Journal of Modern Applied Statistical Methods

Volume 14 | Issue 2 Article 5

11-1-2015

Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions

Donald W. Zimmerman
Carleton University

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Recommended Citation
Zimmerman, Donald W. and Zumbo, Bruno D. (2015) "Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 2 , Article 5.
DOI: 10.22237/jmasm/1446350640

Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions

Cover Page Footnote

Professor Donald W. Zimmerman passed away in December 2013 at age 82. His scholarly work was known for a precision of thought and a desire for mathematical rigor, never shying away from controversial topics. This is one of the last collaborative papers with Professor Zumbo; a collaboration that spanned over 25 years.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 9-26.

ISSN 1538 − 9472

Invited Article
Resolving the Issue of How Reliability Is Related to Statistical Power: Adhering to Mathematical Definitions

Donald W. Zimmerman
Carleton University
Ottawa, ON, CAN

Bruno D. Zumbo
University of British Columbia
Vancouver, BC, CAN

Reliability in classical test theory is a population-dependent concept, defined as a ratio of true-score variance and observed-score variance, where observed-score variance is a sum of true and error components. On the other hand, the power of a statistical significance test is a function of the total variance, irrespective of its decomposition into true and error components. For that reason, the reliability of a dependent variable is a function of the ratio of true-score variance and observed-score variance, whereas statistical power is a function of the sum of the same two variances. Controversies about how reliability is related to statistical power often can be explained by authors' use of the term "reliability" in a general way to mean "consistency," "precision," or "dependability," which does not always correspond to its mathematical definition as a variance ratio. The present note shows how adherence to the mathematical definition can help resolve the issue and presents some derivations and illustrative examples that have further implications for significance testing and practical research.

Keywords: Reliability, power, hypothesis test, error of measurement, true score, error score, observed score, difference score

The relation between the reliability of measurement, as the concept is defined in classical test theory, and the power of statistical hypothesis tests, has been investigated for many years and has engendered controversy that has not been

Professor Donald W. Zimmerman passed away in December 2013 at age 82. His scholarly work was known for a precision of thought and a desire for mathematical rigor, never shying away from controversial topics. This is one of the last collaborative papers with Professor Zumbo.

Dr. Zumbo is Paragon UBC Professor of Psychometrics & Measurement with affiliate appointments in the Department of Statistics and the Institute of Applied Mathematics at the University of British Columbia, and an Associate Editor of this journal. Email him at bruno.zumbo@ubc.ca.

completely resolved. Overall & Woodward (1975, 1976) observed that the paired-samples *t* test based on difference scores can under some conditions have maximum power when the reliability of differences is zero. That finding led to discussion as to how the power of the *t* test and other familiar hypothesis tests depends on the reliability of dependent variables in experiments (Cleary & Linn, 1959; Collins, 1996; Feldt & Brennan, 1989; Fleiss, 1976; Hopkins & Hopkins, 1979; Kopriva & Shaw, 1991; Levin, 1986; Mellenbergh, 1996, 1999; Subkoviak & Levin, 1977; Sutcliffe, 1958; Zimmerman & Williams, 1986; Zimmerman, Williams, & Zumbo, 1993), with presentation of various inconsistent points of view.

The methods introduced by Cohen (1988) have been applied widely to calculate the power of familiar hypothesis tests used in educational and psychological research. In the case of tests based on the normal distribution, such as the Student *t* and ANOVA *F* tests, those methods provide a good approximation to exact results obtained from noncentral *t* and *F* distributions. However, the concept of test reliability and validity defined in classical test theory has not been employed in power analysis with the same degree of precision (see Thomas & Zumbo, 2012).

Researchers and test users often associate the concept of reliability with terms such as dependability, precision, repeatability, and so on, assuming they are consistent with the mathematical definition in classical test theory. The classical definition is based on the decomposition of scores in a population of individuals into true scores and error scores and the relative variability of those components. In the traditional theory, each individual's test score is a sum of a true score and an error score, $X = T + E$, and the total variance (or observed-score variance) with respect to a population of individuals is a sum of the variances of the components, $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Finally, *reliability* is defined as the ratio of the true-score variance and the total variance, $\rho = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$, or equivalently as $\rho^2\left(X, T_X\right)$, the squared correlation between observed scores and true scores (Gulliksen, 1950; Novick, 1966; Lord & Novick, 1968). It is also worth noting that the numerical value of reliability can always be found solely from the ratio of σ_T and σ_E, although the combined values of the two standard deviations may differ in size. This can be seen by defining $\psi = \sigma_T / \sigma_E$ and dividing both the numerator and denominator of $\sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$ by $\sigma_T\ \sigma_E$ to obtain $\rho = \psi / \left(\psi + \psi^{-1}\right)$.

The fact that reliability in classical test theory is a *population-dependent* concept has been emphasized by Mellenbergh (1996, 1999). The concept does not apply to an individual examinee, and this fact is important in considering statistical power. Because reliability is defined as a *ratio* of two components of variance with respect to a population, a given numerical value of reliability can be associated with many different combinations of values of true-score variance and error-score variance. That fact has been at the root of many problems in analyzing how reliability is related to statistical power.

Reliability and variance heterogeneity

A familiar formula in classical test theory enables one to find reliability in one population with a particular observed-score variance when knowing reliability in another population with a different observed-score variance. The formula is

$$\rho_2 = 1 - \frac{\sigma_{X_1}^2}{\sigma_{X_2}^2}\left(1 - \rho_1\right) \tag{1}$$

where the subscripts 1 and 2 denote the respective populations. This equation was derived under the assumption that the change in observed-score variance is accounted for by a change in true-score variance, while error-score variance remains constant (Gulliksen, 1950, p 111; Lord & Novick, 1968, p 130).

In contrast to the familiar approach, if a change in observed-score variance is accounted for by a change in error-score variance, while true-score variance remains constant, the results are described by the equation

$$\rho_2 = \frac{\sigma_{X_1}^2}{\sigma_{X_2}^2}\rho_1 \tag{2}$$

which can be derived easily, although equation (1) is prominent in test theory. Whether it is more reasonable to regard a difference in the observed scores of two groups as resulting from different true-score variances or different error-score variances is problematic. Curiously, test theorists have assumed constant error-score variance in deriving equation (1), but when considering how reliability influences statistical power, have adopted implicitly the assumption underlying the relatively unknown equation (2).

It is well understood in statistics that the power of an hypothesis test is inversely proportional to the variance of any dependent variable, assuming that other determinants, including significance level, sample size, and directionality of the hypothesis, remain constant. Expressed otherwise, the power of an hypothesis test is inversely proportional to the *observed-score variance* considered in test theory, irrespective of how that variance is partitioned into true score variance and error-score variance. For this reason, if observed-score variance does not change, the power of a significance test remains the same, even when the value of the reliability coefficient changes extensively over a wide range.

Although equations (1) and (2) show how reliability changes as observed-score variance changes, for present purposes in considering statistical power, we need just the reverse, that is, equations showing how observed-score variance changes as reliability changes. Simply rearranging equations (1) and (2), we can write

$$\frac{\sigma^2_{X_2}}{\sigma^2_{X_1}} = \frac{1-\rho_1}{1-\rho_2}, \text{ and} \tag{3}$$

$$\frac{\sigma^2_{X_2}}{\sigma^2_{X_1}} = \frac{\rho_1}{\rho_2} \tag{4}$$

These forms show immediately that, if error-score variance is constant, observed-score variance is proportional to reliability, and, if true-score variance is constant, observed-score variance is inversely proportional to reliability. In turn, because of what is known about power functions, that means that, if error-score variance is constant, statistical power is inversely proportional to reliability, and, if true-score variance is constant, statistical power is directly proportional to reliability.

It is possible for a test to have high reliability and still have low power, or, conversely, to have low reliability and have high power (see, for example, the paradox originally discussed by Overall and Woodward (1975, 1976) in the context of difference scores). Furthermore, it is possible for the same reliability coefficient to be associated with different degrees of power and for different reliability coefficients to result in the same power.

A simple example illustrates some possibilities. Table 1 compares hypothetical tests, each having a large number of scores with distributions like those shown in the table. In section A, the test on the left apparently has high true scores and low error scores, so that its reliability might be expected to be high, but, because the *variance* of T_1 is much higher than that of E_1, reliability is only .096. In the test on the right, the reverse is true, and the reliability is .904, even though the true scores at first glance look small. Nevertheless, despite the difference in reliability, the two tests have the same statistical power, because the observed-score variances are the same. In section B, the two tests have the same reliability, .645, because the variances of *T* and *E*, although different, have the same ratio. However, the observed-score variances are different, and the statistical power of the test on the left is greater.

Table 1. A) Score components of two tests having substantially different reliability coefficients and the same statistical power; B) Score components of two tests having the same reliability coefficients and substantially different statistical power.

A						B					
Score Components			**Score Components**			**Score Components**			**Score Components**		
T_1	**E_1**	**X_1**	**T_2**	**E_2**	**X_2**	**T_1**	**E_1**	**X_1**	**T_2**	**E_2**	**X_2**
100	1	101	0	99	99	50	5	55	100	10	110
101	6	107	5	100	105	52	6	58	104	12	116
100	2	102	1	99	100	51	4	55	102	8	110
102	7	109	6	101	107	53	6	59	106	12	118
101	2	103	1	100	101	52	4	56	104	8	112
100	7	107	6	99	105	50	6	56	100	12	112
102	1	103	0	101	101	53	5	58	106	10	116
100	6	106	5	99	104	51	6	57	102	12	114
Variance of T_1 − 0.786			Variance of T_2 − 7.429			Variance of T_1 − 1.429			Variance of T_2 − 5.714		
Variance of E_1 − 7.429			Variance of E_2 − 0.786			Variance of E_1 − 0.786			Variance of E_2 − 3.143		
Variance of X_1 − 8.214			Variance of X_2 − 8.214			Variance of X_1 − 2.214			Variance of X_2 − 8.857		
Reliability − .096			Reliability − .904			Reliability − .645			Reliability − .645		

Power as a composite function of reliability

For investigating the relation of reliability and power, it is more convenient to examine changes in reliability with changes in true-score variance and error-score variance, as opposed to changes in observed-score variance as given by equations (1) and (2). It is then possible to express observed-score variance as a 1-1 function of reliability, provided either true-score variance or error-score variance is held constant. Then, because power is a 1-1 function of observed-score variance, it is possible in turn to express power as a composite function. Under those conditions, power is a monotonic decreasing function of observed-score variance and a monotonic increasing or decreasing function of reliability depending on which component is constant. Of course, the form of the functions depends on properties of the particular hypothesis test considered.

First, begin with the equations $\rho_1 = 1 - \sigma_{E_1}^2 / \left(\sigma_T^2 + \sigma_{E_1}^2\right)$ and $\rho_2 = 1 - \sigma_{E_2}^2 / \left(\sigma_T^2 + \sigma_{E_2}^2\right)$, solve both for σ_T^2, assumed to be constant, and set the two expressions equal. The result is

$$\frac{\rho_1 \sigma_{E_1}^2}{1-\rho_1} = \frac{\rho_2 \sigma_{E_2}^2}{1-\rho_2}$$

Then, solving for ρ_2 gives the result

$$\rho_2 = \frac{1}{1 - \dfrac{\sigma_{E_2}^2}{\sigma_{E_1}^2}\left(1 - \dfrac{1}{\rho_1}\right)} \tag{5}$$

This equation indicates how reliability changes as the variance of the error component changes, while the true-score variance remains fixed.

Alternatively, if σ_T^2 changes while σ_E^2 is constant, a similar derivation give $\rho_1 = \sigma_{T_1}^2 / \left(\sigma_{T_1}^2 + \sigma_E^2\right)$ and $\rho_2 = \sigma_{T_2}^2 / \left(\sigma_{T_2}^2 + \sigma_E^2\right)$, so that $\sigma_{T_1}^2 \left(1-\rho_1\right) / \rho_1 = \sigma_{T_2}^2 \left(1-\rho_2\right) / \rho_2$. Solving for ρ_2 gives the result

$$\rho_2 = \frac{1}{1 - \dfrac{\sigma_{T_1}^2}{\sigma_{T_2}^2}\left(1 - \dfrac{1}{\rho_1}\right)} \tag{6}$$

This equation indicates how reliability changes as true-score variance changes, while error-score variance is constant. Equations (5) and (6) clearly indicate that changes in reliability resulting from changes in either true-score variance or error-score variance depend only on the *ratios* $\sigma_{E_2}^2 / \sigma_{E_1}^2$ or $\sigma_{T_1}^2 / \sigma_{T_2}^2$ relating the old and new score components and not on the individual variances considered separately.

Changes in observed score variability and power with changes in reliability

Table 2 contains results found from equations (5) and (6). The first row at the top, labeled "Initial ρ" is the value of the reliability coefficient, denoted by ρ_1 in the equations, and the entries in the right-hand section of the table are the values of the new reliability coefficient, ρ_2, after a designated change in the error-score variance or true-score variance. The ratio of old-to-new error-score variance, $\sigma^2_{E_1} / \sigma^2_{E_2}$, is located in the first column, and the entry in the table gives the value of the new reliability after the change, assuming that true-score variance remains constant. The same entry in the table is also the value of the new reliability if a change shown by the adjacent entry in the second column is made in the ratio $\sigma^2_{T_1} / \sigma^2_{T_2}$, assuming that error-score variance remains constant. That is, the ratios in the second columns are inverses of those in the first column, and the same change in reliability corresponds to both ratios.

Table 2. Modification of reliability and observed-score variance by changes in error-score variance ($\sigma^2_{E_1} / \sigma^2_{E_2}$) and in true-score variance ($\sigma^2_{T_1} / \sigma^2_{T_2}$). Entries in the five right-hand columns are the modified reliability values (ρ_2) corresponding to variances and variance ratios in the first four columns.

				Initial Reliability (ρ_1)				
$\sigma^2_{E_1} / \sigma^2_{E_2}$	σ^2	$\sigma^2_{T_1} / \sigma^2_{T_2}$	σ^2	.10	.30	.50	.70	.90
0.250	5.000	4.000	1.250	.027	.097	.200	.368	.692
0.286	4.500	3.500	1.286	.031	.109	.222	.400	.720
0.333	4.000	3.000	1.333	.036	.125	.250	.438	.750
0.400	3.500	2.500	1.400	.043	.146	.286	.483	.783
0.500	3.000	2.000	1.500	.053	.176	.333	.538	.818
0.667	2.500	1.500	1.667	.069	.222	.400	.609	.857
1.000	2.000	1.000	2.000	.100	.300	.500	.700	.900
1.500	1.667	0.667	2.500	.143	.391	.600	.778	.931
2.000	1.500	0.500	3.000	.182	.462	.667	.824	.947
2.500	1.400	0.400	3.500	.217	.517	.714	.854	.957
3.000	1.333	0.333	4.000	.250	.562	.750	.875	.964
3.500	1.286	0.286	4.500	.280	.600	.778	.891	.969
4.000	1.250	0.250	5.000	.308	.632	.800	.903	.973

The values of ρ_2 in the right-hand section always *increase* as values of $\sigma^2_{E_1} / \sigma^2_{E_2}$ increase and also as those of $\sigma^2_{T_1} / \sigma^2_{T_2}$ *decrease*. At the same time, the values of σ^2 decrease (and therefore power increases), as those of ρ_1 increase, and vice versa. Also, the same values of ρ_2 are associated with different values of σ^2 (and therefore power).

The relationship can be seen in more detail by plotting graphs of some power functions obtained from simulations. Figure 1 plots power functions of the one-sample Student *t* test under conditions where reliability was either increased or reduced by changing one component of the observed-score variance while the other remained constant. These simulations were programmed using *Mathematica*, version 4.1 (Wolfram, 1999), together with *Mathematica* statistical add-on packages. The program performed *t* tests on sums of "true-score" and "error-score" random variables, selected from $N(0,1)$ and multiplied by constants in order to determine means, variances, and reliabilities. The means increased in increments of $.32\sigma$, and each data point in the figure was found from 20,000 iterations of the sampling procedure.

In both sections of the figure, the true-score and error-score variances were initially equal, so that reliability was .50. The middle curves with filled circles represent these initial reliabilities. In the upper section, reliability was increased to .80 in two ways. In the top curve in that section (triangular symbols), error-score variance was reduced, while true-score variance was constant. In the lower section (square symbols), true-score variance was increased while error-score variance was constant.

In the lower graph, reliability was decreased to .20 in two ways. In the top curve (square symbols), true-score variance was reduced while error-score variance was constant. In the lower curve (triangular symbols), error-score variance was increased while true-score variance was constant. All these curves, with shapes typical of power curves, show that the sum of the two variance components, that is, the observed-score variance, determined the power of the hypothesis test irrespective of how reliability changed as a result of a change in the ratio of the two components.

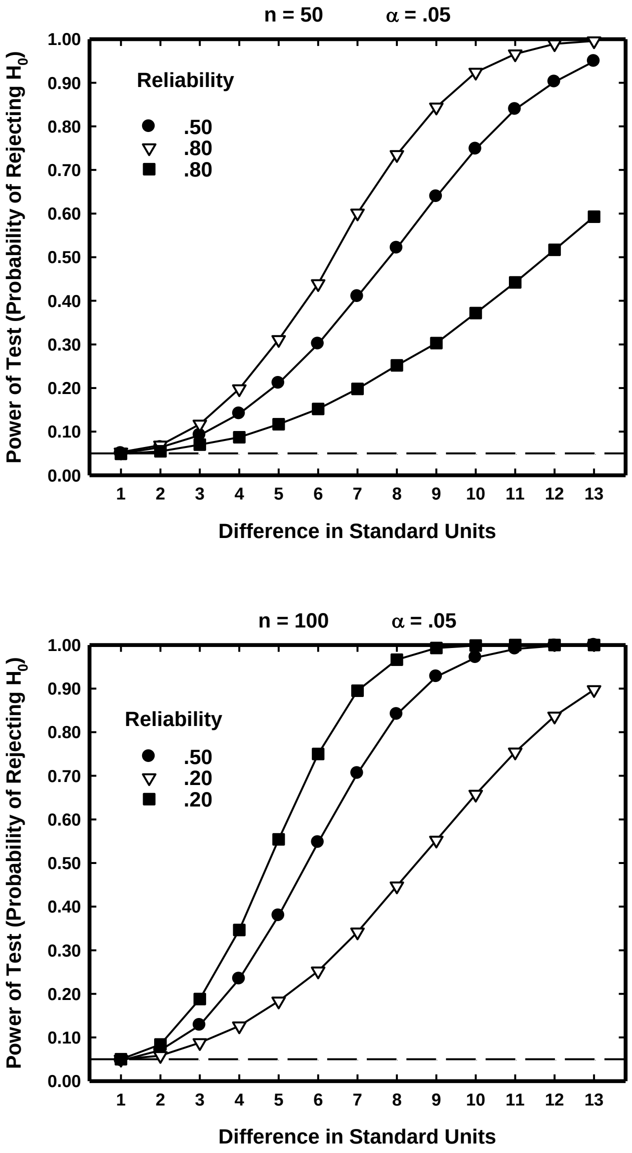

Figure 1. Power functions of the one-sample t test when reliability was increased or decreased by changing component variances. Upper graph: reliability was increased from .50 to .80. The middle curve is for ρ = .50. In the upper curve, error-score variance was reduced while true-score variance remained constant. In the lower curve, true-score variance was increased while error-score variance remained constant. Lower graph: Reliability was reduced from .50 to .20. The middle curve is for ρ = .50. In the upper curve, true-score variance was reduced while error-score variance remained constant. In the lower curve, error-score variance was increased while true-score variance remained constant.

Relations, functions, and composite functions

It is well known that statistical power is a function of several variables, some of which are under the direct control of an experimenter. These include sample size, N, the significance level, α, and the directionality of the hypothesis tested. Of course, different hypothesis tests, parametric and nonparametric, have different power characteristics under various conditions. The relations between N and power and between α and power are functional when the other variables are held constant; that is, each value in the domain of the relation is associated with a single value in its range. Some authors have considered it reasonable to add reliability to the list of determinants. However, as we have seen, reliability influences power only to the extent that it influences observed-score variance.

The association between reliability and power, therefore, is a mathematical relation, but it is not a *function* or a *functional relation*. However, it becomes functional if the variance of one of the two variables determining reliability is held constant. In that case, if the variance of one score component is held constant, power is a composite of two functions, the one between a score component and observed-score variance, and the one between observed-score variance and power. The range of the first function is the domain of the second.

As said before, still another way to express the same relationship is that, all other things equal, statistical power is a function of the sum of the variances of T and E, whereas reliability is a function of the ratio of those two variances. As noted earlier, reliability can be defined as $\psi/(\psi+\psi^{-1})$, where $\psi = \sigma_T/\sigma_E$. That definition makes it clear that reliability can be either large or small at the same time the sum, which determines power, is either large or small, independently of the ratio. The fact that power is determined by the observed-score variance, which is comprised of the sum in the denominator of the expression $\rho = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$ shows that, for a fixed value of σ_E^2, power has its maximum value when $\rho = 0$. But for a fixed value of σ_T^2 power has a maximum when $\rho = 1$.

Reliability of difference scores and statistical power

In order to gain insight into paradoxes concerning difference scores, we shall pursue an approach similar to the above. Rather than directly seeking a relationship between the reliability of differences and the power of an hypothesis test employing differences, we first consider how both are related to observed-score variance and also the reliability coefficients of the two variables determining the differences.

Once again, beginning with what is known, the power of tests on difference scores, $X - Y$, is certainly a decreasing function of the variance of the difference scores. However, reliability depends on partitioning that variance into true and error components and finding ratios, which in turn depend on the similar ratios of both *X* and *Y*. In all cases, both reliability and the power of an hypothesis test can be considered joint functions of the true-score variance and error-score variance of the difference scores. However, power is determined uniquely by their sum and reliability by their ratio, just as in the case of a single variable *X*.

A familiar equation is

$$\rho_D = \frac{\sigma_{T_D}^2}{\sigma_D^2} = \frac{\sigma_{T_X}^2 + \sigma_{T_Y}^2 - 2\rho_{T_X T_Y}\sigma_{T_X}\sigma_{T_Y}}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y} \tag{7}$$

where $D = X - Y$, T_X and T_Y are the true score components of *X* and *Y*, and ρ_D is the reliability of *D*. If $\sigma_{T_X}^2 = \sigma_{T_Y}^2$ and $\sigma_{E_X}^2 = \sigma_{E_Y}^2$, this equation can be solved for σ_D^2 and substitutions made using $\rho_X = \sigma_{T_X}^2 / \left(\sigma_{T_X}^2 + \sigma_{E_X}^2\right)$. The result is

$$\sigma_D^2 = \frac{2\sigma_{T_X}^2}{\rho_X}\left(1 - \rho_{T_X T_Y}\rho_X\right) \tag{8}$$

and an equivalent result is

$$\sigma_D^2 = 2\left[\sigma_T^2\left(1 - \rho_{T_X T_Y}\right) + \sigma_E^2\right] \tag{9}$$

Although the assumption that variances of *X* and *Y* are equal is often unrealistic in practice, it suffices to indicate the form of the relation between reliability and statistical power. Next, the reliability of differences can be written in the form

$$\rho_D = \frac{\rho_X\left(1 - \rho_{T_X T_Y}\right)}{1 - \rho_{T_X T_Y}\rho_X}, \text{ or} \tag{10}$$

$$\rho_D = \frac{\sigma_T^2\left(1-\rho_{T_X T_Y}\right)}{\sigma_T^2\left(1-\rho_{T_X T_Y}\right)+\sigma_E^2} \tag{11}$$

Equation (10) indicates that, if $\rho_{T_X T_Y} = 0$, the reliability of differences is the same as the common reliability of the components.

Equations (8), (9), (10), and (11) have the desirable feature that all combinations of values of the variables on the right-hand side of the equation yield meaningful values of ρ_D and σ_D^2. That is not true in the case of several well-known formulas that involve both ρ_{XY} and ρ_X, because the Cauchy-Schwarz inequality places limits on the values the two can have together (Zumbo, 1999). For example, the relation $\rho_D = \left(\rho_X - \rho_{XY}\right)/\left(1-\rho_{XY}\right)$ is not meaningful for all values of ρ_{XY} and ρ_X.

The above equations provide a convenient way to exhibit the relation between the reliability of differences and statistical power. Table 3 shows results of calculations using equations (9) and (11), comparing the reliability of component scores (ρ_X), the reliability of difference scores (ρ_D), and the observed variance of difference scores (σ_D^2), as a function of σ_T^2 while σ_E^2 is constant (upper section) and of σ_E^2 while σ_T^2 is constant (lower section).

If σ_E^2 is fixed, an increase in ρ_X comes from an increase in σ_T^2, and if σ_T^2 is fixed, it comes from a reduction in σ_E^2. Those outcomes are apparent in the table: As σ_T^2 increased from 0 to 1.8, the reliability coefficients ρ_X and ρ_D both increased, and also the variance of observed scores increased, so that statistical power *decreased*. The same was true for all three values of the correlation between true scores, $\rho(T_X,T_Y)$. On the other hand, as σ_E^2 increased from 0 to 1.8, ρ_X and ρ_D both decreased, but the variance of observed scores still increased, so that power again *decreased*. As σ_T^2 varied, power was greatest when the reliability of differences was 0. However, as σ_E^2 varied, power was greatest when the reliability of differences was 1.

Table 3. Changes in observed variance and reliability of difference scores associated with changes in reliability of component scores.

		$\rho(T_X,T_Y) = -.60$			$\rho(T_X,T_Y) = 0$			$\rho(T_X,T_Y) = .60$		
	σ_T^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2
$\sigma_E^2 = 1$	0.0	.000	.000	2.000	.000	.000	2.000	.000	.000	2.000
	0.2	.167	.242	2.640	.167	.167	2.400	.167	.074	2.160
	0.4	.286	.390	3.280	.286	.286	2.800	.286	.138	2.320
	0.6	.375	.490	3.920	.375	.375	3.200	.375	.194	2.480
	0.8	.444	.561	4.560	.444	.444	3.600	.444	.242	2.640
	1.0	.500	.615	5.200	.500	.500	4.000	.500	.286	2.800
	1.2	.545	.658	5.840	.545	.545	4.400	.545	.324	2.960
	1.4	.583	.691	6.480	.583	.583	4.800	.583	.359	3.120
	1.6	.615	.719	7.120	.615	.615	5.200	.615	.390	3.280
	1.8	.643	.742	7.760	.643	.643	5.600	.643	.419	3.440

		$\rho(T_X,T_Y) = -.60$			$\rho(T_X,T_Y) = 0$			$\rho(T_X,T_Y) = .60$		
	σ_E^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2
$\sigma_T^2 = 1$	0.0	1.000	1.000	3.200	1.000	1.000	2.000	1.000	1.000	0.800
	0.2	.833	.889	3.600	.833	.833	2.400	.833	.667	1.200
	0.4	.714	.800	4.000	.714	.714	2.800	.714	.500	1.600
	0.6	.625	.727	4.400	.625	.625	3.200	.625	.400	2.000
	0.8	.556	.667	4.800	.556	.556	3.600	.556	.333	2.400
	1.0	.500	.615	5.200	.500	.500	4.000	.500	.286	2.800
	1.2	.455	.571	5.600	.455	.455	4.400	.455	.250	3.200
	1.4	.417	.533	6.000	.417	.417	4.800	.417	.222	3.600
	1.6	.385	.500	6.400	.385	.385	5.200	.385	.200	4.000
	1.8	.357	.471	6.800	.357	.357	5.600	.357	.182	4.400

Consider now the relation between increases in reliability and power, reading from top to bottom in the columns in the upper section of the table and from bottom to top in the lower section. When the reliability coefficients of the component tests increased, the reliability of differences also increased, as long as just one column is considered. However, note that the same reliability of the components in many cases is associated with decidedly unlike reliabilities of the differences, depending on whether the change is attributable to a change in true-score variance or error-score variance. Often the values were far apart. Furthermore, the reliability of differences is either greater or less than that of the components, depending on whether the correlation between true scores, $\rho(T_X,T_Y)$, is positive or negative. As the absolute value of that correlation increases, the discrepancy is greater.

The observed scores of the differences, and hence the statistical power, *increases* as reliability increases if the change is attributable to a change in error-score variance and *decreases* if it is attributable to a change in true-score variance. That means that simply selecting a value of reliability, either of differences or the component tests, does not in itself provide information about the statistical power of the differences as a dependent variable. Just as in the case of a single test, the relation between reliability and power is not a *functional* relation unless the variance of one of the components of the scores is held constant.

These conclusions about the relation between power and the reliability of differences are consistent with results obtained by May & Hittner (2003), Overall & Woodward (1975, 1976), and Nicewander & Price (1978, 1983) using different methods. The so-called paradox of low reliability being associated with high power becomes more understandable from inspection of Table 3. That problem also is closely related to another issue that has been extensively treated in the literature, that of the reliability of differences often being considerably less than the reliability of the components. As the table shows, that is not always true, and again, looking at the reliability of the components alone, without further information, is one source of the trouble. The approach in Table 3, in which reliability coefficients are first related to the variances of true scores and error scores, makes it possible to focus on values that realistically would be likely to occur. At any rate, it is clear that an hypothesis test of differences can be powerful even if the reliability of a dependent variable is quite low.

How to increase statistical power: some practical implications

As mentioned before, a possible reason for the controversies surrounding the relation of reliability and statistical power is ambiguity about the precise meaning of the term “reliability” in practical research. The term often is used in a way that conforms to popular usage, and even to widespread usage in various scientific fields, but does not match the mathematical definition given in classical test theory. The root of the difficulty is the fact that reliability, as defined in test theory, is a property of populations of individuals, that is a ratio of statistics applicable to populations, but not to a single individual or experimental object. The “reliability” of a scientific instrument, especially in physical sciences, often refers to its consistency in measuring a single physical object of a certain kind, but that is not the way the term is used in classical test theory.

When one asks the question “How does reliability influence power?” investigators in psychology and education often assume the question is similar to “How does reliability influence validity?” or “How does test length influence reliability?” What is typically desired is a function relating changes in the first variable to changes in the second variable, and many such functions are known in test theory. On the other hand, a researcher in another field, or a statistician, may assume the question is similar to “How does sample size influence power?” or “How does the significance level influence power?” having in mind well-known functions relating those variables.

As emphasized in the present note, there is not a unique way of making the increments in reliability needed to exhibit power as a function of reliability. We can conclude that increasing an instrument’s reliability will contribute to greater power in hypothesis testing only if the change occurs through a reduction of error-score variance that exceeds any increase in true-score variance occurring at the same time.

Suppose a researcher has a choice between two instruments, one with a known reliability coefficient of .90 and the other .80. Before assuming automatically that the first instrument is the better choice, it is prudent to look at the variance of scores that can be expected. If the instrument with lower reliability typically produces scores with considerably less variability, it could still be the better choice. That is especially true if the experiment is designed to detect possible differences among large groups of subjects with respect to an independent variable and is not concerned with short-term fluctuations in measures of individuals.

Another way to look at the problem is to recall that an hypothesis test is essentially a determination, based on probability, of whether or not a difference found between samples can be attributed to chance variability. However, an hypothesis test is blind to the partitioning of variability into contributions from separate components, such as “true scores” and “error scores.” A test statistic such as *t* typically is computed as a ratio of an obtained value to an estimate of variability based on a sampling distribution.

Recommending that the reliability coefficient be increased whenever possible is not always good advice in hypothesis testing, although the conventional emphasis on practical measures to reduce error variance still applies. All other things being equal, the more error of measurement can be avoided in an experiment, the better, and that task certainly should be considered along with other well-known methods of increasing power (see, for example, Wilcox, 2003) that are useful in research. But reducing error is productive, we have seen, only if the same practical steps also reduce observed-score variance. If a more heterogeneous group is tested at the same time error of measurement is less, power does not necessarily increase. For practical usefulness, eliminating error and thereby increasing reliability for a particular population of examinees can be effective, provided the change is made without altering the population.

References

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, *22*(1), 49-55. doi:10.1111/j.2044-8317.1969.tb00419.x

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

Collins, L. M. (1996). Is reliability obsolete? A commentary on “Are simple gain scores obsolete?” *Applied Psychological Measurement, 20*(3), 289-292. doi:10.1177/014662169602000308

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Fleiss, J. J. (1976). Comment on Overall & Woodward’s asserted paradox concerning the measurement of change. *Psychological Bulletin*, *83*(5), 774-775. doi:10.1037/0033-2909.83.5.774

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

Hopkins, K. D., & Hopkins, D. R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education, 13*(4), 463-466. doi:10.1177/002246697901300413

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement, 51*(3), 585-595. doi:10.1177/0013164491513006

Levin, J. R. (1986). Note on the relation between the power of a significance test and the reliability of the measuring instrument. *Multivariate Behavioral Research*, *21*(2), 255-261. doi:10.1207/s15327906mbr2102_6

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

May, K., & Hittner, J. B. (2003). On the relation between power and reliability of difference scores. *Perceptual and Motor Skills, 97*(3.1), 905-908. doi:10.2466/PMS.97.7.905-908

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293-299. doi:10.1037/1082-989X.1.3.293

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement, 23*(1), 87-89. doi:10.1177/01466216990231007

Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of statistical tests. *Psychological Bulletin, 85*(2), 405-409. doi:10.1037/0033-2909.85.2.405

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests. *Psychological Bulletin, 94*(3), 524-513. doi:10.1037/0033-2909.94.3.524

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(3), 1-18. doi:10.1016/0022-2496(66)90002-2

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*(1), 85-86. doi:10.1037/h0076158

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin, 83*(5), 776-777. doi:10.1037/0033-2909.83.5.776

Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a significance test. *Journal of Educational Measurement, 14*(1), 47-52. doi:10.1111/j.1745-3984.1977.tb00028.x

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 23*(1), 9-17. doi:10.1007/BF02288974

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37-43. doi: 10.1177/0013164411409929

Wilcox, R. B. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

Wolfram, S. (1999). *The Mathematica Book* (4th ed.). NY: Wolfram Media/Cambridge University Press.

Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin, 100*(1), 123-124. doi:10.1037/0033-2909.100.1.123

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), 1-10. doi:10.1177/014662169301700101

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.). *Advances in Social Science Methodology*, *5*, (pp. 269-304). Greenwich, CT: JAI Press.