Paradoxes, Contradictions, and Illusions

Lloyd G. Humphreys and Fritz Drasgow
University of Illinois

There is no contradiction between a powerful significance test based on a difference score and the necessity for reliable measurement of the dependent measure in a controlled experiment. In fact, the former requires the latter. In this paper we review the conclusions that were drawn by Humphreys and Drasgow (1989) and show that Overall's (1989) "contradiction" is an illusion derived from imprecise language.
Index terms: analysis of covariance, baseline correction, control of individual differences, difference scores, measurement of change, reliability of the marginal distribution, statistical power, within-group reliabilities.

In a recent paper (Humphreys & Drasgow, 1989), we argued that researchers should "always worry about the measurement properties of dependent variables" (p. 420). Overall (1989) disagreed with our distressing recommendation, evidently concluding "Don't worry (about reliability); be happy." Unfortunately, Overall's arguments fail to address our main points and we must continue to insist on the importance of reliability.

Three different lines of argument led to our conclusion that reliability is important. Because Overall (1989) did not accurately represent these arguments, we will review each in turn.

Dependent Variables

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. The "contradiction" that Overall (1989) found in our earlier article is a result of his confusing these two meanings of "dependent variable".

Equations 1 through 4 in Humphreys and Drasgow (1989, p. 419) show that different conclusions can be drawn about the importance of the reliability of an outcome measure and a particular mathematical variable. First, these equations show that the power of a t test *is* directly related to the reliabilities of the outcome measures in the one-sample, pretest-posttest design (the standard error used in the t test is a function of the squared standard errors of measurement of the outcome variable). Second, given a statistical assumption that is ordinarily incorrect for data collected with this experimental design, the equations

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 429–431

0146-6216/89/040429-03$1.40

show that the power of the *t* test is *not* related to the reliability of the mathematical variable that is analyzed (the difference score between pretest and posttest in the numerator of the *t* test).

What advice should researchers glean from these two discrepant conclusions? When designing a study, the first conclusion clearly points to the importance of careful measurement of the outcome variable. The second conclusion has no implications for designing a one-sample, pretest-posttest study; it is a statistical curio.

Revivifying Reliability

It is possible to define the reliability of a dependent measure in a controlled experiment in a way that eliminates the statistical curio. Specifically, reliability can be computed using the marginal distribution of the dependent variable, rather than the conditional (within groups) distributions. Overall (1989) objected to this definition because "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" (p. 426).

Should treatment effects be considered as contributing to true-score variance? Treatment effects are surely neither random nor systematic *error*. Moreover, it is well known that reliability, as ordinarily defined, is subpopulation dependent. Thus, it is only a small extension from the traditional statement that a measure has as many reliabilities as there are subpopulations to the new conception in which there are as many reliabilities as there are subpopulations and treatment effects.

The dependence of reliability on subpopulation (and effect size in the new definition) provides motivation to use better indices of measurement accuracy. The standard error of measurement is less dependent on the range of true scores in the sampled subpopulation and can therefore be considered a more direct assessment of the measurement quality of the dependent variable. The information function of item response theory can be used to determine conditional standard errors of measurement given different true scores. Consequently, it can be viewed as an index of measurement accuracy that is subpopulation independent.

Models for Difference Scores

The model used by Overall and Woodward (1975) and repeated in the first three equations of Humphreys and Drasgow (1989) assumes that true scores on the posttest differ from true scores on the pretest by 0 in the control group and by a constant in the experimental group. Subtracting pretest scores from posttest scores thus eliminates all true-score variance, and allows a powerful *t* test despite an unreliable difference score. But this model is undoubtedly wrong because people continuously learn, forget, grow, tire, and change. Consequently, attention should be shifted from the statistical curio to a study of power as a function of (1) the degree of control of individual differences obtained by forming difference scores and (2) the reliability of the difference scores.

We agree with Overall (1989) that the use of residual scores (as in the analysis of covariance) can lead to biased significance tests (i.e., the nominal alpha level does not equal the actual alpha level) in quasi-experimental designs. We also agree that the use of difference scores can potentially control individual differences effectively. These statements, however, need qualification. In Equation 7 of Humphreys and Drasgow (1989), the correlation on the left side cannot be interpreted psychologically without knowledge of the three correlations and the two variances on the right side. A statistically significant and nontrivial correlation on the left can be a function of significant and nontrivial differences in correlations, in variances, or both. A trivial and statistically nonsignificant correlation on the left can be a function of significant and nontrivial differences in correlations and in variances that have opposing effects.

The concerns listed above are relevant for Overall's (1989) unqualified recommendation that difference scores be used in quasi-experimental designs involving the assignment of intact groups to experimental conditions. This recommendation may stem from the assumption of identical true scores on pretest and posttest, which is rarely correct. Moreover, the effect described above can magnify or obscure differences. Thus, we recommend that researchers always compute all of the statistics on the right side of Equation 7 of Humphreys and Drasgow (1989) and then estimate ρ_{xd}.

References

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

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.

Author's Address

Send requests for reprints or further information to Lloyd G. Humphreys, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820, U.S.A.