A Note on Correlations Corrected for Unreliability and Range Restriction

Nambury S. Raju, Daniel V. Lezotte and Benjamin K. Fearing, Illinois Institute of Technology
T. C. Oshima, Georgia State University

This note describes a procedure for estimating the range restriction component used in correcting correlations for unreliability and range restriction when an estimate of the reliability of a predictor is not readily available for the unrestricted sample. This procedure is illustrated with a few examples.
Index terms: unreliability, range restriction, corrected correlations

In a recent article, Raju and Brand (2003) proposed an asymptotic formula for computing the standard error of a correlation coefficient corrected for unreliability in x (predictor) and y (criterion) and direct range restriction in x. One of the statistics needed for computing the corrected correlation and its standard error is the ratio (k) of the standard deviation of true scores on x in the unrestricted sample (S_{t_x}) to the standard deviation of true scores on x in the restricted sample (s_{t_x}). Although s_{t_x} is generally known (or can be computed readily when the standard deviation of observed scores [s_x] and their reliability [r_{xx}] are known in the restricted sample), S_{t_x} may be difficult to estimate in practice because the reliability of x in the unrestricted sample (R_{xx}) may be unavailable. The purpose of this note is to describe a procedure for estimating R_{xx}, and hence S_{t_x} and k, for use in computing corrected correlations and their standard errors.

Corrected Correlation

Using Raju and Brand's (2003) notation, the corrected correlation ($\hat{\rho}_{xy}$) may be expressed as

$$\hat{\rho}_{xy} = \frac{kr_{xy}}{\sqrt{r_{xx}r_{yy} - r_{xy}^2 + k^2r_{xy}^2}}, \tag{1}$$

where r_{xy} is the restricted and attenuated correlation between the x and y in a restricted sample, and r_{xx} and r_{yy} are the sample-based restricted predictor and criterion reliabilities, respectively. As previously noted, k is the ratio of the unattenuated, unrestricted standard deviation to the unattenuated, restricted standard deviation of x; k can also be stated as the ratio of the unrestricted true score standard deviation (S_{t_x}) to the restricted true score standard deviation (s_{t_x}) on x. That is,

$$k = \frac{S_{t_x}}{s_{t_x}}, \tag{2}$$

Applied Psychological Measurement, Vol. 30 No. 2, March 2006, 145–149
DOI: 10.1177/0146621605283342

which can also be expressed as

$$k = \frac{S_{t_x}}{s_{t_x}} = \frac{(S_x)(\sqrt{R_{xx}})}{(s_x)(\sqrt{r_{xx}})} = k^* \sqrt{\frac{R_{xx}}{r_{xx}}}, \tag{3}$$

where

$$k^* = \frac{S_x}{s_x}, \tag{4}$$

which is the ratio of the standard deviation of observed scores on x in the unrestricted sample to the standard deviation of observed scores in the restricted sample. To compute k, according to equation (3), one needs the standard deviations of observed scores and their reliabilities from the unrestricted and restricted samples because, according to classical test theory, the true score standard deviation is the product of the standard deviation of observed scores and the square root of their reliability. In practice, s_x, r_{xx}, and S_x are much more readily available than R_{xx}. In such situations, it is still possible to estimate R_{xx} and hence be able to compute k in equation (3) and the corrected correlation in equation (1).

Estimating Reliability in the Unrestricted Sample

If sufficient predictor data (at the item level) are available for the unrestricted sample, it is generally possible to obtain an estimate of the unrestricted reliability (R_{xx}). In the absence of such data, it is still possible to obtain an estimate of R_{xx}. According to Lord and Novick (1968, p. 30), the unrestricted reliability may be expressed as a function of k^* and r_{xx}. That is,

$$R_{xx} = 1 - \left(\frac{1}{k^*}\right)^2 (1 - r_{xx}). \tag{5}$$

Substituting equation (5) into equation (3), one obtains

$$k = \sqrt{\frac{(k^*)^2 + r_{xx} - 1}{r_{xx}}}. \tag{6}$$

Compared to equation (3), equation (6) does not involve R_{xx} and hence can be used to compute k when only the standard deviations of observed scores in the restricted and unrestricted samples and the reliability in the restricted sample are known.

In practice, an investigator may use either equation (3) or equation (6) for the range restriction component (k) in equation (1) for a corrected correlation: equation (3) if the unrestricted and restricted reliabilities of x are known or equation (6) if only the restricted reliability of x is known. In either case, it is assumed that the restricted and unrestricted standard deviations of observed scores (k^*) are known.

The distinction between k and k^* is important; k^* represents the ratio of observed score standard deviations, whereas k denotes the ratio of true score standard deviations. Because k^* is much more readily available than k in practice, one may be tempted to use it to estimate the corrected correlation as well as its standard error. Unfortunately, the use of k^* in place of k in equation (1) can lead to inaccurate estimates of corrected correlations. Corrected correlations based on k^* will always be less than or equal to the corrected correlations based on k. To show this algebraically, equation (1) (dividing both the numerator and denominator by k) should be rewritten as

$$\hat{\rho}_{xy} = \frac{r_{xy}}{\sqrt{\frac{(r_{xx}r_{yy} - r_{xy}^2)}{k^2} + r_{xy}^2}}. \tag{7}$$

According to equation (3), k^* will be less than k whenever R_{xx} is greater than r_{xx}. Hence, the denominator in equation (7) (or equation (1)) will be bigger with k^* than with k, thus leading to an underestimate of a corrected correlation with k^*.

Standard Error of a Corrected Correlation

According to Raju and Brand (2003), an asymptotic standard error (SE) of a corrected correlation may be expressed as

$$SE(\hat{\rho}_{xy}) = \frac{k\sqrt{r_{xx}r_{yy}(r_{xx} - r_{xy}^2)(r_{yy} - r_{xy}^2)}}{\sqrt{(n-1)}[r_{xx}r_{yy} - r_{xy}^2 + k^2 r_{xy}^2]^{3/2}}, \tag{8}$$

where n is the number of subjects in the sample. The standard error formula also includes k, and hence the use of k^* in place of k in equation (8) will also affect an estimate of the standard error of a corrected correlation. The use of k^* in equation (8) can result in either an overestimate or an underestimate of SE compared with the SE estimate with k. In general, the under- and overestimations will be small.

Examples

In an attempt to offer some guidance to practitioners on the degree of over- and underestimation with the use of k^* in place of k in equations (1) and (8), several examples are presented in Table 1. In preparing this table, three values of r_{xy} (.600, .400, and .200), three values of k^* (1.250, 1.500, and 2.000), two values of r_{xx} (.900 and .700), and two values of r_{yy} (.700 and .500) were used. A value from one set (or statistic) was crossed with all values from each of the other three sets, resulting in 36 different examples. In all these examples, the sample size (n) was set at 100. Estimates of $\hat{\rho}_{xy}$ and $SE(\hat{\rho}_{xy})$ with k and k^* for the 36 examples are shown in Table 1. As expected, estimates of $\hat{\rho}_{xy}$ with k^* are smaller than the estimates with k. Differences in the two estimates (estimate with k minus estimate with k^*) vary from a minimum of .000 (Examples 4, 8, and 12) to a maximum .054 (Example 36). In all cases, the two estimates of $\hat{\rho}_{xy}$ differ in the second decimal. In three cases, the $\hat{\rho}_{xy}$s are actually greater than 1.00 but rounded back to 1.00, as shown in Table 1. The differences in the estimates of $SE(\hat{\rho}_{xy})$ vary from –.011 (Examples 23 and 24) to .009 (Example 31). Most of the differences in the $SE(\hat{\rho}_{xy})$ estimates are negative, meaning that the use of k^* will generally overestimate the standard error of a corrected correlation. The number of instances of underestimation of $SE(\hat{\rho}_{xy})$ with k^* is small, all confined to the small observed correlation of 0.200. In fact, using equation (8), it can be shown that when $r_{xy} = 0$, an estimate of $SE(\hat{\rho}_{xy})$ with k^* (in view of equation (3)) will be less than or equal to an estimate of $SE(\hat{\rho}_{xy})$ with k.

In conclusion, the computation of corrected correlations and their standard errors in the Raju and Brand (2003) framework must be based on k; k may be obtained with equation (3) when the reliability of x in the unrestricted sample is known or with equation (6) when only the reliability of x in the restricted sample is known. It is incorrect to use k^* as the range restriction component in

Table 1
Examples of Corrected Correlations and Sampling Error Variances Errors Associated With k and k^*

Example	r_{xy}	k^*	k	r_{xx}	r_{yy}	$\hat{\rho}_{xy}$ (with k^*)	$\hat{\rho}_{xy}$ (with k)	$SE(\hat{\rho}_{xy})$ (with k^*)	$SE(\hat{\rho}_{xy})$ (with k)
1	0.600	1.250	1.275	0.900	0.700	0.822	0.827	0.056	0.055
2	0.600	1.250	1.275	0.900	0.500	0.928	0.931	0.044	0.043
3	0.600	1.250	1.343	0.700	0.700	0.901	0.913	0.052	0.047
4	0.600	1.250	1.343	0.700	0.500	1.000	1.000	0.039	0.034
5	0.600	1.499	1.545	0.900	0.700	0.866	0.872	0.046	0.044
6	0.600	1.499	1.545	0.900	0.500	0.949	0.951	0.033	0.031
7	0.600	1.499	1.668	0.700	0.700	0.928	0.941	0.039	0.033
8	0.600	1.499	1.668	0.700	0.500	1.000	1.000	0.027	0.022
9	0.600	2.000	2.082	0.900	0.700	0.918	0.923	0.031	0.029
10	0.600	2.000	2.082	0.900	0.500	0.970	0.972	0.020	0.018
11	0.600	2.000	2.299	0.700	0.700	0.958	0.967	0.024	0.019
12	0.600	2.000	2.299	0.700	0.500	1.000	1.000	0.015	0.011
13	0.400	1.250	1.275	0.900	0.700	0.589	0.597	0.103	0.103
14	0.400	1.250	1.275	0.900	0.500	0.680	0.688	0.107	0.106
15	0.400	1.250	1.343	0.700	0.700	0.657	0.683	0.108	0.105
16	0.400	1.250	1.343	0.700	0.500	0.754	0.777	0.109	0.103
17	0.400	1.499	1.545	0.900	0.700	0.658	0.670	0.100	0.099
18	0.400	1.499	1.545	0.900	0.500	0.744	0.754	0.097	0.095
19	0.400	1.499	1.668	0.700	0.700	0.722	0.758	0.099	0.093
20	0.400	1.499	1.668	0.700	0.500	0.809	0.837	0.094	0.084
21	0.400	2.000	2.082	0.900	0.700	0.759	0.772	0.086	0.084
22	0.400	2.000	2.082	0.900	0.500	0.830	0.840	0.075	0.072
23	0.400	2.000	2.299	0.700	0.700	0.812	0.848	0.080	0.069
24	0.400	2.000	2.299	0.700	0.500	0.878	0.904	0.067	0.056
25	0.200	1.250	1.275	0.900	0.700	0.309	0.315	0.143	0.145
26	0.200	1.250	1.275	0.900	0.500	0.364	0.370	0.163	0.165
27	0.200	1.250	1.343	0.700	0.700	0.349	0.372	0.158	0.165
28	0.200	1.250	1.343	0.700	0.500	0.410	0.434	0.180	0.186
29	0.200	1.499	1.545	0.900	0.700	0.364	0.373	0.161	0.164
30	0.200	1.499	1.545	0.900	0.500	0.424	0.435	0.180	0.182
31	0.200	1.499	1.668	0.700	0.700	0.408	0.445	0.175	0.184
32	0.200	1.499	1.668	0.700	0.500	0.474	0.514	0.194	0.200
33	0.200	2.000	2.082	0.900	0.700	0.462	0.477	0.185	0.188
34	0.200	2.000	2.082	0.900	0.500	0.530	0.545	0.197	0.198
35	0.200	2.000	2.299	0.700	0.700	0.512	0.565	0.195	0.198
36	0.200	2.000	2.299	0.700	0.500	0.583	0.637	0.203	0.200

Note. The sample size is 100 for all examples.

the Raju and Brand framework. The use of k^* (instead of k) will lead to a slight underestimation of corrected correlations and slight under- or overestimation of the standard errors of corrected correlations. Table 1 offers several examples of what to expect when k^* is used in place of k in estimating corrected correlations and their standard errors. The 36 examples in Table 1 do not cover a wide range of other possibilities, but these examples are realistic enough to offer practitioners some useful guidance on what to expect when k^* is used in place of k in estimating corrected correlations and their standard errors. A comparison of these results with the extensive Monte Carlo assessments reported in Raju, Burke, Normand, and Langlois (1991) for the corrected correlations with k (Tables 4-7) and in Raju and Brand (2003) for the sampling variances also with k (Table 3) may be helpful in understanding the accuracy of equations (1) and (8) when used with k^*. Despite the small differences in accuracy, it is recommended that practitioners use k whenever the true score standard deviation is available in the unrestricted sample or estimate k with equation (6) rather than using k^* in place of k.

References

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

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27,* 52-71.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76,* 432-446.

Acknowledgments

The authors express their appreciation to three anonymous reviewers and the editor for their helpful comments on an earlier version of this article.

Author's Address

Address correspondence to Daniel V. Lezotte, Institute of Psychology, Illinois Institute of Technology, Chicago, IL 60616-3793; e-mail: lezotte@iit.edu.