Reliability of Measurement and Power of Significance Tests Based on Differences

Donald W. Zimmerman, Carleton University

Richard H. Williams, University of Miami

Bruno D. Zumbo, University of Ottawa

The power of significance tests based on difference scores is indirectly influenced by the reliability of the measures from which differences are obtained. Reliability depends on the relative magnitude of true score and error score variance, but statistical power is a function of the absolute magnitude of these components. Explicit power calculations reaffirm the paradox put forward by Overall & Woodward (1975, 1976)—that significance tests of differences can be powerful even if the reliability of the difference scores is 0. This anomaly arises because power is a function of observed score variance but is not a function of reliability unless either true score variance or error score variance is constant. Provided that sample size, significance level, directionality, and the alternative hypothesis associated with a significance test remain the same, power always increases when population variance decreases, independently of reliability. *Index terms: difference scores, error of measurement, power, significance tests, t test, test reliability, true scores.*

The relation between the reliability of difference scores and the power of significance tests based on difference scores has been a troublesome issue in psychometrics for more than a decade. Overall & Woodward (1975, 1976) showed that a paired-samples Student *t* test can be powerful, even though the reliability of the difference scores from which the *t* statistic is calculated is 0. This extreme example engendered controversy, and several authors (e.g., Fleiss, 1976; Nicewander & Price, 1983; Zimmerman & Williams, 1986) expressed opinions on the issue.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 1, March 1993, pp. 1-9

0146-6216/93/010001-09$1.70

Recently, this issue has again come into prominence in an interchange of views (Humphreys & Drasgow, 1989a, 1989b; Overall, 1989a, 1989b). The issue is investigated here using concepts of statistical power analysis developed by Cohen (1988, 1990), originally in 1969. Calculations are presented here that demonstrate how the statistical power associated with difference scores is influenced by reliability. The results of these calculations reemphasize the importance of the paradox noted by Overall & Woodward (1975, 1976). The present paper extends to difference scores some methods originally used to find relations between reliability and power associated with a single measurement (Williams & Zimmerman, 1989; Zimmerman & Williams, 1986).

Determinants of the Power of Significance Tests

The power function of a significance test is determined by sample size, population variance, significance level, the alternative hypothesis, and directionality, as well as the use of information in sample data by the test statistic. The influence of population variance and the alternative hypothesis can be combined into a measure of effect size (Cohen, 1988). Therefore, the question "How does statistical power depend on reliability?" can be rephrased to "Everything else being equal, how does power change as reliability changes?" The answer is that statistical power does not change at all under these conditions. If the variables just mentioned have fixed values, then the power of a significance test is completely

determined and is independent of the reliability of measurement.

This is not really a serious paradox; it means, however, that the wrong question has been asked. The problem is that reliability can change even when population variance does not, or reliability can stay the same while population variance changes. The concept of reliability arises from partitioning observed score variance into true and error components. The reliability coefficient is by definition the proportion of observed score variance that is true score variance, or the proportion of total variance that is accounted for by variation among individuals or experimental objects, and not by error of measurement (e.g., Lord & Novick, 1968).

The power of a significance test, on the other hand, is a function of the total observed score variance and does not depend on how that variance is partitioned into true and error components. Thus, power changes as reliability changes only if observed score variance changes at the same time. Therefore, in order to make the original question meaningful ("How does statistical power depend on reliability?"), it is necessary to assume that a change in reliability can be attributed to a change in true score variance, a change in error score variance, or some known combination of the two.

To be more specific, statistical power is a monotonically decreasing function of population variance, apart from any consideration of reliability. If true score variance is fixed, then reliability increases as error of measurement decreases; but under these conditions power increases, because population variance decreases. Conversely, if error score variance is fixed, then reliability increases as true score variance increases, but under these conditions power decreases, because population variance increases. Mathematically, this means that statistical power is related to the reliability of measurement, but is not a function of the reliability of measurement unless either true score variance or error score variance is constant. In other words, reliability is determined by the relative magnitude of true and error components of variance, but the power of a significance test is determined by the absolute magnitude of the total variance.

This conceptualization of the relation between reliability and power is general and not exclusive to an experimental design involving treatment effects or differences between groups. For example, it is meaningful when applied to a significance test of the hypothesis that a population mean has some specific value. It applies to both parametric and nonparametric tests—one sample, two sample, and many sample versions. It also applies to both simple and composite hypotheses, as well as simple and composite alternatives. In all cases, the concepts are essentially the same as outlined above. This generality reveals the limitations of the idea suggested by Humphreys & Drasgow (1989a) that experimental treatment effects are part of true score variance (see also the rebuttal by Overall, 1989b). If this notion were applied to a one-sample significance test, it would be necessary to associate a different reliability coefficient with every conceivable alternative hypothesis.

Power of Significance Tests Based on Differences

When the relation between reliability and power of significance tests based on difference scores is investigated, some further complications arise. This paper considers, first, how the power of a test based on differences, such as the paired-samples t test, is related to the power of the same test applied to separate measures. Also examined is how the power of a test based on differences is related to the reliability of the separate measures. Once again, power is a function of the variance of the difference, but not a function of the reliability of the difference. This is because the variance of a difference depends not only on the variances of the separate measures, but also on the correlation between the measures—that is, if $D = Y - X$, then

$$\mathrm{Var}(D) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\rho_{XY}[\mathrm{Var}(X)\mathrm{Var}(Y)]^{1/2}. \quad (1)$$

Moreover, reliability influences the correlation ρ_{XY} by virtue of the relation

$$\rho_{XY} = \rho(T_X, T_Y)(\rho_{XX'}\rho_{YY'})^{1/2}, \quad (2)$$

that is, by attenuation of the correlation between true scores. If the reliabilities of the pre-measures increase because error of measurement decreases, then the variance of difference scores decreases, because the variance of X [Var(X)] and the variance of Y [Var(Y)] decrease and also because ρ_{XY} increases. Also, the power associated with the difference scores increases. Unfortunately, these relations are not obvious from the usual formulas for the reliability of difference scores.

Power Considered as a Function of Reliability: Some Counterexamples

The relation between reliability and statistical power can be demonstrated by explicitly calculating the power of some significance tests when reliability is known. The examples in this section make assumptions about true score variance, error score variance, and reliability, and the method of power calculations introduced by Cohen (1988), Welkowitz, Ewen, & Cohen, (1976), and Howell (1987) is employed. The results of the calculations are shown in Table 1.

Each column in the table represents a separate power calculation, based on one-sample Student t tests performed on difference scores and on the pre-measures. Each test, based on $N = 30$, was assumed to be performed at the .05 significance level, was directional, and the difference between the true mean and the hypothesized mean was 1. It was also assumed that the true score and error score variance were the same for the pre- and post-measures. Therefore, they have the same observed score variance and the same reliability coefficient. The table gives only the values for the pre-measures—the identical values for the post-measures are omitted. The correlation between pre-measures and post-measures was assumed to be .60.

The variables in Table 1 (and in the equations below) are:

For the pre-measures:

Table 1
Variances of Differences, Effect Size, and Power for Three Examples With Different Values of Pre-Measure Variances ($\rho_{XY} = .60$)

	Example		
Statistic	1	2	3
Pre-Measure			
Var(T)	12	12	32
Var(E)	8	3	8
Var(X)	20	15	40
$\rho_{XX'}$	.60	.80	.80
Differences			
Var(T_D)	0	6	16
Var(E_D)	16	6	16
Var(D)	16	12	32
$\rho_{DD'}$	0	.50	.50
Effect Size			
γ_X	.22	.26	.16
γ_D	.25	.29	.18
δ_X	1.22	1.41	.87
δ_D	1.37	1.58	.97
Power			
$P(X)$	.52	.60	.37
$P(D)$	.56	.64	.40

Var(T), the true score variance;
Var(E), the error score variance;
Var(X), the observed score variance; and
$\rho_{XX'}$, the reliability.

For the difference scores:

Var(T_D), the true difference score variance

$$\text{Var}(T_D) = 2\text{Var}(T_X)[1 - \rho(T_X, T_Y)]; \quad (3)$$

Var(E_D), the corresponding error score variance, where Var(E) = 2[Var(E)];
Var(D), the observed difference score variance

$$\text{Var}(D) = 2\text{Var}(X)(1 + \rho_{XY}); \quad (4)$$

and

$\rho_{DD'}$, the reliability of the differences scores, [Var(T)/Var(D)], or

$$\rho_{DD'} = \frac{\rho_{XX'} - \rho_{XY}}{1 - \rho_{XY}}. \quad (5)$$

The effect size statistics are:

γ_X, Cohen's measure of effect size for the pre-measures,

$$\gamma_X = \frac{\mu_1 - \mu_0}{[\mathrm{Var}(X)]^{1/2}}; \quad (6)$$

γ_D, Cohen's measure of effect size for the differences,

$$\gamma_D = \frac{\mu_{D_1} - \mu_{D_0}}{[\mathrm{Var}(D)]^{1/2}}; \quad (7)$$

δ_X, Cohen's measure that takes sample size into consideration for the pre-measures,

$$\delta_X = \gamma_X(N)^{1/2}; \quad (8)$$

and

δ_D, Cohen's measure that takes sample size into consideration for the differences,

$$\delta_D = \gamma_D(N)^{1/2}. \quad (9)$$

Power statistics:

$P(X)$, the power of a one-sample Student t test based on the pre-measures; and

$P(D)$, the power of a one-sample t test based on the differences. This test is frequently used to determine the significance of the difference in the means of the pre- and post-measures.

Values of $P(X)$ and $P(D)$, which are functions of δ_X and δ_D, were obtained from tables such as those provided by Cohen (1988), Howell (1987), and Welkowitz et al., (1976). The columns in Table 1 are for three examples that represent various assumptions about $\mathrm{Var}(T)$ and $\mathrm{Var}(E)$.

The relation between reliability and power is not a functional relation, which is apparent from inspection of Table 1. The values in the first and second columns demonstrate that as the reliability of the differences increases from 0 to .50, the power of the differences increases from .56 to .64. Comparison of the first and third columns shows that as the reliability of the differences increases from 0 to .50, the power of the differences decreases from .56 to .40. The same reliability coefficients are associated with different values of $P(X)$ and $P(D)$, which contradicts the notion that reliability and power are functionally related.

Table 1 can be considered from another point of view. All entries in the last six rows depend solely on the values of $\mathrm{Var}(X)$ and $\mathrm{Var}(D)$. Provided these remain constant, no changes in the entries in any other rows would affect entries in the last six rows. The proportion of the variance of X that is true score variance [$\mathrm{Var}(T)$ + $\mathrm{Var}(E)$] determines the reliability of X. In turn, the proportion of the variance of D which is true score variance, determines the reliability of D. However, these proportions have nothing to do with the power calculations—only the sum of $\mathrm{Var}(T)$ and $\mathrm{Var}(E)$ [i.e., $\mathrm{Var}(X)$] influences the calculations. The same is true for difference scores—only the sum of $\mathrm{Var}(T_D)$ and $\mathrm{Var}(E_D)$ [i.e., $\mathrm{Var}(D)$] affects power calculations.

The fact that power is not necessarily low when the reliability of differences is low was emphasized by Overall & Woodward (1975, 1976) and by Overall (1989a, 1989b). Example 1 in Table 1 can be regarded as one version of the original paradox discussed by Overall & Woodward (1975). For this example, $\rho_{DD'} = 0$ and $\rho_{XX'} = .60$. Nevertheless, the statistical power [$P(D)$] associated with the differences is substantial (.56) and is actually higher than the power associated with the pre-measures [$P(X) = .52$].

As another example (using the same assumptions as in Table 1), if $\mathrm{Var}(T) = 49$ and $\mathrm{Var}(E) = 1$, then $P(X) = .33$ and $P(D) = .37$. In this case, the reliability coefficients of both pre-measures and differences are very high, but the power associated with both pre-measures and differences is quite low. Thus, a low reliability of either pre-measures or differences does not preclude high statistical power, but high reliability of either pre-measures or differences does not guarantee high statistical power. The latter result is counterintuitive in somewhat the same sense as the original paradox of Overall & Woodward (1975). Like their examples, those in Table 1 are not intended primarily to be typical of practical testing situations, but are provided as counterexamples to the proposition that power is a function of reliability.

Statistical Power as a Function of Population Variance

Population variance is a major determinant of

the power of a significance test, which can be seen in power calculations in mathematical statistics. Reliability influences power only insofar as it influences population variance, which is why reliability was omitted above as a determinant of power. The same reasoning shows how the reliability of differences is related to the power associated with differences.

The other determinants of statistical power were all assumed to be constant in making the calculations in Table 1. If sample size, significance level, directionality, and the alternative hypothesis all remain constant, then the power of a significance test is indeed a monotonically decreasing function of observed variance. The same is true for the power associated with differences: Assuming that the other parameters are constant, $P(D)$ is a monotonically decreasing function of $\mathrm{Var}(D)$. How reliability affects the power of tests of differences, therefore, depends on how reliability affects the variance of differences.

Expressed in another way, a significance test "sees" only the variability of measures, however it arises. Without further information, there is no way of knowing whether the variability of N scores is accounted for by sampling error, error of measurement, or some combination of the two. Most investigators admit that a difference between two highly variable groups of examinees is difficult to detect even with a perfectly reliable measure. Conversely, a small difference between two homogeneous groups can sometimes be detected by an unreliable measure.

Power Associated with Pre-Measures and Differences

In order to demonstrate the dependence of power on reliability, it is necessary to examine first how reliability is related to observed variance. As mentioned before, a change in observed score variance can be accounted for by a change in true score variance, error score variance, or some combination of the two.

It is natural to attribute an increase in reliability to a reduction of error of measurement, even though some authors have not been explicit about this point. Textbooks (e.g., Gulliksen, 1950; Lord & Novick, 1968) present equations that show the dependence of reliability on "group heterogeneity" or "range of talent." However, this dependence frequently is treated as a kind of nuisance that complicates the meaning of test reliability. Although it is possible for reliability to increase when true score variance increases and error score variance remains fixed, this is probably not what most authors mean in discussions of the influence of error of measurement on statistical power.

If true score variance is assumed to be constant and increased reliability is attributed to a reduction of error score variance, then statistical power of pre-measures is a monotonically increasing function of reliability. Every possible value of the initial (pre-measure) true score variance determines a different function. Similar reasoning extends to the case of difference scores. If the true score variance of a difference score remains fixed, then the associated statistical power is a monotonically increasing function of the reliability of the difference.

Furthermore, under the same assumptions, the power associated with a difference score is a monotonically increasing function of the reliability of the pre-measures. In this case, the function depends on still another parameter—the correlation between pre- and post-measures.

Power Values for Difference Scores

In these derivations, $\mathrm{Var}(T)$ is held fixed. It is assumed that pre- and post-measures have the same power (and that Cohen's γ is the same for both), and that pre- and post-measures have the same reliability coefficients. These simplifications do not restrict the generality of the results.

Using Equations 6 and 8 for Cohen's γ and δ, respectively,

$$\delta_D = \frac{\delta_X}{[2(1 - \rho_{XY})]^{1/2}} . \tag{10}$$

Equation 10 shows that the power associated with

differences depends strongly on the correlation between pre- and post-measures. The power associated with differences exceeds that associated with the pre-measures when ρ_{XY} is high, because a correlation reduces the variance of a difference.

Correlations can be attenuated by a person × time interaction in a repeated measures design, by measurement error, or both. In a simulation study, Overall & Ashby (1991) varied interaction and measurement error independently and showed that it makes no difference whether lack of perfect correlation results from interaction with perfect reliability or from unreliability with no interaction. In other words, the analysis "sees" only the correlation, whatever its source.

Table 2 shows $P(D)$ as a function of $P(X)$ for selected values of ρ_{XY}. Equation 10 shows that when $\rho_{XY} = .50$, $P(D) = P(X)$. Table 2 shows that when $\rho_{XY} > .50$, $P(D) > P(X)$, and when $\rho_{XY} < .50$, $P(D) < P(X)$. This result contrasts markedly with the more familiar relationship in which a high value of ρ_{XY} reduces the reliability of difference scores relative to the reliability of the pre-measures.

Table 2
$P(D)$ as a Function of $P(X)$ and ρ_{XY}

	ρ_{XY}				
$P(X)$	0	.20	.40	.60	.80
.10	.10	.10	.10	.10	.11
.20	.15	.16	.18	.22	.34
.30	.20	.22	.26	.34	.54
.40	.26	.30	.36	.47	.71
.50	.32	.37	.44	.58	.83
.60	.38	.44	.53	.69	.91
.70	.45	.53	.63	.78	.96
.80	.55	.63	.73	.88	.99
.90	.66	.74	.84	.95	.99

Reliability of Differences as a Composite Function of Reliability of Pre-Measures

The identity

$$\rho_{DD'} = \frac{\rho_{XX'} - \rho_{XY}}{1 - \rho_{XY}} \tag{11}$$

can be interpreted as expressing $\rho_{DD'}$ as a function of $\rho_{XX'}$ with ρ_{XY} as a parameter. After inspection of this simple equation, test theorists (e.g., Gulliksen, 1950) began to comment on the unreliability of difference scores. Again, however, it is important to be clear as to what question is being asked and what is being assumed. If the question is: "When $\rho_{XX'}$ changes, how does $\rho_{DD'}$ change?" then, as mentioned above, it is natural to assume that $\rho_{XX'}$ changes because error score variance is reduced. Furthermore, in examining the effect of the change on $\rho_{DD'}$, it is natural to assume that the correlation between the true scores of the separate measures remains fixed. However, these assumptions have not been explicit.

From this perspective, ρ_{XY} is itself a function of $\rho_{XX'}$ because of attenuation of the correlation between true scores, as mentioned in connection with Overall & Ashby's (1991) simulation study. For this reason, Equation 11 does not have the form $Y = f(X, a)$, where X is a variable and a is a constant. Rather, it has the form $Y = f[X, g(X)]$, where g is another function. It is an identity for any given values of $\rho_{DD'}$, $\rho_{XX'}$, and ρ_{XY} that characterize the parameters of probability distributions of random variables and differences between random variables. However, this identity does not reveal how a *change* in $\rho_{XX'}$ is associated with a change in $\rho_{DD'}$ when the correlation between true scores is fixed. In order to answer this more practical question, a derivation in which ρ_{XY} changes as $\rho_{XX'}$ changes is needed.

Table 3 shows $\rho_{DD'}$ as a function of $\rho_{XX'}$ with ρ_{XY} treated as a parameter, where the relation is given by the familiar equation:

$$\rho_{DD'} = \frac{\rho_{XX'} - \rho_{XY}}{1 - \rho_{XY}} . \tag{12}$$

Many entries in the table give the impression that $\rho_{DD'}$ is often far less than $\rho_{XX'}$. Not all combinations of $\rho_{XX'}$ and ρ_{XY}, however, yield meaningful values of $\rho_{DD'}$. The asterisks in Table 3 indicate these meaningless combinations in which reliability is a negative number. The reason for these impossible results is that, because of the Cauchy-

Schwartz inequality (Feller, 1966, p. 151), ρ_{XY} cannot exceed $\rho_{XX'}$. This is another consequence of attenuation resulting from error of measurement.

Table 3
$\rho_{DD'}$ as a Function of $\rho_{XX'}$ for Values of the Correlation of Observed Pre-Measures and Observed Post-Measures (ρ_{XY}) and Their True Score Components [$\rho(T_X, T_Y)$]

	ρ_{XY}				$\rho(T_X,T_Y)$			
$\rho_{XX'}$	.20	.40	.60	.80	.20	.40	.60	.80
.10	*	*	*	*	.08	.06	.04	.02
.20	0	*	*	*	.17	.13	.09	.05
.30	.13	*	*	*	.26	.20	.15	.08
.40	.25	0	*	*	.35	.29	.21	.12
.50	.38	.17	*	*	.44	.38	.29	.17
.60	.50	.33	0	*	.55	.47	.38	.23
.70	.63	.50	.25	*	.65	.58	.48	.32
.80	.75	.67	.50	0	.76	.71	.62	.44
.90	.88	.83	.75	.50	.88	.84	.78	.64

A more informative way of exhibiting the relation between reliability of pre-measures and differences also is provided in Table 3. The correlation between true scores is treated as a parameter, and all combinations of $\rho_{XX'}$ and $\rho(T_X, T_Y)$ are admissible. The entries are obtained from

$$\rho_{DD'} = \frac{\rho_{XX'} - \rho_{XX'}\rho(T_X, T_Y)}{1 - \rho_{XX'}\rho(T_X, T_Y)}, \tag{13}$$

which expresses the dependence of $\rho_{DD'}$ on $\rho_{XX'}$ in a form that is more useful in answering practical questions. For example, if $\rho_{XX'}$ is increased from .50 to .90, and $\rho(T_X, T_Y)$ remains fixed at .40, how is $\rho_{DD'}$ influenced? In this situation, $\rho_{DD'}$ increases from .38 to .84. Of course, as $\rho_{XX'}$ changes, ρ_{XY} will change, although the latter variable is not included explicitly in the equation. Using Equation 13, the reliability of difference scores does not seem quite as anomalous as when using Equation 11, especially when $\rho(T_X, T_Y)$ is not extremely high.

Power Associated with Differences and Reliability of Pre-Measures

The reliability of a difference depends inversely on the magnitude of ρ_{XY}. On the other hand, powerful significance tests are associated with high values of ρ_{XY}. In examining how changes in $\rho_{XX'}$ influence the power associated with differences, it is necessary to take into account how changes in reliability influence ρ_{XY}.

Using superscripts A and B to represent arbitrarily selected values of reliability and the same superscripts to represent the corresponding values of ρ_{XY} and δ_D,

$$\frac{\delta_D^A}{\delta_D^B} = \frac{(1 - \rho_{XY}^B)^{1/2}(\rho_{XX'}^A)^{1/2}}{(1 - \rho_{XY}^A)^{1/2}(\rho_{XX'}^B)^{1/2}}, \tag{14}$$

where the change in ρ_{XY} is given by

$$\rho_{XY}^B = \frac{\rho_{XY}^A \rho_{XX'}^B}{\rho_{XX'}^A}. \tag{15}$$

Combining these results,

$$\frac{\delta_D^A}{\delta_D^B} = \frac{(\rho_{XX'}^A - \rho_{XX'}^B \rho_{XY}^A)^{1/2}}{(\rho_{XX'}^B - \rho_{XX'}^B \rho_{XY}^A)^{1/2}}. \tag{16}$$

Once again, this result describes a family of functions with parameters. Table 4 shows the change in the power associated with differences that occurs as a consequence of a change from a pre-measure reliability to post-measure reliability. It is assumed, as before, that the change in reli-

Table 4
Power for $\rho_{XX'}^B$ and Three Levels of $P(D)$ as a Function of $\rho_{XX'}^A$ for Two Values of ρ_{XY}

ρ_{XY} Value and $\rho_{XX'}^A$	$\rho_{XX'}^B$	$\rho_{D'}^A$ .40	.60	.80
$\rho_{XY} = 0$				
.50	.65	.48	.70	.89
	.80	.55	.77	.94
	.95	.61	.83	.96
.65	.80	.46	.68	.87
	.95	.52	.74	.91
.80	.95	.45	.66	.86
$\rho_{XY} = .50$				
.50	.65	.60	.83	.96
	.80	.88	.98	.99
	.95	.99	.99	.99
.65	.80	.55	.77	.94
	.95	.74	.93	.99
.80	.95	.52	.74	.91

ability is accounted for entirely by reduction in error variance. It is not necessary to include particular values of true score variance or error score variance when the change is exhibited in this way.

The remaining columns represent the initial power (.40, .60, or .80), and the entries in the table are the augmented power resulting from the increase in reliability. Table 4 shows that improvement in reliability of pre-measures by reducing error of measurement can result in substantial increases in the power of significance tests of differences.

Further Implications

These results, based on explicit power calculations, show that augmenting the reliability of measurement by reducing error score variance can make significance tests of differences more powerful. Table 4 shows that increasing reliability of pre-measures as much as .15 or .30 can increase power substantially. However, the calculations in these tables do not imply that high reliability guarantees high statistical power. Again, it is emphasized that reliability is the proportion of the total variance that is accounted for by variation among individuals, as opposed to variation among replicate measurements on an individual. On the other hand, power depends on effect size—that is, the magnitude of the treatment effect relative to population variance.

In other words, power can be low, even though reliability is high, if effect size is small. Of course, increasing an already high reliability coefficient to a still higher value (say, from .94 to .98) yields some modest increase in power (say, from .10 to .15). High reliability and high power cannot be simply equated.

Humphreys & Drasgow (1989a, 1989b) and Overall (1989a, 1989b) generally agreed that the reliability of measures is always a matter of concern. Humphreys & Drasgow insisted that reliability is always "directly" related to power. Once again, however, it is important to pay close attention to the way questions are phrased. If asked, "Are there any circumstances under which, given a choice, a lower reliability coefficient would be preferred by a researcher?" The answer is most certainly "yes."

Suppose a researcher designing an experiment to detect a treatment effect has a choice between examinees in Group A or Group B. The true score variance of Group A is 20, the true score variance of Group B is 5, and the error score variance of both groups is 5. For Group A, reliability is .80, and for Group B it is .50. But, everything else being equal, Group B provides a more powerful significance test, because the total variance is less.

On the other hand, if the reliability of either group decreased because of an increase in error of measurement, power would be lost. Reduction of reliability is beneficial only when it is accompanied by reduction of total variance. Although manipulating true score variance may have this outcome, it is normally preferable to achieve the same end by controlling the error score variance associated with a measurement technique. If that is done, then the same technique can be applied in other experiments, whatever the true score variance may be, and the power of significance tests always will be enhanced.

References

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (3rd ed.). Hillsdale NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. II). New York: Wiley.

Fleiss, J. J. (1976). Comment on Overall & Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin, 83,* 774–775.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Howell, D. C. (1987). *Statistical methods for psychology* (2nd ed.). Boston: Duxbury Press.

Humphreys, L. G., & Drasgow, F. (1989a). Some comments on the relation between reliability and statistical power. *Applied Psychological Measurement, 13,* 429–431.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of significance tests. *Psychological Bulletin, 94,* 524–533.

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

Overall, J. E. (1989b). Distinguishing between measurements and dependent variables. *Applied Psychological Measurement, 13,* 432–433.

Overall, J. E., & Ashby, B. (1991). Baseline corrections in experimental and quasi-experimental clinical trials. *Neuropsychopharmacology, 4,* 273–281.

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.

Welkowitz, J., Ewen, R. B., & Cohen, J. (1976). *Introductory statistics for the behavioral sciences* (2nd ed.). New York: Academic Press.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology, 116,* 359–369.

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

Acknowledgments

This research was supported by a Carleton University Research Grant to the first author and by a Social Sciences and Humanities Research Council of Canada Fellowship to the third author.

Author's Address

Send requests for reprints or further information to Donald W. Zimmerman, 2738 Garber St., Berkeley CA 94705, U.S.A.; E-mail Bruno D. Zumbo, zumbo@acadvm1.uottawa.ca.