Correction for Attenuation of the Multiple Correlation Coefficient Given Non-Independent Error Scores

Brenna Curley **Debra Wetcher-Hendricks**
Moravian University

The multiple correlation coefficient measures the linear relationship between two or more independent variables and a single dependent variable. Like other correlations, the multiple correlation value relies upon observed scores. Researchers attempting to consider the relationship between true scores have relied upon standard corrections for attenuation. However, these standard estimates for the multiple correlation coefficient corrected for attenuation ignore any potentially correlated error in the observed data. In this work we propose a new estimator for the multiple correlation coefficient that accounts for the error in observed data where this error is allowed to covary. We explore properties of the estimator via simulation and illustrate its effectiveness with a real-world data application.

Formulas for partial and part correlation coefficients that correct for attenuation while accounting for correlated error scores were published by Wetcher-Hendricks (2006). However, existing formulas that estimate true-value multiple correlation coefficients still rely upon the erroneous psychometric assumption of independent error scores. A process similar to that used for estimates of the partial and part correlation coefficients can be used to derive a formula for the multiple correlation coefficient that corrects for attenuation and accounts for covarying error scores.

The multiple correlation coefficient, in and of itself, indicates the linear relationship between two or more independent variables and a single dependent variable. The multiple independent variables, collectively called the canonical variable, often combine to produce a stronger relationship with the dependent variable than any of the independent variables does alone. The formula to compute the standard multiple correlation consists of a systematic arrangement of pairwise correlation coefficients. The equation,

$$\rho_{x.yz} = \sqrt{\frac{\rho_{xy}^2 - 2\rho_{xy}\rho_{xz}\rho_{yz} + \rho_{xz}^2}{1-\rho_{yz}^2}}, \quad (1)$$

shows this arrangement for a situation involving two independent variables, Y and Z. With additional independent variables, both the numerator and denominator expand to include relevant terms. Bacon (1938), and B. R. B. (1924), among others, provide various approaches to making these expansions.

The multiple correlation coefficient corrected for attenuation described in this piece has foundations in two statistical principles. First, data gathered by researchers does not describe the truth of the situation investigated. Rather, an observed data point represents the true value along with an error value. Error consists of unpredictable and unmanageable factors that affect the data we observe. The observed data, therefore, does not necessarily represent the actual aptitude, conduct, sentiment, etc. we are trying to measure; but, instead is a combination of the true measurement along with the error that taints that measure. The basic psychometric principle that an observed score, X, equals the sum of its true score and its error score ($X = T_x + E_x$) reflects this relationship (Allen & Yen, 2001).

An important point to make with regard to error scores distinguishes between the psychometric value of E and the residual value (also sometimes called an error score). One can determine residual values by finding the difference between an observed Y value and its corresponding predicted value, for a given observed X value, based on the estimated regression model. The ability to obtain residual values allows for analyses of the error scores (estimated by the residuals) that address model diagnostics such as robustness of the error score variances and scedasticity of the error scores. Consequently, approaches to adjust for irregularities in residual values, most notably the Huber-White Sandwich Estimator which corrects for heteroscedasticity, exist (Dudgeon, 2017; Freedman, 2006; Szpiro, Rice, & Lumley, 2010). Many statistical software programs, in fact, make these functions available to users. However, users should not confuse the error addressed by these functions with the psychometric error created by the disparity between the observed and true score. Although statistical software programs exist for Huber-White Sandwich estimators in the presence of this psychometric error (see for example Hardin and Carroll (2003) and Hardin, Schmiediche, and Carroll (2003)), these models and estimates rely on additional instrumental variables or regression calibration methods and further focus on estimating the model standard errors whereas our work focuses more specifically on the multiple correlation coefficient. To date, no statistical software program can provide the desired estimates for the correlation in multiple linear regression that adequately accounts for psychometric error. In what follows, error scores will refer to the psychometric error not the residuals.

Accounting for the error score is necessary because the presence of psychometric error has an attenuating effect upon correlation coefficients (Carroll, Ruppert, & Stefanski, 1995). In other words, correlation coefficients calculated using observed scores tend to be smaller than values calculated from the true-score values, if these values could be obtained. Spearman (1904) made strides towards improving correlation coefficient estimates by defining an estimator for the true-value pairwise correlation coefficient. The coefficient corrected for attenuation, defined by Spearman's estimate, consistently lies closer to the actual true-score value than the observed-score value does (Hakstian, Schroeder, & Rogers, 1988). The derivation of the correction for attenuation, however, relies upon the psychometric assumption of independent error scores.

The assumption of independent error scores emerges as the second statistical principle to consider. Error scores, in reality, covary. Consider the example of SAT tests, as discussed by Wetcher-Hendricks (2006). Factors such as illness, nervousness, or distractions in the setting (e.g. a warm room, noise in the hallway) that affect participants' performance on the test contributes to an error score. But, these factors likely exist while the individuals complete both the verbal and the mathematics portions of the test, leading to a predictable relationship, or covariance, between the error scores on these two portions.

If error scores did not covary, then simply replacing the uncorrected pairwise values in the multiple correlation coefficient formula with values corrected using Spearman's formula would produce an accurate estimate of the multiple correlation between true scores. However, in reality, an accurate representation of the multiple correlation coefficient requires correcting the pairwise values for attenuation without making the assumption of independent error scores. Zimmerman and Williams (1977) developed a formula for pairwise correlations that corrects for attenuation in light of covarying error scores. In what follows, we derive an estimator for the multiple correlation coefficient that best estimates the true-score relationship by replacing the pairwise coefficients in the multiple correlation formula with the Zimmerman and Williams (1977) formula. We follow the derivation with an illustration of the effectiveness of the new estimator via both simulation and with a real-world data application.

Methods

Development of the new multiple correlation coefficient formula, which corrects for attenuation without making the assumption of independent error scores, begins with the standard multiple correlation coefficient shown in (1). In this equation ρ_{xy} , ρ_{xz} , and ρ_{yz} represent the observed pairwise correlations. An estimator that corrects the multiple correlation coefficient for attenuation, however, does not rely on these observed-score pairwise coefficients. Rather, these pairwise coefficients are adjusted to estimate the true score. The correction for attenuation that comes from Spearman (1904) estimates pairwise correlations between true scores, T_x and T_y by,

$$\rho_{T_xT_y}=\frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \tag{2}$$

with $\rho_{xx'}$ and $\rho_{yy'}$ representing the reliability values of X and Y, respectively. Still, simply replacing the pairwise observed coefficients with the versions corrected using Spearman's formula does not provide an accurate estimate of the true-score multiple correlation because it assumes independent error scores.

An estimator that corrects the multiple correlation coefficient for attenuation without assuming independent errors, relies upon the true-score values obtained with the Zimmerman and Williams (1977) formula,

$$\rho_{T_xT_y} = \frac{\rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \tag{3}$$

The importance of the Zimmerman and Williams (1977) formula, in fact, "does not lie in practicality, but theory, given that it allows researchers to understand the process needed to increase the accuracy of the correction for attenuation" (Wetcher-Hendricks, 2006). In a theoretical context, the error score correlation coefficient ($\rho_{E_xE_y}$) in the numerator can be easily computed and input into (3). Replacing the pairwise coefficients in (1) with the corresponding true-score values defined in (3) gives,

$$\rho_{T_x.T_yT_z} = \sqrt{\frac{\left(\frac{\rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}\right)^2 - 2\left(\frac{\rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}\right)\times\left(\frac{\rho_{xz} - \rho_{E_xE_z}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{zz'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{zz'}}}\right)\times\left(\frac{\rho_{yz} - \rho_{E_yE_z}\sqrt{1-\rho_{yy'}}\sqrt{1-\rho_{zz'}}}{\sqrt{\rho_{yy'}}\sqrt{\rho_{zz'}}}\right) + \left(\frac{\rho_{xz} - \rho_{E_xE_z}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{zz'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{zz'}}}\right)^2}{1-\left(\frac{\rho_{yz} - \rho_{E_yE_z}\sqrt{1-\rho_{yy'}}\sqrt{1-\rho_{zz'}}}{\sqrt{\rho_{yy'}}\sqrt{\rho_{zz'}}}\right)^2}}, \tag{4}$$

with the multiple correlation coefficient subscript updated to $\rho_{T_x.T_yT_z}$ to represent the fact that the components of the equation represent true, not observed, pairwise values. Simplification of the expression in (4) follows from first creating a common denominator, separately within the overall numerator and the denominator:

$$\rho_{T_x.T_yT_z} = \sqrt{\frac{\frac{\rho_{zz'}\left(\rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}\right)^2 - 2\left(\rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}\right)\times\left(\rho_{xz} - \rho_{E_xE_z}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{zz'}}\right)\times\left(\rho_{yz} - \rho_{E_yE_z}\sqrt{1-\rho_{yy'}}\sqrt{1-\rho_{zz'}}\right) + \rho_{yy'}\left(\rho_{xz} - \rho_{E_xE_z}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{zz'}}\right)^2}{\rho_{xx'}\rho_{yy'}\rho_{zz'}}}{\frac{\rho_{yy'}\rho_{zz'} - \left(\rho_{yz} - \rho_{E_yE_z}\sqrt{1-\rho_{yy'}}\sqrt{1-\rho_{zz'}}\right)^2}{\rho_{yy'}\rho_{zz'}}}}. \tag{5}$$

By combining and cancelling like terms in (5) we get the simplified formula,

$$\rho_{T_x.T_yT_z} = \sqrt{\frac{\rho_{zz'}\phi_{xy}^2+\rho_{yy'}\phi_{xz}^2-2\phi_{xy}\phi_{xz}\phi_{yz}}{\rho_{xx'}\left(\rho_{yy'}\rho_{zz'}-\phi_{yz}^2\right)}} \tag{6}$$

with, $$\phi_{xy} = \rho_{xy} - \rho_{E_xE_y}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{yy'}}\ , \tag{7}$$

$$\phi_{xz} = \rho_{xz} - \rho_{E_xE_z}\sqrt{1-\rho_{xx'}}\sqrt{1-\rho_{zz'}} \quad \text{, and} \tag{8}$$

$$\phi_{yz} = \rho_{yz} - \rho_{E_yE_z}\sqrt{1-\rho_{yy'}}\sqrt{1-\rho_{zz'}}\ . \tag{9}$$

The final estimator defined in (6)-(9) produces the new corrected multiple correlation coefficient. This estimator differs from the multiple correlation coefficient corrected for attenuation using Spearman's method (obtained by simply combining (1) and (2)) as it estimates the relationship between the true scores of the three variables, *X*, *Y* and *Z* while additionally accounting for error scores that covary.

Applications

In this section, we illustrate properties and the effectiveness of the new multiple correlation coefficient estimator, defined by the formula in (6)-(9), through both a simulation study and a real-world data application. The simulations illustrate properties (e.g., bias and variability) of the estimator assuming multivariate normal data. We follow the simulation study with a real-world data application to further illustrate the estimator's effectiveness.

Simulations

The simulated data is used to investigate the performance of the new estimator under scenarios with varying correlations of both the true scores and the error scores. We compare the performance of the estimator to the observed-score and Spearman formula estimators for the multiple correlation coefficient as defined in (1) and the combination of (1) and (2), respectively.

We follow the setup described in the Introduction, in which we assume data are observed with additive error. Simulations pertain to scenarios with both the true data and the errors assumed to be normally distributed, correlated random variables. Specifically, the simulation model represents a scenario with both the true scores $\boldsymbol{T} = (T_x, T_y, T_z)'$ and the errors $\boldsymbol{E} = (E_x, E_y, E_z)'$ following multivariate normal (MVN) distributions. That is, we let $\boldsymbol{E} \sim MVN(\boldsymbol{0}, \Sigma_{EE})$ in which,

$$\Sigma_{EE} = \begin{bmatrix} 1 & r_{e_{xy}} & r_{e_{xz}} \\ & 1 & r_{e_{yz}} \\ & & 1 \end{bmatrix}, \tag{10}$$

and $\boldsymbol{T} \sim MVN(\boldsymbol{0}, \Sigma_{TT})$ in which,

$$\Sigma_{TT} = \begin{bmatrix} 1 & r_{t_{xy}} & r_{t_{xz}} \\ & 1 & r_{t_{yz}} \\ & & 1 \end{bmatrix}, \tag{11}$$

for some chosen r_e and r_t values.

Comparisons between the observed-score estimator, the Spearman formula estimator and the new formula estimator can then take place. We set r_e = 0.2, 0.6, or 0.8 and r_t = 0.5 or 0.9; for an example of a case with no correlation of the errors, we consider a scenario in which r_t = 0.7 and r_e= 0. The next step involves generation of a single realization of the true data $\boldsymbol{T} = (T_x, T_y, T_z)'$ and generation of k = 1,...,M replicates for the errors $\boldsymbol{E} = (E_x, E_y, E_z)'$, each with n = 1000 (See Table 1.) The observed data are generated using the principle (stated in the Introduction) that the observed score equals the sum of the true value and its error score (e.g., $X = T_x + E_x$). Computing the bias and standard error based on the simulated data allows for evaluating the performance of each estimator for the different scenarios. The estimated bias is,

$$\widehat{Bias}(\hat{\rho}) = \frac{1}{M}\sum_{k=1}^{M}(\hat{\rho}_k - \rho), \tag{12}$$

in which $\hat{\rho}_k$represents the multiple correlation estimate for the kth replicate sample and ρ represents the true-score value as defined by using the true-score data to compute the correlations in (1).

The t-ratio measures the magnitude of the bias. This ratio appears as,

$$t_{bias} = \frac{\widehat{Bias}\,(\hat{\rho})}{\widehat{SE}\,(\hat{\rho})}, \tag{13}$$

with $\widehat{SE}(\hat{\rho}) = \left[\widehat{Var}(\hat{\rho})/M\right]^{1/2}$ for

$$\widehat{Var}(\hat{\rho}) = \widehat{MSE}(\hat{\rho}) - \left[\widehat{Bias}(\hat{\rho})\right]^2, \tag{14}$$

and estimated mean squared error,

$$\widehat{MSE}(\hat{\rho}) = \frac{1}{M}\sum_{k=1}^{M}(\hat{\rho}_k - \rho)^2. \tag{15}$$

As shown in Table 1, when no error correlation exists (last row of the table), the Spearman formula estimator and the new formula estimator perform similarly. However, as supported by the small absolute magnitude of the estimated bias, the new formula estimator outperforms the observed-score and Spearman formula estimators in all other scenarios with covarying error scores. The only scenario in which the new estimator has some evidence of bias pertains to an error correlation that remains small compared to the data correlation; however, the magnitude of the bias still supports large improvement over the observed-score and Spearman formula estimators. In general, the new formula estimator tends to overestimate the truth (bias tends to be positive) except when the error correlation is larger than the data correlation. The observed-score estimator tends to underestimate the true-score value in all instances except when the error correlation is larger than the data correlation. In all scenarios, the Spearman formula estimator tends to overestimate the truth; further, the Spearman formula estimator tends to drastically overcorrect the observed-score estimator to the point at which it has the largest absolute magnitude in bias. The Spearman formula estimator only shows improvement, as compared to the observed-score estimator, in cases where the error correlation value is small.

Real-World Applications

The independent Swedish foundation, Gapminder, has compiled global data from a variety of sources, all accessible for free at https://www.gapminder.org/data/ (Gapminder, 2020). In this application, we consider life expectancy (X) as the dependent variable, predicted by the independent variables of CO2 emissions per

Table 1. Simulation results (M = 500 replicates) for scenarios in which data (n = 1000) are generated to have moderate or strong associations. Errors are generated to have either weak, moderate or fairly strong correlations. For reference, the last row includes a scenario with no error correlation

			Observed			Spearman			New		
r_t	r_e	Truth	Est.	Bias	t_{bias}	Est.	Bias	t_{bias}	Est.	Bias	t_{bias}
0.5	0.2	0.6034	0.4446	-0.1588	-162.54	0.7812	0.1779	95.93	0.6045	0.0012	0.820
	0.6		0.6361	0.0327	45.03	1.0679	0.4646	239.60	0.6025	-0.0009	-0.741
	0.8		0.7244	0.1211	220.56	1.2045	0.6011	388.74	0.6020	-0.0014	-1.196
0.9	0.2	0.9274	0.6339	-0.2935	-382.24	1.0676	0.1402	76.38	0.9311	0.0038	3.908
	0.6		0.8058	-0.1216	-257.42	1.3344	0.4070	267.93	0.9279	0.0005	0.838
	0.8		0.8863	-0.0411	-143.61	1.4536	0.5262	317.47	0.9275	0.0001	0.238
0.7	0	0.7686	0.4365	-0.3321	-338.03	0.7747	0.0061	3.40	0.7728	0.0042	2.926

Table 2. Pairwise Correlations for the Gapminder Data.

	ρ_{xy}	ρ_{xz}	ρ_{yz}
Observed	0.7554	0.6547	0.7113
Error	0.4683	0.2510	0.4213
True	0.8211	0.7527	0.8092

Table 3. Reliability for the Gapminder Data.

$\rho_{xx'}$	$\rho_{yy'}$	$\rho_{zz'}$
0.7519	0.8578	0.8514

person (Y) and percent of the population living in urban areas (Z). Gapminder formally defines these variables as follows:

- Life expectancy pertains to the average age (in years) that an individual would live if the current mortality patterns for a given year would continue. Source information: https://www.gapminder.org/data/documentation/gd004/
- Carbon Dioxide (CO_2) emissions measures the amount of CO_2 (tonnes per person) being emitted from the burning of fossil fuels in a country. Source: Carbon Dioxide Information Analysis Center (CDIAC) at https://cdiac.ess-dive.lbl.gov/ through www.gapminder.org. Values are highly right-skewed, so data are log-transformed for all analysis.
- The percent of the population living in urban areas, measured as a percent of the total population of a country in a given year, compiled by The World Bank with percents calculated by the Statistics Division of the United Nations Department of Economic and Social Affairs and are smoothed by the UN Population Division. Source: The World Bank at https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS through www.gapminder.org.

This application uses 46 years-worth of available data (1969-2014). Although the Gapminder site provides data for 194 countries around the world, seven countries have missing values in all years for at least one variable, decreasing the sample size to 187. Values for 2014, the most recent year for which Gapminder provides data for all three variables, constitute observed scores. The mean of scores from the years 1970-2013 constitutes each country's corresponding true score, based upon the psychometric principle that the long-run average is equivalent to the true score (Allen & Yen, 2001). The difference between the true and observed scores for a country represents the error score for that country. This piece's Supplementary Materials includes tables in both .csv and .pdf formats that contain observed, true, and error scores for all variables used.

Table 2 contains the pairwise coefficients obtained from this data and used in subsequent calculations of multiple correlation coefficients. To determine reliability values, we rely upon test-retest logic (Babbie, 2017), calculating the correlations between the most recent (2014) and earliest (1969) years of data for the variables collected. Table 3 contains these reliability values.

Of particular interest is the multiple correlation value, $\rho_{T_x.T_yT_z}$, as this exercise attempts to demonstrate that the value produced by the new formula estimates the true-score multiple correlation coefficient better than both the observed-score coefficient and the coefficient that incorporates Spearman's correction. The closest estimate of $\rho_{T_x.T_yT_z}$in this application is the true-value coefficient of 0.8348, obtained using the true data with (1). Steps to calculate the new formula value, using the data available in the Supplementary Materials to this piece and following the process demonstrated by equations (4)-(9), appear in the Appendix.

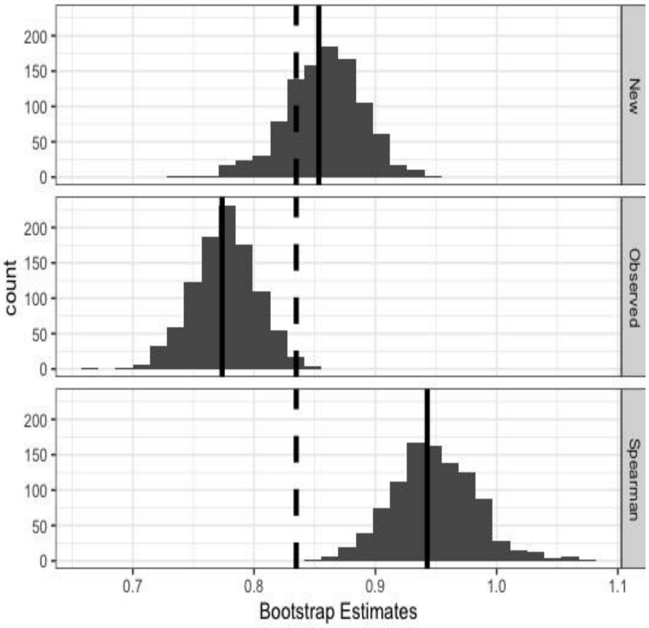
Table 4 contains the observed-score, Spearman formula, and new formula estimates. Standard errors and 95% percentile confidence intervals are computed based on 1000 bootstrap samples. (e.g., We sample the dataset with replacement 1000 times, each sample the same size, *n* = 187, as the original. See Efron and Tibshirani (1994), for example.) Histograms of the bootstrap samples appear as Figure 1. Estimates further support the patterns seen in the simulations. The new formula estimate slightly overestimates, but falls closest to the true value of the multiple correlation. The observed-score estimate underestimates the truth and the Spearman formula value drastically overestimates the truth. The 95% percentile confidence intervals further support the effectiveness of the new formula value as the interval for this estimate is the only one that contains the true-value multiple correlation coefficient.

Overall, the closest estimate of the strength of the true linear relationship of life expectancy with carbon dioxide emissions and percent of population living in urban areas is the true-value coefficient of 0.8348. Knowledge of how the actual true-score coefficient compares to the proposed estimates, supports the new formula's legitimacy. The outcome of this exercise, along with the simulations, suggests that researchers should use the new estimator whenever possible to determine multiple correlation coefficient values.

Table 4. Comparisons of multiple correlation formula estimates. The value produced by the new formula most closely estimates the true value of the multiple correlation, which is 0.8348.

	Estimate	SE	95% CI
Observed Score	0.7736	0.0258	(0.7231, 0.8253)
Spearman	0.9428	0.0342	(0.8832, 1.0164)
New Formula	0.8534	0.0311	(0.7890, 0.9134)

Figure 1. Bootstrap distributions for the new formula, observed-score, and Spearman formula estimates. The dashed line represents the true value of 0.8348 and the solid lines represent the three different point estimates.



Discussion

The methods and examples provided here serve two purposes. First, they illustrate the need to account for covarying error scores when estimating the multiple correlation between true scores. Prior proposed estimators — both the standard observed multiple correlation coefficient and the coefficient corrected according to Spearman's formula (1904) – ignore any covariance between the errors. Second, this piece's methods and examples offer an approach to acknowledging covarying error scores in the form of a new formula for an estimator that corrects the multiple correlation coefficient for attenuation. This formula produces values that lie nearer to $\rho_{T_x.T_yT_z}$than the values produced with previously used formulas.

Further consideration of the first point, the need to account for correlated errors, is highlighted by the fact that linear relationships between observed scores can actually reflect quite a bit of error-score correlation. A comparison of the observed-score and error-score correlation values in Table 2 for the real-world data application provides such evidence. Additionally, it stands to reason that error scores and true scores may covary despite the psychometric assumption that $\rho_{T_xE_x} = 0$. The new formula described here does not address this possibility; however perhaps it should receive attention.

The second point, noting the new formula's improvement upon existing estimators of the multiple correlation between true scores, both validates and challenges psychometric theory. Theoretically, true-score values correlate more strongly than observed-score values do; and, accordingly, a coefficient corrected for attenuation should exceed the uncorrected coefficient. This exercise confirms that using both the standard correction formula (Spearman, 1904) and the new formula increases the correlation from the uncorrected value. The simulation and the real-world data example in the Applications section of this text, in fact, demonstrate that both of these values tend to overestimate the correlation between true scores. Psychometric theory, however, does not address overestimation of true-score values, suggesting that correlations corrected for attenuation lie between the observed-score and true-score coefficients. Although this tendency to overestimate deserves additional attention, it does not discount from the merit of the new formula. Both the simulated data example and the authentic data application indicate that the Spearman value grossly overestimates the true-score value with the Spearman-corrected coefficient actually lying farther from the true-score value than the observed-score correlation coefficient does. In contrast, the new-formula value overestimates the true-score correlation only slightly and, therefore, provides the best estimate of this value.

Admittedly, obtaining this best estimate requires knowledge about information to which the typical researcher does not have access. Values of the terms $\rho_{E_xE_y}$, $\rho_{E_xE_z}$, and $\rho_{E_yE_z}$, generally remain unknown in practical situations. These terms are part of the new formula due to its reliance on Zimmerman and Williams (1977) correction for attenuation of pairwise coefficients with the presence of error-score covariance. As stated in the Methods section of this piece, Zimmerman and Williams' correction pertains to theoretical contexts in which researchers can obtain error-score correlation values. The presence of error-score correlations in the new formula makes it most germane to theoretical contexts as well. The examples presented in the Applications section of this piece use psychometric principles to determine error scores simply for the purpose of demonstrating the new formula's effectiveness. Those who wish to use the formula in applied situations, in which error scores of data generally remain unknown and cannot be obtained using psychometric principles, can take advantage of programs such as LISREL8-11 structural equation modeling software (Jöreskog & Sörbom, 2006) to estimate error components of observed scores. Having obtained error scores, and with knowledge of reliability values, one can use the new correction formula for multiple correlation coefficients. Because the arithmetic required by the formula, although manageable, can become tedious, those wishing to obtain the corrected value might appreciate applications that perform the calculations. Supplements to this piece provide an R function (requiring user input of the observed data, error data, and reliability values) and an Excel function (requiring user input of observed pairwise coefficients, error-score coefficients, and reliability values) that provide the multiple correlation coefficient corrected with the new formula.

Although the assumption of independent errors simplifies calculations, researchers simply cannot make this assumption. In some cases, error scores may correlate so slightly that the new formula proposed in this piece and incorporating Spearman's (1904) formula into the multiple correlation equation produce almost identical values. However, one cannot expect such negligible error-score correlations to always exist as they often don't. Use of the new formula indicates recognition of the potential for a relationship between error scores to impact the perceived linear relationship. Accounting for correlated errors, as illustrated in this work, is, thus, crucial to understanding the underlying relationships between variables.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

B. R. B. (1924). Multiple correlation. *The Journal of Educational Research*, 149–154.

Babbie, E. R. (2017). *The basics of social research*. Cengage Learning.

Bacon, H. (1938). Note on a formula for the multiple correlation coefficient. *The Annals of Mathematical Statistics*, *9*(3), 227–229.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models*. Chapman & Hall.

Dudgeon, P. (2017). Some improvements in confidence intervals for standardized regression coefficients. *Psychometrika*, *82*(4), 928–951.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, *60*(4), 299–302.

Gapminder. (2020). *Data*. www.gapminder.org/data/. (Accessed June, 2020)

Hakstian, A. R., Schroeder, M. L., & Rogers, W. T. (1988). Inferential procedures for correlation coefficients corrected for attenuation. *Psychometrika*, *53*(*1*), 27–43.

Hardin, J. W., & Carroll, R. J. (2003). Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *The Stata Journal*, *3*(*4*), 342–350.

Hardin, J. W., Schmiediche, H., & Carroll, R. J. (2003). The regression-calibration method for fitting generalized linear models with additive measurement error. *The Stata Journal*, *3*(4), 361–372.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL for windows [computer software]. *Lincolnwood, IL: Scientific Software International*.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101.Szpiro, A. A., Rice, K. M., & Lumley, T. (2010). Model-robust regression and a bayesian "sandwich" estimator. *The Annals of Applied Statistics*, *4*(4), 2099–2113.

Wetcher-Hendricks, D. (2006). Adjustments to the correction for attenuation. *Psychological Methods*, *11*(2), 207.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology*, *16*(*2*), 135–152.

Send correspondence to: Brenna Curley
Moravian University
Email: curleyb@moravian.edu

Appendix: New Formula Calculations

The operations shown here demonstrate the substitutions and arithmetic one could use to compute the new formula estimate defined in (6)-(9). The calculations illustrated below provide evidence that the formula does, in fact, produce the value, 0.853, identified in the real-world data application presented in this piece. However, the R and Excel functions suggested within the Discussion (and shared as part of the Supplementary Materials) produce the new corrected multiple correlation coefficient more efficiently than performing the calculations by hand. Both of these built functions were used to verify the computed value below.

Calculations begin by inserting the pairwise values (see Table 2) into equations (7)-(9) and performing the indicated operations:

$$\phi_{xy} = 0.755 - 0.468\sqrt{1-0.752}\,\sqrt{1-0.858}\ , \tag{16}$$

$$\phi_{xz} = 0.655 - 0.251\sqrt{1-0.752}\,\sqrt{1-0.851}\ , \tag{17}$$

and

$$\phi_{yz} = 0.711 - 0.421\sqrt{1-0.858}\,\sqrt{1-0.851}\ . \tag{18}$$

The indicated arithmetic operations produce values (rounded to three decimal places) of
0.667 for ϕ_{xy}, 0.607 for ϕ_{xz}, and 0.650 for ϕ_{yz}. Then, using these values, along with reliability values (also rounded to three decimal places), in equation (6) leads to the following calculations:

$$\rho_{T_x.T_yT_z} = \sqrt{\frac{0.851(0.667)^2-2(0.667)(0.607)(0.650)+0.858(0.607)^2}{0.752[(0.858)(0.851)-0.650^2]}}. \tag{19}$$

The resulting value, 0.853, represents an estimate of the multiple correlation between true scores without making the assumption of independent error scores.