Commentary on the Commentaries of Collins and Humphreys

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

The critiques of Collins (1996) and Humphreys (1996) certainly throw light on properties of gain scores and difference scores that have led to controversies in the past. Collins' examples reveal that familiar formulas for the reliability of differences do not adequately reflect the precision of measures of change, because they do not allow for intraindividual change. Some additional examples are provided here, and a similar argument is applied to the reliability of a single test. As Collins implies, these arguments indeed disclose flaws, not only in the conventional approach to the reliability of gains and differences, but also in the basic concept of reliability in classical test theory. *Index terms: change scores, classical test theory, difference scores, gain scores, intraindividual differences, measurement of growth, reliability, test theory, validity.*

Collins (1996) has made some important observations and presented instructive examples and graphs to support her argument. It is clear from her Figure 1 that a precise measure of change can have a reliability coefficient of 0. This observation is closely related to one made by Overall & Woodward (1975) many years ago: Under certain conditions, significance tests performed on difference scores are most powerful when the scores have reliabilities of 0. The Overall and Woodward "paradox," which engendered controversy at the time, also involved ambiguities in the definition and interpretation of test reliability.

Collins proceeded to ask: "Who cares whether gain scores are reliable or not if this says nothing about whether they are precise measures of change?" Perhaps we can obtain further insight into this issue by extending Collins' critique as follows. First, reliability can be defined as: (1) the proportion of observed score variance that is true score variance, (2) the correlation between parallel measurements, and (3) the squared correlation between true scores and observed scores. Under the assumptions of classical test theory, these three definitions are mathematically equivalent (Lord & Novick, 1968). If reliability, by any one of these definitions, is high (or low), then it is also high (or low) by the other definitions. This is true of difference scores, as well as ordinary test scores.

Let us now rephrase Collins' question in three ways:

1. Who cares whether or not true gain-score variance is a large part of observed gain-score variance, if this says nothing about the precision of gain scores?
2. Who cares whether or not the correlation between parallel measures of gains is high, if this says nothing about the precision of the measures?
3. Who cares whether or not the correlation between true gains and observed gains is high, if this says nothing about the precision of the measures?

At first glance, the mutual equivalence of these questions and the one asked by Collins appears to cast doubt on her approach, but further reflection reveals that it does not. Examples similar to Collins' Figure 1 reveal that the correlation between highly precise measures of change, which are parallel by the conventional definition, can be 0. Of course, "precision," like "reliability," is difficult to define with precision.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 295–297

0146-6216/96/030295-03$1.40

How should this term, or for that matter, "consistency" or "stability" be defined mathematically? Are there perhaps several equivalent definitions?

These questions all emphasize that Collins' critique has exposed something fundamental in test theory. The questions are relevant, not only to gain scores and difference scores, but also to scores on a single test. It is legitimate to ask, "Who cares whether or not the correlation between scores on parallel forms of a test is high, if this says nothing about the precision of the scores on those forms?" and so on.

Collins noted that a familiar reliability formula (her Equation 2) does not take intraindividual variability into account. In fact, none of the three definitions mentioned above reflects this source of variation. This indeed provides an explanation of some of the anomalous properties of gain scores.

There is a slight difficulty with Collins' interpretation of her Figure 1. She makes a supposition that there is no error and that the instrument, according to her Equation 2, therefore has a reliability of 0. But absence of error, together with constancy of true gains, means that both the numerator and denominator of Equation 2 are 0, and reliability is undefined. However, the example is still important and does show that highly precise measures can have 0 reliability. The variance of true gain scores in the example is 0 because there are no interindividual differences in intraindividual change. But if there is any error variance at all, no matter how small, the variance of observed gain scores will be nonzero, and reliability by Equation 2 will be 0. Similarly, if two sets of identical true gain scores are associated with small error variances of the same magnitude, then the corresponding observed scores will conform to the classical definition of parallel measurement. The correlation between these two parallel measures of gains will be 0 even though they are both precise.

Now, let us apply the same argument to a single test. Suppose six individuals have exactly the same true score (say, 100) and extremely small random error scores (in the range of, say, $\pm 10^{-10}$). The reliability of this test is 0, although one would regard the scores as highly precise and relatively free from error. Or, looking at the problem in another way, suppose the gain scores in Collins' Figure 1 comprise the initial data in an experiment, and that a researcher does not know that they originated from pretest and posttest measures. Then, the same argument applies to these six measures, which could be scores on one test. Furthermore, it is possible for coefficient α to be 0 when calculated from the item statistics of a test that is highly precise in Collins' sense.

Many authors have believed difference scores to be unreliable and misleading, and at the same time assumed that the concept of the reliability of a single test score is meaningful. Apparently, Humphreys (1996) is one of these authors. Although he rightly emphasizes the linear dependence of gain scores on their components and has doubts about the reliability of gain scores, he does not question reliability at the more fundamental level explored by Collins. It is encouraging, however, that Humphreys recognizes the importance of the scale of pretest and posttest measures in determination of the reliability of differences, which is consistent with our investigations of λ, the ratio of standard deviations of pretest and posttest scores.

All these considerations, especially the points raised by Collins, imply that the concept of reliability in classical test theory was not given a mathematically satisfactory definition in the first place. It is not coextensive with what one regards as "precision." The classical model does not quite capture the gist of what authors attempt to describe by this term and also by "consistency," "stability," and so on, and neglect of intraindividual variability is one reason. Perhaps this shortcoming of the model is the root of continuing controversies surrounding gain scores and difference scores. It also accounts for the "paradox" noted by Overall & Woodward (1975) more than two decades ago.

The mathematical formalism of correlations and variance ratios has proved to be useful for some purposes in measurement theory, but in other contexts it has failed. Although gains and differences have considerable practical significance in research, the vexatious problems associated with their investigation

compel behavioral scientists to look more closely at the axioms of the classical test theory model. The observations made by Collins (1996), Humphreys (1996), and others certainly suggest that theorists should explore alternatives and extend the theory of measurement of change conceptually in directions not envisioned by Spearman, Yule, Pearson, and their contemporaries.

References

Collins, L. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete?" *Applied Psychological Measurement, 20,* 289–292.

Humphreys, L. G. (1996). Linear dependence of gain scores on their components imposes constraints on their use and interpretation: Comment on "Are simple gain scores obsolete?" *Applied Psychological Measurement, 20,* 293–294.

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

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for the measurement of change. *Psychological Bulletin, 82,* 85–86.

Author's Address

Send requests for reprints or further information to Richard H. Williams, Department of Educational and Psychological Studies, School of Education, University of Miami, 317-B Merrick Building, P.O. Box 248065, Coral Gables FL 33124, U.S.A. E-mail: rwilliams@umiami.ir.miami.edu.