Contradictions Can Never a Paradox Resolve

John E. Overall
University of Texas Medical School

The fact that difference scores tend to be less reliable than the original measurements from which they are calculated should not be a matter of concern in testing the significance of treatment-induced change. The reliabilities of the *original measurements* are important because unreliability attenuates correlation, and substantial correlation between prescores and postscores is required for difference scores to be of value in controlling for individual differences. Reliability notwithstanding, difference scores provide superior control over true baseline differences in quasi-experimental research, whereas the analysis of covariance (ANCOVA) is generally preferable for baseline control in randomized experimental designs. *Index terms: analysis of covariance, baseline correction, difference scores, measurement of change, reliability.*

The comment by Humphreys and Drasgow (1989) is generous insofar as it begins with an acknowledgment that simple difference scores, D_i, can provide the basis for powerful tests of significance even though reliabilities of the difference scores may be 0. It is difficult to justify that acknowledgment with their subsequent conclusion that "it is always important to have high reliability of dependent variables." Difference scores *are* dependent variables when tests of significance are performed on them.

Reliability is indeed a revered concept in psychometric literature. Humphreys and Drasgow themselves seemingly understand the issues quite well, but they are afraid that "substantive researchers with sketchy psychometric educations" will be misled by any mention of a case in which reliability as usually defined is not a relevant concern. This, we feel, is overly protective of "common man" to the point of being somewhat Machiavellian.

The "new formulation of the reliability of a difference" by Humphreys and Drasgow deserves a comment in this regard. Reliability is generally understood to be a property of a measuring instrument as it is applied in some specified population. Implicit in this conception is the idea that reliability should be estimated from measurements on a random or representative sample from that specified population. Most readers of this journal appreciate the fact that reliability estimates can be manipulated by increasing the heterogeneity of the samples from which measurements derive. To include *treatment effects* as components of the true-score variance would seem like the ultimate manipulation. Reliability would no longer be a simple attribute of the measuring instrument but rather a definition of the treatment effect,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 426–428

0146-6216/89/040426-03$1.40

and will vary as a function of the treatments chosen for comparison. That is indeed a new concept of reliability!

Overall and Woodward (1975) were concerned about a spreading belief that the use of difference scores in any form should be avoided because of unreliability. It is important to understand that this was the context in which zero reliability was offered as an ad absurdum example. Going back several decades, psychologists had decried the fact that difference scores tend to be less reliable than the original measurements from which they are calculated (Guilford, 1954; Lord, 1963; Webster & Bereiter, 1963). The blind reverence for reliability, which apparently survives today, led increasing numbers to the unfounded generalization that any use of difference scores for measurement of change must be suspect.

It was against that backdrop that Overall and Woodward offered an extreme example to demonstrate that reliability of *difference scores* is not essential for powerful tests of significance of treatment effects. Again, the aim was to refute the then-widespread but irrelevant concern over the unreliability of *difference scores* when used to control for baseline differences in testing the significance of treatment-induced change.

What we did say is well represented in the initial paragraphs of the comment by Humphreys and Drasgow. We have never said, nor did we ever imply, that researchers should be unconcerned about the reliabilities of the original measurements from which difference scores may be computed. This was stated as clearly as we then knew how in the conclusion to our original note on the subject: "The reliability of the original prescores and postscores is a valid concern, but this is not true of the *decrease* in reliability resulting from combining of measurement errors in the testing of group difference scores."

We have repeated the disclaimer in rebuttals to subsequent unfounded criticisms, and we repeat it here: *Reliability of the original measurements from which difference scores are computed is a matter of concern*. It is a matter of concern because unreliability attenuates the correlation between prescores and postscores, and substantial correlation is important if difference scores are to enhance precision in the evaluation of treatment effects. As Humphreys and Drasgow correctly point out, tests on difference scores can be less powerful than tests on outcome scores alone if the pre/post correlation is low in a randomized experimental design. Given that other assumptions are met, baseline correction by analysis of covariance (ANCOVA) will provide conservative alpha protection and enhanced power even though the pre/post correlation may be low in a randomized experimental design.

The situation is quite different in quasi-experimental designs in which treatments are administered to samples from different predefined populations. Campbell and Stanley (1963) and others (Campbell & Boruch, 1975; Campbell & Erlebacher, 1970; Overall & Woodward, 1977) have convincingly demonstrated the inadequacies of ANCOVA as a means for baseline correction in quasi-experimental research where treatment groups represent samples from populations that differ in baseline values on the variable used to measure treatment effects. Contrary to Lord's (1967) conclusion that there is no adequate solution to the problem, Overall and Ashby (1989) recently completed an extensive monte carlo investigation comparing three methods of baseline correction, in which it was demonstrated that simple difference scores provide the basis for appropriately conservative tests of a null hypothesis that is defined as "no treatment-induced change in the preexisting difference between groups" in a quasi-experimental design. The ANCOVA and percentage change analyses produced seriously nonconservative results in the simulated quasi-experimental design, as predicted from the previous literature. The investigation further revealed that preliminary tests of significance undertaken to document the absence of a statistically significant baseline difference provide no protection against the nonconservative biases of ANCOVA tests in quasi-experimental designs.

In closing, we plead not to be characterized as universally advocating the use of difference scores for measurement of change in all situations, just as we ask not to be represented as advocates of

measurement error. In randomized experimental designs, ANCOVA provides more uniform alpha protection and superior power. However, there are circumstances in which simple difference scores provide the only viable baseline correction, and in those situations there is no need to be concerned that difference scores have lower reliability than the original measure from which they are calculated. The lower reliability results from reduced representation of true individual differences in the error term that is used to evaluate the significance of treatment effects.

References

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 experience: Some critical issues in assessing social programs*. New York: Academic Press.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. *Disadvantaged Child, 3,* 185–210.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

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

Humphreys, L. G., & Drasgow, F. (1989). Some comments on the relation between reliability and statistical power. *Applied Psychological Measurement, 13,* 419–425.

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. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Overall, J. E., & Ashby, B. (1989). *Measurement of change in clinical trials*. Manuscript submitted for publication.

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

Overall, J. E., & Woodward, J. A. (1977). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin, 84,* 588–594.

Webster, H., & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

Author's Address

Send requests for reprints or further information to John E. Overall, Department of Psychiatry, University of Texas Medical School, P. O. Box 20708, Houston TX 77225, U.S.A.