Modifying Spearman's Attenuation Equation to Yield Partial Corrections for Measurement Error—With Application to Sample Size Calculations

Educational and Psychological

Measurement

2018, Vol. 78(1) 70–79

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417713571

journals.sagepub.com/home/epm

W. Alan Nicewander¹

Abstract

Spearman's correction for attenuation (measurement error) corrects a correlation coefficient for measurement errors in either-or-both of two variables, and follows from the assumptions of classical test theory. Spearman's equation removes all measurement error from a correlation coefficient which translates into "increasing the reliability of either-or-both of two variables to 1.0." In this inquiry, Spearman's correction is modified to allow partial removal of measurement error from either-or-both of two variables being correlated. The practical utility of this partial correction is demonstrated in its use to explore increasing the power of statistical tests by increasing sample size versus increasing the reliability of the dependent variable for an experiment. Other applied uses are mentioned.

Keywords

reliability, Spearman correction, classical test theory, statistical power

Introduction

One of psychology's earliest contributions to statistics is the Spearman (1904) correction for attenuation, which corrects a correlation coefficient for measurement errors

¹Pacific Metrics, Inc., Monterey, CA, USA

Corresponding Author:

W. Alan Nicewander, Pacific Metrics, Inc., 1 Lower Ragsdale Dr, Ste 1150, Monterey, CA 93940, USA.

Email: anicewander@pacificmetrics.com

in either-or-both of the variables. This correction of correlations for measurement error follows from the assumptions of classical test theory,

$$\rho(\tau_x, \tau_y) = \frac{\rho(x, y)}{\sqrt{\rho(x, x')}\sqrt{\rho(y, y')}} \quad \text{or} \quad \rho^2(\tau_x, \tau_y) = \frac{\rho^2(x, y)}{\rho(x, x')\rho(y, y')}, \quad (1)$$

where τ_x and τ_y represent true values of X and Y , free of measurement errors, and ε_x and ε_y are measurement errors assumed to be uncorrelated with τ_x or τ_y and other observed variables (see Lord & Novick, 1968). The quantities,

$$\rho(x, x') = \sigma^2(\tau_x) / [\sigma^2(\tau_x) + \sigma^2(\varepsilon_x)] = \sigma^2(\tau_x) / \sigma^2(x) \quad (2a)$$

and

$$\rho(y, y') = \sigma^2(\tau_y) / [\sigma^2(\tau_y) + \sigma^2(\varepsilon_y)] = \sigma^2(\tau_y) / \sigma^2(y) \quad (2b)$$

are the reliability coefficients for X and Y .

Derivation of a Partial Correction for Measurement Error

Equation (1) answers the question, “If X and Y had perfect reliabilities equal to 1.00, what would the correlation between X and Y be?” However, suppose one has two measures X and Y that have unsatisfactory reliabilities, then a relevant question might be, “If X could be improved to have reliability equal to .90, and Y improved to have reliability equal to .80, what would the correlation between X and Y be?” Questions about the degree to which correlations can be improved by increasing the reliabilities of the X and Y variables to higher than current values (but less than 1.00) can be answered using the modified spearman correction derived below. This partial correction for measurement error is quite simple, and it is somewhat surprising that it has not been derived in the past; however, the author’s search of the literature did not discover any previous derivations of this equation.

Let X_c and Y_c be unobservable variables for which

$$\begin{aligned} \rho(x, x') &\leq \rho(x_c, x'_c) \leq 1, \\ \rho(y, y') &\leq \rho(y_c, y'_c) \leq 1, \end{aligned} \quad (3)$$

and

$$\rho(\tau_{x_c}, \tau_x) = \rho(\tau_{y_c}, \tau_y) = 1. \quad (4)$$

In other words, X_c and Y_c are variables having higher reliabilities than the original variables, X and Y , and the true scores for X_c and X correlate perfectly as do the true scores for Y and Y_c . Now, imbedding $\rho^2(x_c, y_c)$ in Spearman’s equation in order to obtain a solution for this correlation,

$$\rho^2(\tau_{x_c}, \tau_{y_c}) = \frac{\rho^2(x_c, y_c)}{\rho(x_c, x'_c)\rho(y_c, y'_c)}. \quad (5)$$

In (5), we seek to solve for $\rho^2(x_c, y_c)$; however, this appears to be unsolvable because it contains two unknowns, $\rho^2(x_c, y_c)$ and $\rho^2(\tau_{x_c}, \tau_{y_c})$. Using assumption in (4), we can substitute $\rho^2(\tau_{x_c}, \tau_{y_c}) = \rho^2(\tau_x, \tau_y)$ in (5) and solve for $\rho^2(x_c, y_c)$, viz.,

$$\begin{aligned} \rho^2(\tau_{x_c}, \tau_{y_c}) &= \rho^2(\tau_x, \tau_y) = \frac{\rho^2(x_c, y_c)}{\rho(x_c, x'_c)\rho(y_c, y'_c)}, \text{ or,} \\ \rho^2(x_c, y_c) &= \rho^2(\tau_x, \tau_y)[\rho(x_c, x'_c)\rho(y_c, y'_c)] \\ &= \frac{\rho^2(x, y)}{\rho(x, x')\rho(y, y')} [\rho(x_c, x'_c)\rho(y_c, y'_c)]. \end{aligned} \quad (6)$$

For example, assume that variables X and Y have $\rho^2(x, y) = .10$, and their reliabilities are .60 and .70, respectively: (a) if all measurement errors were removed from X and Y , the squared correlation would be $\rho^2(\tau_x, \tau_y) = .10/ (.6)(.7) = .238$; (b) if all measurement errors in X alone were eliminated, the correlation between τ_x and Y would be $\rho^2(\tau_x, y) = .10/ (.6) = .167$. In practice, the above values of correlation can never be attained because X and Y cannot be modified to give perfect reliabilities, but what would the squared correlation be if the reliabilities of X and Y were increased to .80 and .90, respectively? Using (6), the modified Spearman correction is

$$\begin{aligned} \rho^2(x_c, y_c) &= \rho^2(\tau_x, \tau_y)[\rho(x_c, x'_c)\rho(y_c, y'_c)] \\ &= .238(.8)(.9) = .173. \end{aligned}$$

Thus, a squared correlation of .10 becomes .238 under a complete correction for unreliability in both variables, and .173 under partial correction.

An Application of the Partial Spearman Correction: Increasing the Power of Statistical Tests for Linear Statistical Models

It is now demonstrated how the modified Spearman attenuation correction may be used in determining the decrease in sample size produced by increasing the reliability of the dependent variable (DV) measure. It has long been known that measurement error in an experiment's DV decreases the power of statistical tests—or conversely, that increasing the reliability of a DV can increase power and reduce the sample size required to achieve a specified degree of statistical power. In one of the earliest studies of measurement error, Sutcliffe (1958) showed that measurement error in a DV does not bias estimates of effects but does decrease the power of statistical tests. Cleary and Linn (1969) and Marcoulides (1993) discussed the role of DV reliability in maximizing power subject to constraints on sample size. Maxwell's (1980) research demonstrated that increasing reliability increases the magnitude of the

noncentrality parameter for an F -test—and thereby increases statistical power. Maxwell recommended that (in cases where the DV is an educational or psychological test) increasing the length of a measure will increase DV reliability and statistical power. The approach taken here is apparently unique in that it blends classical test theory (i.e., Spearman's attenuation equation) with linear statistical models in order to examine reliability–power issues. Specifically, F -ratios are expressed as functions of squared, multiple correlations; these multiple correlations are then subjected to total-or-partial correction for DV unreliability using Spearman's equation, and the modification thereof, proposed here.

In what follows, examples are used to illustrate how increasing DV reliability can reduce the sample size required for attaining specified degrees of statistical power of an F -test. Examples are presented in the context of one-way ANOVA models. Suppose an F -ratio is being used in an ANOVA with J -treatment conditions, and n observations per treatment (for a total sample size of $N = Jn$): The simple ANOVA model fitted to the data may be written as

$$Y = \beta_0 + \sum_{j=1}^J \beta_j X_j + \epsilon_1, \quad (7)$$

where, in this context, the X s are (1, 0) marker variables for presence or absence of treatments, and the β_j s are treatment effects. The null hypothesis for testing that the treatment effects are all zero is

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_J = 0. \quad (8)$$

The observed F -ratio for a test of this hypothesis is

$$F = \frac{R_{xy}^2/(J-1)}{(1-R_{xy}^2)/(N-J)} = \frac{R_{xy}^2}{(1-R_{xy}^2)} \left(\frac{N-J}{J-1} \right), \quad (9)$$

where R_{xy}^2 is the squared, multiple correlation (for several X -variables and a single Y) that is associated with the model above; the numerator and denominator degrees of freedom are $df_{num} = J-1$ and $df_{den} = N-J$. In order to correct the F -ratio for measurement error in Y , all that needs to be done is to correct R_{xy}^2 for measurement error in Y using Spearman's modified correction in (1), viz.,

$$F_{adjusted} = \frac{R_{xy}^2/\rho_y}{1-(R_{xy}^2/\rho_y)} \left(\frac{N-J}{J-1} \right) = \frac{R_{xy}^2}{(\rho_y - R_{xy}^2)} \left(\frac{N-J}{J-1} \right), \quad (10)$$

where ρ_y is a reliability coefficient associated with the variable, Y :

$$\rho_y = \begin{cases} 1, & \text{if no correction for the reliability of } Y \text{ is wanted;} \\ \rho(y, y'), & \text{if a complete Spearman correction is wanted;} \\ \rho(y_c, y'_c), & \text{if a partial Spearman adjustment is wanted.} \end{cases}$$

(Note: Nicewander (1975)—see also Guttman (1945)—have shown that $\rho(y, y') \geq R_{y,x}^2$, and hence, in theory, the F -ratio in (10) can never be negative.) The partial Spearman correction can be used in statistical power calculations wherein the effects of DV reliability can be evaluated in terms of the sample size needed to achieve a certain degree of power. In order to perform power calculations for varying values of sample size, n (per treatment group) and reliability adjustments, one needs the following:

1. An effect size—Cohen's $f^2 = R_{xy}^2 / (1 - R_{xy}^2)$ —was used here (Cohen, 1988). Specify a value of DV reliability— $\rho(y, y')$, $\rho(y_c, y'_c)$ or 1.0—and values of J and $N = Jn$ are chosen.
2. An F -ratio—adjusted for some specified value of DV reliability—is then computed using the values above,

$$F_{adj.} = \frac{R^2}{[\rho_y - R^2]} \left(\frac{N - J}{J - 1} \right) \quad (11)$$

3. A critical value of F —for testing the null hypothesis at some level, α , of Type I error—is determined:

$$F_{critical} = F_{\alpha, df_{num}, df_{den}} \quad (12)$$

4. An estimate of the noncentrality parameter, λ —adjusted for the specified reliability, ρ_y —is computed in order to specify the appropriate noncentral F -distribution for computing power,

$$\lambda_{adj.} \simeq (J - 1)F_{adj.} \quad (13)$$

(See Liu & Raudenbush, 2004.).

5. Finally, a value of power is computed, viz.,

$$\text{Power} = 1 - \int_0^{F_{\alpha, df_{num}, df_{den}}} u \varphi(u | df_{num}, df_{den}, \lambda_{adj}) du, \quad (14)$$

where

$$\varphi(u | df_1, df_2, \lambda_{adj})$$

is a noncentral p df for F .

In the examples below, the sample size, n , for producing power = .80 was calculated using a do-loop (*do while (power <= .80)*) in SAS® that encompassed Steps 1

through 5, using the SAS routines, FINV (to compute the critical values of F), and PROBF (to compute power).

Examples of Using the Complete and Partial Spearman Adjustments in Power Calculations

In the following tables, there are either two or four treatment conditions. The “original” DV, Y , has reliability equal to .60, and the effect, on power, of partially adjusting reliability to .70, .75, .80, .85, and .90 were examined—as well as the case where there was complete (Spearman) adjustment of reliability to a value of 1.00, or no correction whatsoever. The value of the original reliability specified to be .60 is based principally on the fact that in psychometrics, .60 is generally considered to be low, and one might consider revamping the measure in order to decrease measurement error. Often, .70 is considered as minimally acceptable. The value of Cohen’s effect size used here was .11 (corresponding to an R^2 of .10), and by convention, .11 would be considered to be a small effect (Cohen, 1988). The sample size (per treatment condition) needed to achieve power equal to .80, $n(.8)$ at an α level of .05 is noted. Also, as the reliability of Y is increased from .60 to .70, .75, .80, .85, .90, and 1.00, the percentage reduction in $n(.8)$ is noted, which indicates increased power as measured by savings in sample size.

Results

Two-Treatment Experiment

For this case, the effect size was small ($f^2 = .11$; $R^2 = .10$), and the reliability of the DV was .6. The first section of Table 1 summarizes the results for the analysis of the two-treatment design. The sample size, $n(.8)$, was equal to 38 per treatment condition with no reliability adjustment, and a perfectly reliable Y required far fewer observations, $n(.8) = 22/\text{treatment}$ —a decrease in sample size of 42% as indicated. If the reliabilities of Y were .9, .85, .8, .75, or .7, the sample size calculations yielded $n(.8)$ values of 24, 28, and 32, respectively, with corresponding reductions in $n(.8)$ of 37%, 32%, 26%, 21%, and 16% compared with no reliability correction in Y .

Four-Treatment Experiment

In the second section of Table 1, the results for the statistical analysis of this experimental design were very similar to those from the two-treatment results. With reliability = .60, $n(.8)$ was equal to 27/treatment with no adjustment to reliability, and equal to 16/treatment if Y had no measurement error (perfect reliability)—a percentage reduction in $n(.8)$ of 41%. Reliabilities for Y of .9, .85, .8, .75, and .7 yielded $n(.8)$ values of 18, 19, 20, 22, and 23, respectively, with corresponding reductions in $n(.8)$ of 33%, 30%, 26%, 18%, and 15% compared with no reliability correction in Y .

Table 1. Sample Size, per Group, $n(.8)$, Necessary to Attain Power = .80 for Various Levels of Reliability Correction for a Small Effect Size, $f^2 = .11$

Small effect ($f^2 = .11$); $R^2 = .10$; $\rho(y, y') = .60$			
Number of groups (J)	Reliability correction in Y	$n(.8)$, $\alpha = .05$	Reduction in $n(.8)$
2	No correction	38	—
2	Complete correction	22	42%
2	$\rho(y_c, y'_c) = .90$	24	37%
2	$\rho(y_c, y'_c) = .85$	26	32%
2	$\rho(y_c, y'_c) = .80$	28	26%
2	$\rho(y_c, y'_c) = .75$	30	21%
2	$\rho(y_c, y'_c) = .70$	32	16%
4	No correction	27	—
4	Complete correction	16	41%
4	$\rho(y_c, y'_c) = .90$	18	33%
4	$\rho(y_c, y'_c) = .85$	19	30%
4	$\rho(y_c, y'_c) = .80$	20	26%
4	$\rho(y_c, y'_c) = .75$	22	18%
4	$\rho(y_c, y'_c) = .70$	23	15%

The results on sample size shown in Table 2 are commensurate with those given in Table 1: The percentage decreases in $n(.8)$, as effect size increases in Table 2, are virtually identical to those in Table 1 for the same degree of increase in the DV reliability. However, even though the percentage decreases in $n(.8)$ are nearly the same as those in Table 1, the absolute decreases in $n(.8)$ are considerably smaller, and become trivially small as effect size becomes large—as can be seen in Table 3 where the results for the largest effect size (.33) are summarized. In cases of large effect size, increasing the reliability of the DV will only be economically viable in cases where the subjects needed for an experiment are extremely rare, and increasing sample size is much more difficult and expensive than increasing the reliability of the DV.

Summary and Conclusions

Spearman's correction for attenuation has proven useful over the years primarily because it allows researchers to determine what the linear relationship would be between two variables, X and Y , if all measurement errors were removed from one or both of the variables. The correction is especially useful for looking at the strength of theoretical relationships between variables undistorted by measurement errors; for example, "Does the ACT college admission tests measure the same underlying construct as the Wechsler Adult Intelligence Scale?" The Spearman equation, modified in the present inquiry, allows the estimation of the effect, on correlations, of partial corrections for unreliability. These partial corrections are probably more useful in

Table 2. Sample Size, per Group, $n(.8)$, Necessary to Attain Power = .80 for Various Levels of Reliability Correction for a Medium Effect Size, $f^2 = .18$.

Medium effect ($f^2 = .18$); $R^2 = .15$; $\rho(y, y') = .60$			
Number of groups (J)	Reliability correction	$n(.8), \alpha = .05$	Reduction in $n(.8)$
2	No correction	24	—
2	Complete correction	14	42%
2	$\rho(y_c, y'_c) = .90$	16	33%
2	$\rho(y_c, y'_c) = .85$	17	29%
2	$\rho(y_c, y'_c) = .80$	18	25%
2	$\rho(y_c, y'_c) = .75$	19	21%
2	$\rho(y_c, y'_c) = .70$	20	17%
4	No correction	18	—
4	Complete correction	10	44%
4	$\rho(y_c, y'_c) = .90$	12	33%
4	$\rho(y_c, y'_c) = .85$	13	28%
4	$\rho(y_c, y'_c) = .80$	13	28%
4	$\rho(y_c, y'_c) = .75$	14	22%
4	$\rho(y_c, y'_c) = .70$	15	17%

Table 3. Sample Size, per Group, $n(.8)$, Necessary to Attain Power = .80 for Various Levels of Reliability Correction for a Large Effect Size, $f^2 = .33$.

Large effect ($f^2 = .33$); $R^2 = .25$; $\rho(y) = .60$			
Number of groups (J)	Reliability correction	$n(.8), \alpha = .05$	Reduction in $n(.8)$
2	No correction	14	—
2	Complete correction	8	43%
2	$\rho(y_c, y'_c) = .90$	9	36%
2	$\rho(y_c, y'_c) = .85$	10	29%
2	$\rho(y_c, y'_c) = .80$	10	29%
2	$\rho(y_c, y'_c) = .75$	11	21%
2	$\rho(y_c, y'_c) = .70$	12	14%
4	No correction	10	—
4	Complete correction	6	40%
4	$\rho(y_c, y'_c) = .90$	7	30%
4	$\rho(y_c, y'_c) = .85$	7	30%
4	$\rho(y_c, y'_c) = .80$	8	20%
4	$\rho(y_c, y'_c) = .75$	8	20%
4	$\rho(y_c, y'_c) = .70$	9	10%

applied situations than in the examination of theoretical relationships. For example, the partial correction allows one to deal with questions such as, “If we can improve the reliability of these measures a bit, what would their correlation look like?” For

example, the correction could be useful for dealing with questions such as, “If we could increase the reliability of course grades to a level approaching that of the ACT, what would the validity coefficient (correlation between ACT scores and grades) be for the ACT?” Another example of the applied usage of the partial correction for attenuation is, of course, the one used here in which it was used to examine the degree to which the sample size in an experiment can be reduced by increasing the reliability of the DV.

A final note on the role of DV reliability in statistical tests: Increasing the reliability coefficient of a DV in an experiment does not necessarily mean that the power of a statistical test will be increased. For given treatment effects in an experiment, the power of a statistical test is determined by magnitude of the “within treatment” variance, and this variance is composed of the sum of true-and-error variances. Spearman’s correction for attenuation—and the partial correction derived here—can increase power by decreasing the magnitude of the measurement error variance (i.e., increasing the reliability coefficient). However, when experimental “controls for individual differences”—such as the analysis of difference scores in a two-treatment repeated measures experiment, blocking in ANOVA, and so on are used—“controlling individual differences” translates into “reducing true variance.” Therefore, there are situations where increased statistical power is accompanied by a decrease in the reliability of the DV (Nicewander & Price, 1983).

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

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, 49-52.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.
- Liu, X., & Raudenbush, S. (2004). A note on the noncentrality parameter and effect size estimates for the F test in ANOVA. *Journal of Educational and Behavioral Statistics*, 29, 251-255.
- Lord, F. M. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational and Behavioral Statistics*, 18, 197-206.

- Maxwell, S. F. (1980). Dependent variable reliability and determination of sample size. *Applied Psychological Measurement*, 4, 253-259.
- Nicewander, W. A. (1975). A relationship between Harris factors and Guttman's sixth lower bound to reliability. *Psychometrika*, 40, 197-203.
- Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin*, 94, 524-533.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.
- Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 23, 9-17.