Further Comments on Reliability and Power of Significance Tests

Lloyd G. Humphreys
University of Illinois, Urbana-Champaign

The controversy about the relationship between reliability and the power of significance tests exists because statisticians obtain numerical solutions by varying independently the parameters of the power of statistical tests. In contrast, researchers have empirical limitations placed on them in varying the same parameters. Reliability and power can legitimately be decoupled by selection of the population from which to sample (Zimmerman & Williams, 1986), but this is an undependable way to increase power (Humphreys, 1991). Reducing population variance by selection of the sample can be considered a special case of (and a crude approximation to) the analysis of covariance, which is also a more effective way of controlling individual differences in true scores than the use of difference scores. Both the regressed differences and the raw differences are less reliable within treatments than their components, but can have more power in statistical tests. As the reliability of derived scores increases, however, power increases. *Index terms: difference scores, error of measurement, planning experiments, power, reliability, significance tests, t tests, true scores.*

The controversy over the reliability and power issue is not a matter of principle. Instead, it represents different points of view toward statistics—as a branch of mathematics or as a research tool having empirical referents. Consider the following quotation from Zimmerman, Williams, & Zumbo (1993): ". . .a significance test 'sees' only the variability of measures, however it arises" (p. 5). The statement is true, but researchers cannot afford to be blind to the determinants of their numbers.

Zimmerman et al. (1993) illustrated the effects of varying a particular parameter of a formula without reference to the empirical implications of that variation. This approach may convey a mathematical understanding of a parameter's function, but it is confusing to persons designing, conducting, or interpreting an experiment.

Zimmerman et al. asked "Everything else being equal, how does power change as reliability changes?" (p. 1). "Everything else" did not include the choice of dependent variable measure, which was selected by the experimenter from possibilities that included derived scores such as the raw score difference between pre- and post-measures. When the specific dependent measure is included along with the parameters fixed by Zimmerman et al.—sample size, significance level, directionality, and the alternative hypothesis—reliability cannot change. If reliability changes, the variance of the observed scores will necessarily change. Zimmerman et al. implicitly assumed that the dependent measure is not held constant. In this way reliability and observed variance can be decoupled.

Observed Variance and Reliability

Zimmerman et al. posited that ". . .power always increases when population variance decreases, independently of reliability" (p. 1). The last three words can be deleted from the above statement as it applies to classical (interval) scales of measurement merely by fixing the measure of the dependent variable. To apply it to psychometric tests, however, requires more than deleting three words. Reducing measurement error in a test, perhaps by rewriting a subset of inadequate multiple-choice alternatives, increases observed variance. True score variance increases more rapidly than the observed variance, so that

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 1, March 1993, pp. 11-14

0146-6216/93/010011-04$1.45

reliability and power both increase. The convention that assumes interval scales of classical measurement theory is widely used, so the authors can be forgiven for their avoidance of test theory. However, researchers must be aware of fundamental differences.

Some stipulations. The observed score variance of any measure is required for the estimate of the population variance, which is required in turn for the standard error of a mean. From this, *t* tests, effect sizes, and estimates of power can be calculated. In this sense, the observed variance is primary, but in this sense only. Also, neither reliability nor observed variance is the primary basis for the selection of the measure of the dependent variable. The primary consideration is the relevance to or validity for the assessment of the treatment effect. An interpretation of Humphreys & Drasgow (1989) that, of two different measures of the dependent variable such as reaction time and errors, the more reliable one will always have more power is clearly fallacious.

It hardly seems necessary to stipulate that the Humphreys & Drasgow (1989) conception of the reliability of the marginal distribution of the dependent measure does not estimate the within-group reliabilities, nor do the latter estimate the former. If there is a sufficient range of item difficulties in a test of mechanical comprehension so that both genders can be measured adequately, reliabilities within-gender will be smaller than the reliability in the total sample. The gender difference in means is not measurement error variance, and it affects the size of correlations with other measures in the same population.

There is no difference in principle when two or more groups are defined by experimental treatment differences. The only issue is whether the Humphreys & Drasgow (1989) conception is useful in planning or interpreting an experiment. It allows experimenters to estimate the true score correlation between the experimental treatment and its effect on a particular dependent measure. It also allows the same estimates for covariates. Reliability theory becomes consistent. Paradoxes are not needed, but it is unconventional.

The misleading Table 1. The decoupling of reliability and observed variance just described occurs in Table 1 in Zimmerman et al. There are no possible operations that can produce the numbers in that table—if it is assumed that there is only one dependent measure represented. The authors held r_{XY} constant at .60, and the difference between means to 1.0. Assuming that Columns 1 and 2 represent the same measure with different levels of error variance, r_{XY} becomes .80 in Column 2, and power in that column is underestimated. With respect to Column 3, the number of parallel measures of a dependent variable can be increased and the variance of true scores in the sum will be increased. To retain the metric in which the variance of true scores in Column 2 was measured, the mean of the n parallel measures is required. In the latter metric, the variance of true scores remains constant for a particular dependent measure and error variance decreases. Without obtaining the mean score, the difference in means of one unit cannot hold.

Zimmerman et al.'s Table 1 illustrates that freedom to put numbers that have no empirical reference to a specific dependent measure in a formula allows the demonstration of the independence of reliability and power. Their table does not, however, provide assistance to investigators who are concerned about power in a particular experiment.

Table 2. Zimmerman et al.'s Table 2 displays useful ranges of values comparing the power of a difference score with the power of the post-measure as a function of the pre-post correlation. Humphreys & Drasgow (1989) called attention to the same principle, but did not discuss it in detail. A difference score controls individual differences in true scores more effectively than the sole use of the post-measure does down to an r_{XY} of approximately .50. When raw score variances of pre- and post-measures differ, the value of .50 can be larger or smaller, depending on the relative sizes of the two variances.

That a difference score can be more powerful than the post-measure alone under appropriate

circumstances is widely accepted, including the reduction in reliability of the difference score within groups. To reduce measurement error in pre- and post-measures, it is useful to remember that the headings of the columns in Table 2 are not fixed. When a difference score in its present form cannot provide adequate control of differences in true scores, it may be feasible to increase r_{XY} by reducing measurement error in both *X* and *Y*.

The use of sample statistics in the preceding paragraphs is justified because the principles hold in every sample. If estimation of the population value for a given statistic is not required, it would be better to write formulas in sample notation. It is easy to forget that there are as many estimates of a population parameter for a given measure as there are reliabilities of that measure.

Table 3. Zimmerman et al.'s Table 3 is interesting with respect to the importance of fixing the measure of the dependent variable. The columns headed r_{XY} do not fix the dependent measure, but the columns headed $r(T_X,T_Y)$ do. Within the last four columns [headed $r(T_X,T_Y)$], the reliability of the raw difference increases as error variance in *X* and *Y* decreases. For the first three columns of true score correlations between *X* and *Y*, the difference score is not more powerful than the post-measure score until the reliabilities of *X* and *Y* reach .90 in the third column. That is, a true score correlation of .60 is attenuated by measurement error to levels that contraindicate the use of the difference score to control individual differences. In turn, this indicates the importance to an investigator of knowing the reliability of the score selected for possible use as a dependent measure.

Table 4. I was unable to reproduce the values in the body of Zimmerman et al.'s Table 4 from the discussion and formulas presented in the text, but the direction of change is correct. Power increases as reliability increases. Researchers are always well advised to look beneath the numbers for their determinants. Table 4 includes systematically arranged examples of the operation of the principle emphasized by Humphreys & Drasgow (1989).

Control of Individual Differences

Raw difference scores. When people are assigned at random to treatment groups, a difference between pre- and post-measures should not be used to increase power. Such use can, under appropriate circumstances, increase power, but a more effective derived score—a regressed difference score—is available. In the absence of random assignment, the decision is more complex (Wainer, 1991).

Furthermore, no derived score will allow complete control over individual differences in true scores when a treatment intervenes between the pre- and post-measures. Humans are dynamic organisms. Some change in true scores is inevitable between a measure obtained before the treatment and one obtained later.

Samples with low variability. Zimmerman et al. correctly pointed out that sampling from a population having a restricted range of talent with respect to the dependent measure can increase power and will decrease reliability. As mentioned earlier, this is the only way in which true score variance can be increased or decreased once the dependent measure has been fixed. The statement about range of talent and power in Humphreys & Drasgow (1989) was corrected and amplified (see Humphreys, 1991). Power is indeed increased when the treatment is under experimental control and when the treatment does not interact with individual differences on the dependent measure. That is, the difference between means is constant at every level of score on the dependent measure.

Differences between obtained and estimated scores. These difference scores might also be called regressed difference scores and are the basic elements in the analysis of covariance. Individual differences are controlled in experiments by subtracting the post-measure estimated from the pre-measure from the observed post-measure. Covariance analysis also adjusts differences between treatment groups, to the extent permitted

by the reliability and validity of the pre-measure, which inevitably arise from random assignment.

Searching for a homogeneous population from which to sample can be considered a special case of, and a crude approximation to, the analysis of covariance. For the control of individual differences in a true experiment, which by definition requires random assignment to treatments, the analysis of covariance is the method of choice. Compared to raw differences, the regressed difference is scale free and consistently more reliable in the same set of data. Some degree of control is exercised down to a small positive correlation between pre- and post-measures. The correlation need be only large enough to compensate for the loss of one degree of freedom for each covariate.

The obvious advantage of covariance analysis over searching for a population with a restricted range of talent is that it can be applied to samples from any population. Researchers claiming generality for a principle derived from research using a wide range of talent are on firmer ground. Covariance analysis allows a test of the hypothesis that effect size does not vary with level of talent, and its potential effective control of individual differences does indeed increase power. Researchers, however, should pay as much attention to the reliabilities of the covariate and the dependent measure in covariance analysis as Humphreys & Drasgow (1989) described for the raw difference. The reliability of a regressed difference in the marginal distribution of these scores in an experiment involving two or more groups can also be conceptualized in the manner they suggested for the raw difference.

References

Humphreys, L. G. (1991) The relation of power of statistical tests to range of talent: A correction and amplification. *Applied Psychological Measurement, 15,* 267.

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

Wainer, H. (1991) Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin, 109,* 147–151.

Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of statistical tests. *Psychological Bulletin, 100,* 123–124.

Zimmerman, D. W., Williams, R. H., & Zumbo, B.D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17,* 1–9.

Acknowledgments

Preparation of this manuscript was supported by the Department of Psychology, University of Illinois, Urbana-Champaign.

Author's Address

Send requests for reprints or further information to Lloyd G. Humphreys, 603 E. Daniel Street, Champaign IL 61820, U.S.A.