Distinguishing Between Measurements and Dependent Variables

John E. Overall
University of Texas Medical School

Humphreys and Drasgow (1989b) recognize two types of dependent variables: the original measurements collected in an experiment and mathematical variables that are subjected to statistical analysis. Overall and Woodward (1975) were explicitly concerned with the latter, whereas Humphreys and Drasgow contend that they were concerned with reliability of the original measurements from which difference scores may be computed. These are quite different matters. Criticisms should focus on points of disagreement, and there has never been any disagreement concerning the importance of reliability of the original measurements. The notion that treatment effects should be considered a part of the true variance for calculation of reliability estimates is rejected as stemming from their failure to understand the basic difference between reliability and validity. *Index terms: control of individual differences, difference scores, measurement of change, reliability of the marginal distribution, statistical power, within-group reliabilities.*

Humphreys and Drasgow (1989b) contend that "the term 'dependent variable' can be used in at least two different senses: (1) the outcome measure that is collected by a researcher in an experiment, and (2) the mathematical variable that is subjected to statistical analysis" (p. 429). Whereas Overall and Woodward (1975) were concerned explicitly with the power of *tests of significance* performed on simple difference scores, Humphreys and Drasgow (1989a) alternated the two meanings without distinguishing between them. First, they considered the logic of Overall and Woodward (1975) concerning power of statistical tests on totally unreliable difference scores, and then without pause they offered the conclusion that reliability of dependent variables is always a matter of concern. When challenged about the apparent contradiction, they have now claimed that the dependent variable whose reliability concerned them is the outcome variable that is collected by the researcher, not the mathematical variable that is subjected to statistical analysis.

If there is to be criticism, it should concern a matter of disagreement. We have stated repeatedly that "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" (Overall, 1989, p. 427; Overall & Woodward, 1975). To quote Overall and Woodward (1976): "We are not advocating that imprecise measurement is good. We are contending that reduction in reliability

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 432–433

0146-6216/89/040432-03$1.40

associated with calculation of difference scores is not a cause for concern with regard to testing the significance of treatment effects.''

There is a curious discontinuity in Humphreys and Drasgow's (1989b) reasoning concerning the unreliability of difference scores in the presence versus absence of true examinees × times interaction. They contend that the *decreased* reliability attendant to calculation of difference scores is inconsequential for the design of experiments *only* in the presence of ''a statistical assumption that is ordinarily incorrect for [the] data collected'' (p. 429). The ''assumption'' to which they refer is a constant difference between pretest and posttest true scores for all examinees in a treatment group (i.e., no true examinee × times interaction within treatment groups). Had they perused the literature only a little further, they would have found that Overall and Woodward (1976) clearly rejected the ''assumption'' that power of tests on unreliable difference scores is dependent on the complete absence of a true interaction component in the error term.

As has been emphasized repeatedly, Overall and Woodward (1975) used as an ad absurdum example the extreme case in which calculation of difference scores removed all true variance. Little could we foresee that intelligent readers would believe that the conclusions do not generalize to cases in which the reliability of difference scores is merely reduced to *near 0*, as in the case of a small but realistic true examinees × times interaction. If zero reliability of difference scores is of no concern for tests of significance, then merely *reduced* reliability should certainly be of no greater concern.

The utility of difference scores as a means of controlling for individual differences depends on the magnitude of the correlation between prescores and postscores; this correlation may be attenuated *either* by measurement error or by true examinees × times interaction. It does not matter which, or in what combination. Other things equal, including equal prescore and postscore variances, analysis of difference scores enhances power of tests of significance only if the pre-post correlation exceeds .5. In general, higher correlation increases the utility of difference scores.

Finally, let the record show that this writer originally suggested to the Editor that Professor Humphreys and his colleague be allowed to withdraw without embarrassment their ''new concept of reliability''. Reliability is concerned with accuracy in discrimination among individuals, and it has traditionally been considered to be a property of the measuring instrument. To include experimentally induced treatment effects as a component of the true variance that defines reliability blurs the distinction between reliability and validity, and it renders reliability a gauge of effect size rather than an attribute of the measuring instrument. An instrument would have as many ''reliabilities'' as there are experimental treatments to be evaluated. Whereas that appears acceptable to Humphreys and Drasgow, surely it will not supplant the more conventional notion of reliability. Perhaps a different name might make their treatment-dependent concept more acceptable (e.g., quasi-validity or partially discriminant validity).

References

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

Humphreys, L. G., & Drasgow, F. (1989b). Paradoxes, contradictions, and illusions. *Applied Psychological Measurement, 13,* 429–431.

Overall, J. E. (1989). Contradictions can never a paradox resolve. *Applied Psychological Measurement, 13,* 426–428.

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

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin, 83,* 776–777.

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.