Some Comments on the Relation Between Reliability and Statistical Power

Lloyd G. Humphreys and Fritz Drasgow
University of Illinois

Several articles have discussed the curious fact that a difference score with zero reliability can nonetheless allow a powerful test of change. This statistical legerdemain should not be overemphasized for three reasons. First, although the reliability of the *difference* score may be unrelated to power, the reliabilities of the variables used to create the difference scores are directly related to the power of the test. Second, with what some will regard as additional legerdemain, it is possible to define reliability in the context of a difference

score in such a way that power is a direct function of reliability. The third and most serious objection to the conclusion that the reliability of a difference score is unimportant is that the underlying statistical model used in its derivation is rarely appropriate for psychological data. *Index terms: control of individual differences, difference scores, reliability, reliability of the marginal distribution, statistical power, within-group reliabilities.*

Overall and Woodward (1975) showed that a difference score with zero reliability can nonetheless allow a powerful test of change from pretest to posttest. This surprising result can be seen easily for the one-sample, pretest-posttest design by writing the pretest score for the i th examinee as

$$Y_{i1} = \tau_i + E_{i1} \quad , \quad (1)$$

where τ_i is the i th examinee's true score on the pretest and E_{i1} is the examinee's error score on the pretest. The key assumption underlying this result is that posttest scores Y_{i2} have the form

$$Y_{i2} = \tau_i + \alpha + E_{i2} \quad , \quad (2)$$

where α , a constant for all examinees, results from the experimental treatment. Difference scores D_i are obtained by

$$D_i = Y_{i2} - Y_{i1} = \alpha + E_{i2} - E_{i1} \quad . \quad (3)$$

From Equation 3 it is obvious that the reliability of D is 0 because individual differences in true scores have been eliminated. Moreover, a powerful significance test

$$t = \frac{\bar{D}}{\{[\text{Var}(E_1) + \text{Var}(E_2)]/n\}^{1/2}} \quad (4)$$

can be obtained with a large sample size n .

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 419-425
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/040419-07\$1.60

As a consequence of the above analysis and some related work, Nicewander and Price (1983) stated that "large reliability coefficients are not necessarily desirable in controlled experiments" (p. 524) and "reliability coefficients are not useful in evaluating experimental studies" (p. 524). This article presents three comments concerning such conclusions. These comments vary in their conceptual and statistical sophistication, but all point to the importance of reliability in an experimental setting.

Misdirection

The conclusion that reliability is unimportant in an experiment is akin to a magician's misdirection. Equations 3 and 4 indeed show that the reliability of the *difference score* D is irrelevant. However, the denominator of Equation 4 contains the sum of the variances of the pretest and posttest error scores. Consequently, the *reliabilities of the pretest and posttest* are directly related to the power of the t test in Equation 4.

Some researchers may not carefully attend to the distinction between the reliability of D and the reliabilities of Y_1 and Y_2 . If researchers generalize Nicewander and Price's conclusions about D to the pretest and posttest Y_1 and Y_2 , they will *incorrectly* assume that the reliabilities of Y_1 and Y_2 are irrelevant; Equation 4 clearly shows that power is directly related to the reliabilities of Y_1 and Y_2 .

Incorrectly generalizing Nicewander and Price's conclusions to the pretest and posttest could lead to disastrous results if researchers decide to assess dependent variables by very unreliable measures. Such a course of action would increase $\text{Var}(E_1)$ and $\text{Var}(E_2)$ and consequently decrease power.

Fighting Fire With Fire: Some Legerdemain to Revivify Reliability

In this section the reliability of a difference score is conceptualized in a way that avoids Overall and Woodward's (1975) paradox. This conceptualization departs from the common understanding concerning the definition of a measure's reliability. Nevertheless, it does provide internally consistent theory in which power is *always* directly related to reliability. Moreover, it provides unequivocal advice to substantive researchers with sketchy psychometric educations: *Always worry about the measurement properties of dependent variables.*

Equivalence of Data Analysis Statistics

It is convenient for the purposes of this discussion to express the relation between an independent variable and the measure of the dependent variable as a point-biserial correlation. (If there are three or more levels of treatment, a correlation is used for each dummy variable.) For testing the hypothesis of a zero relation in the population, it is immaterial whether a standardized difference between means or a point-biserial correlation is used. The relationship between the two for the independent-groups t test is

$$t = \frac{r_{xy}(n_1 + n_2 - 2)^{1/2}}{(1 - r_{xy}^2)^{1/2}} = \frac{\bar{Y}_2 - \bar{Y}_1}{s_y[(n_1 + n_2)/n_1 n_2]^{1/2}}, \quad (5)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the pretest and posttest sample means and s_y is the pooled within-treatment estimate of the standard deviation.

The standardized difference in means in the right-hand expression in Equation 5 provides a measure of effect size without an upper limit. In contrast, the difference between the means in the point-biserial correlation is standardized by dividing by the sample standard deviation of the marginal distribution. This provides, after multiplying by $(pq)^{1/2}$, for an effect size with an upper limit of unity. However, there is a one-to-one, albeit markedly nonlinear, relation between the two ways of measuring effect size. It is

well known that power is directly related to effect size as measured using a standardized mean difference (see, e.g., Kraemer, 1985, p. 183). It follows, therefore, that power is directly and monotonically related to effect size as measured in terms of the point-biserial correlation.

Reliabilities of Components and the Correlation With Differences

In the development that follows, it is assumed that examinees are randomly assigned to either an experimental or a control condition, and a pretest and posttest are administered to each examinee. Let X denote the independent variable with two levels of treatment, Y_1 and Y_2 represent pretest and posttest measures, D represent the observed difference between pretest and posttest, and primes represent parallel measurements. It is also necessary to distinguish between true-score and observed-score parameters. Appropriate symbols for each, and the relations between them found in classical measurement theory, are as follows:

	Mean	Variance	Covariance	Correlation
True score parameter	μ_y	$\sigma_{\tau y}^2$	$\sigma_{x\tau y}$	$\rho_{x\tau y}$
Observed score parameter	μ_y	$\sigma_{y y'}^2 / \rho_{yy'}$	$\sigma_{x\tau y}$	$\rho_{x\tau y} (\rho_{yy'})^{1/2}$

(6)

As usual, the error components of the observed scores Y_1 and Y_2 are assumed to be mutually independent, independent of true scores, and independent of the error component (if any) of X .

Now, with the aid of the relations in Equation 6 and the algebra of linear combinations, the correlation between the obtained difference scores and the independent variable can be written in terms of the components of the difference as

$$\rho_{xd} = \frac{\rho_{x\tau 2} \sigma_{\tau 2} - \rho_{x\tau 1} \sigma_{\tau 1}}{\left(\frac{\sigma_{\tau 2}^2}{\rho_{y_2 y_2'}} + \frac{\sigma_{\tau 1}^2}{\rho_{y_1 y_1'}} - 2\sigma_{\tau 2 \tau 1} \right)^{1/2}} \quad (7)$$

Although measurement error attenuates correlations and increases variance, these effects are cancelled in covariances. The variances that appear in the denominator of the fraction are increased by measurement error, and the correlation of the difference with the independent variable is decreased.

A New Formulation of the Reliability of a Difference

Although Equation 7 involves a difference only indirectly in terms of its components, a difference can have no properties other than those of its components. Consequently, classical measurement theory requires the following equality:

$$\rho_{xd} = \rho_{x\delta} (\rho_{dd'})^{1/2} \quad (8)$$

where δ is the true score corresponding to the observed difference score D . The relation given in Equation 8 leads to an interesting question: How must the reliability of a difference score be defined if Equation 8 is to be true? With such a definition, the paradox vanishes and what remains is a direct relation between reliability of the dependent measure and power. Such a simple relation is unlikely to mislead substantive researchers.

The expression for the reliability of a difference between pretest and posttest that satisfies Equation 8 could be written in terms of the components of the difference, but this would unduly complicate the equation. For simplicity, differences between pretest and posttest scores are used. Subscripts of the difference score refer to the two treatment groups. The reliability required by Equation 8 is

$$\rho_{dd'} = \frac{\sigma_{d_1d_1} + \sigma_{d_2d_2} + (\mu_{d_1} - \mu_d)^2 + (\mu_{d_2} - \mu_d)^2}{\sigma_{d_1}^2 + \sigma_{d_2}^2 + (\mu_{d_1} - \mu_d)^2 + (\mu_{d_2} - \mu_d)^2} \quad (9)$$

Equation 9 uses the usual assumptions of a parallel measurement operation. Insofar as the treatment effect is concerned, the assumption is that the experimental controls operating in the first measurement situation can be reproduced in the second. The difference between means resulting from the experimental treatment contributes non-error variance to both the numerator and denominator.

The ratio $\sigma_{d_1d_1}/\sigma_{d_1}^2$ is the reliability of the difference between pretest and posttest scores for group 1. Similarly, $\sigma_{d_2d_2}/\sigma_{d_2}^2$ is the reliability of the difference score for group 2. These within-group reliabilities can approach 0 as the control of individual differences in true scores approaches 100% (Overall & Woodward, 1975). However, the between-group variances $(\mu_{d_1} - \mu_d)^2 + (\mu_{d_2} - \mu_d)^2$ appear in both parts of the fraction, contributing nonzero variance to the marginal distribution of the difference. As long as there is an experimental effect, the reliability of the marginal distribution approaches 1.00 as within-group reliabilities approach 0. As it is the marginal distribution that is correlated with the independent variable, it is the reliability of the marginal distribution that affects the size of the correlation. Therefore, the power of the statistical test increases directly with the reliability defined by Equation 9 and there is no paradox.

The Question of the Model

Overall and Woodward (1975) assumed identical true scores on the pretest and posttest for each examinee. Consequently, only error remains when the pretest is subtracted from the posttest to define the difference score. This shows why investigators should be concerned with the reliabilities of the components of the difference. More is involved, however, than measurement error.

Examinees continuously learn, forget, grow, tire, and change. Hence Equation 2 is ordinarily invalid; a conceptually more appropriate model is

$$Y_{i2} = \tau_i + \delta_i + \alpha + E_{i2} \quad , \quad (10)$$

where δ_i is the change in examinee i from pretest to posttest.

Equation 7 assumed that the true scores of pretest and posttest were different. They are expected to be positively correlated, but this correlation will only occasionally approach unity. This means that the reliability of the difference in Equation 9 is unlikely to approach unity because of the problem of achieving vanishingly small measurement error on the pretest and posttest occasions, and because residual true-score variance is ordinarily present in the difference. Hence it is useful to look at examples of the effects of reliabilities of components, reliability of the difference, and the true-score correlation between components. This exercise will portray the separable direct influences on power of degree of control of individual differences by the use of difference scores and the reliability of these differences.

Numerical Examples of the Phenomena

Table 1 relates five levels of within-group reliabilities to the correlations of differences with the independent variable and to the reliabilities of the difference scores. This is done for two levels of effect size (in standard-score units) and three levels of pretest-posttest correlations. The highest level of the pretest-posttest correlation for a given effect size is the maximum possible, while the two lower levels are arbitrary selections representing a range of possible values. Within-group true-score variances were set equal to 1.00, and the within-group reliabilities were taken as constant across experimental and control groups and across pretest and posttest.

Table 1
Effects of Within-Group Reliabilities, Pretest-Posttest
Correlations, and Effect Sizes on ρ_{xd} and ρ_{dd}

Effect and Effect Size ^a	$\rho_{\tau_1\tau_2}$	Within-Group Reliabilities				
		.99	.90	.80	.70	.50
Effect on ρ_{xd}						
$\mu_{\tau_2} - \mu_{\tau_1} = .40$ or $\rho_{xd} = .196$						
	.981 ^b	.815	.390	.272	.211	.140
	.800	.306	.252	.210	.178	.129
	.500	.196	.179	.162	.146	.115
$\mu_{\tau_2} - \mu_{\tau_1} = 1.10$ or $\rho_{xd} = .482$						
	.876 ^b	.968	.759	.614	.511	.362
	.700	.646	.571	.501	.440	.334
	.500	.506	.468	.427	.387	.309
Effect on ρ_{dd}						
$\mu_{\tau_2} - \mu_{\tau_1} = .40$ or $\rho_{xd} = .196$						
	.981 ^b	.664	.153	.074	.045	.020
	.800	.953	.648	.450	.323	.170
	.500	.981	.821	.671	.543	.338
$\mu_{\tau_2} - \mu_{\tau_1} = 1.10$ or $\rho_{xd} = .482$						
	.876 ^b	.937	.576	.377	.261	.131
	.700	.972	.760	.585	.451	.260
	.500	.983	.839	.699	.575	.367

^aFor convenience pretest and posttest within-group variances were set equal to 1.00 in these computations. Given this metric and random assignment to conditions, the true score effect size is the mean difference in posttest scores or is the difference between pre- and posttest scores. The point biserial represents the posttest difference and is presented for purposes of comparison with the ρ_{xd} values.

^bThis value is the maximum possible given the effect size in the first column and random assignment to treatments; the lower values were selected to cover a range of more realistic values.

Table 1 illustrates several different principles:

1. The point-biserials decrease across the rows of the table as the reliability of the difference decreases, showing the direct relation between power and reliability of the difference.
2. Moving down the columns of the table, as the correlation between pretest and posttest decreases, the size of the point biserials at each effect size also decreases, showing the direct relation between power and control of individual differences.
3. As the correlation between pretest and posttest increases, the reliability of the difference becomes more sensitive to small decreases in within-group reliabilities.
4. The contribution of the treatment effect to the reliability of the dependent measures, hence to the power of the statistical test, can be swamped by moderate decreases in within-group reliabilities when the treatment effect is weak.

5. Under several combinations of parameters that appear in the table, the comparison of posttest means is more powerful than the comparison of the mean differences. The reliabilities of pretest and posttest and the size of the pretest-posttest true-score correlation are critical elements in the decision to use a difference as a dependent measure.
6. Reliabilities of, say, .15 can be modest by usual standards when the second variable (i.e., the experimental treatment) is measured with high reliability. In the present example the latter is 1.00, which is not an unrealistic value for the independent variable in an experiment.
7. Each ρ_{xd} in a given row of the table corrects to a common value after allowing for the attenuation induced in the correlation by measurement error in the difference score. This common value is the maximum correlation possible for a given effect size and a given test-retest correlation.

Related Issues

The above discussion has shown that there is a direct relation between power and the reliability of difference scores in an experiment. The effects on power due to the control of individual differences were distinguished from the effects of that control on within-group reliabilities. Note, however, that difference scores have limitations in the control of individual differences. In Equation 7 difference scores become the equivalent in power to the direct comparison of posttest means at about $\rho_{r1r2} = .50$.

This discussion has not covered research in which random assignment to treatments is absent or in which experimentally controlled treatments are absent as well. The discussion of the direct relation between reliability and power readily generalizes to the first of these two cases. The size of the effect increases the reliability of the differences, and power is still directly related to the reliability of the dependent measure in the quasi-experiment. For the second case, the direct relation between reliability and power has been widely known for many years.

Also omitted from this discussion is Zimmerman and Williams' (1986) recent note on reliability and power. They posited a second paradoxical relationship under the assumption of a constant effect size in two populations differing in range of talent. The larger true-score variance and higher reliability in the wider range, given their assumption, was associated with a decrease in power. There is a problem, however, with the derivation that led to this conclusion. Given identical independent variables and identical measures of the dependent variable, it is not possible to hold effect size constant from one range of talent to another. Power is directly related to reliability of a given dependent measure across ranges of talent as well as within each range.

Conclusions

A message can be extracted from Nicewander and Price (1983) and from Overall and Woodward (1975) that low reliability of a dependent measure in experimental research is not a problem. Unfortunately, this message is all too likely to foster bad research. The present paper attempts to convey the unambiguous message that it is always important to have high reliability of dependent variables. Increases in reliability—or its obverse, decreases in random measurement error in the marginal distribution of the dependent measure in an experiment—always increase power for a fixed effect size.

References

Kraemer, H. C. (1985). A strategy to teach the concept and application of power to statistical tests. <i>Journal of Educational Statistics</i>, 10, 173–195.	Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. <i>Psychological Bulletin</i>, 94, 524–533.

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

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.