedures that are specifically designed for measurement of

Figure 1
Examples of Patterns of Growth



change over time. Development of these approaches will be a challenge. The theories must find a way to evaluate separately intraindividual change and interindividual differences in intraindividual change. They must find a way to distinguish real change over time from random fluctuations in error components masquerading as change. They must also provide methods for measurement not only of simple pretest-posttest gain, but of growth over multiple times of observation, which is becoming increasingly important as researchers turn to growth curve modeling (Willett & Sayer, 1994). In my opinion, the most fruitful starting point is a model of the change process, with different measurement theories necessary for different types of change processes. Some progress in this direction has been made by Fischer (1976), Embretson (1991), and Collins (1996), but much remains to be done.

A quarter of a century later, we still do not have a complete answer to the first part of Cronbach & Furby's (1970) question, "How should we measure change?" But we know much more than we once did, thanks to Williams & Zimmerman (1996), Rogosa et al. (1982), and others who have worked at clarifying these issues. Most of us remain convinced that the answer to the second part of their question, "Should we?" is a resounding "Yes!" I am optimistic that in the next several years we will see more significant progress in this fascinating area.

References

Collins, L. M. (1996). Measurement of change in research on aging: Old and new issues from an individual growth perspective. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 38–56). San Diego: Academic Press.

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

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 184–203). Washington DC: American Psychological Association.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York: Wiley.

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

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.

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

Author's Address

Send requests for reprints or further information to Linda M. Collins, The Methodology Center, The Pennsylvania State University, S-159 Henderson Building, University Park PA 16802, U.S.A. Email: lmc8@psu.edu.