Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 21, Number 6, March 2016 ISSN 1531-7714

A Comparison of Three Approaches to Correct for Direct and Indirect Range Restrictions: A Simulation Study

Andreas Pfaffel, *University of Vienna*
Barbara Schober, *University of Vienna*
Christiane Spiel, *University of Vienna*

A common methodological problem in the evaluation of the predictive validity of selection methods, e.g. in educational and employment selection, is that the correlation between predictor and criterion is biased. Thorndike's (1949) formulas are commonly used to correct for this biased correlation. An alternative approach is to view the selection mechanism as a missing data mechanism. The aim of this study was to compare Thorndike's formulas for direct and indirect range restriction scenarios with two state-of-the-art approaches for handling missing data: full information maximum likelihood (FIML) and multiple imputation by chained equations (MICE). We conducted Monte-Carlo simulations to investigate the accuracy of the population correlation estimates in dependence of the selection ratio and the true population correlation in an experimental design. For a direct range restriction scenario, the three approaches are equally accurate. For an indirect range restriction scenario, the corrections using FIML and MICE are more precise than when using Thorndike's formula. The higher the selection ratio and the true population correlation, the higher the precision of the population correlation estimates. Our findings indicate that both missing data approaches are alternative corrections to Thorndike's formulas, especially in the case of indirect range restriction.

A common methodological problem in the evaluation of the predictive validity of a selection method (e.g., a psychometric test or an interview), is that of estimating the population correlation ρ between a selection method (predictor) and a certain criterion for success based on a sample of selected individuals. This so-called *range restriction problem* in correlation analysis arises because the observed selected sample is not random, and therefore not representative of the applicant population (Sackett & Yang, 2000; Thorndike, 1949). As an inherent effect of the selection, values for the criterion variable are available only for selected applicants. This problem, for example, occurs in the evaluation of the predictive validity of an admission test in higher education, because data of academic success are only available for applicants who were admitted to the program. Another example is personnel selection, when we want to estimate the correlation between a knowledge test and job performance, but we only have job performance data from those individuals who were hired. Consequently, the Pearson correlation coefficient r obtained from a selected sample is a biased estimation of the population correlation ρ (Alexander, 1990; Bobko, 1983; Duan & Dunlap, 1997; Raju & Brand, 2003; Sackett & Yang, 2000). Hence, this biased estimate r has to be corrected to provide a more valid estimate of ρ.

Thorndike (1949), following Pearson (1903) and Lawley (1943), presented formulas to correct the biased sample correlation r_{XY} between a predictor X and a criterion Y for the two most common selection scenarios, typically in educational and employment selection: (A) The explicit or direct range restriction scenario (DRR), in which the selection is based directly on the predictor variable *X*, and (B) the incidental or indirect range restriction scenario (IRR), in which the selection is based on a third variable *Z*, different to the predictor of interest (for a detailed description of DRR and IRR scenarios see the next subsection 'Range Restriction Scenarios: Direct and Indirect'). Thorndike's formulas have been widely studied (Duan & Dunlap, 1997; Holmes, 1990; Linn, 1983; Ree, Carretta, Earles, & Albert, 1994), and have often been applied to correct for range restriction, e.g. in predictive validity studies of large-scale testing programs such as the Graduate Record Examination (GRE) (Chernyshenko & Ones, 1999), or the Graduate Management Admission Test (GMAT) (Sireci & Talento-Miller, 2006). Correcting for range restriction has also been applied in other fields, e.g. in predicting job performance (SjöBerg, SjöBerg, Näswall, & Sverke, 2012), or to predict scores on a practical driving-license test (Wiberg & Sundström, 2009). Range restriction is also an important issue in validity generalization (Hunter, Schmidt, & Le, 2006; Murphy, 2003).

An alternative approach correcting for range restriction is to view the selection mechanism as a missing data mechanism (Mendoza, 1993; Wiberg & Sundström, 2009), see subsection 'Range Restriction as a Missing Data Mechanism'. There are many advantages to view the selection mechanism as a special case of missing data, as comprehensive statistical literature on dealing with missing data exists, and a variety of techniques and research results are available (Little & Rubin, 2002; Rubin, 1996, 2004; Schafer & Graham, 2002). So far, this state-of-the-art approach in dealing with missing values has been very seldom used for range restriction problems (Pfaffel, Kollmayer, Schober, & Spiel, 2016; Wiberg & Sundström, 2009). Wiberg and Sundström (2009) applied this approach to data from a Swedish driving-license test to correct for a DRR scenario. Their findings indicate that the missing data approach provides an effective estimate of the population correlation. However, Wiberg and Sundström (2009) pointed out that simulations of different population correlations and different selection ratios are necessary to investigate the accuracy of the correction of the proposed missing data approach.

In the present paper, we apply this missing data approach to both a DRR scenario and an IRR scenario, and compare this approach with Thorndike's (1949) correction formulas. First, we describe the mechanisms of loss of criterion data in the case of DRR and IRR scenarios and show the data matrix used for the correction. Second, we describe the theoretical assumptions necessary to apply a missing data approach to the two scenarios. Third, we investigate the accuracy of this proposed correction by conducting Monte Carlo simulations, which allow for a comparison of the corrected correlation with the true population correlation in an experimental design. Finally, the results of the comparison of the three approaches are discussed.

Range Restriction Scenarios: Direct and Indirect

The most straightforward selection scenario is the direct range restriction (DRR) scenario (Sackett & Yang, 2000; Thorndike, 1949). Selection is based directly on the predictor variable *X* from the top down, assuming a positive relationship between predictor *X* and criterion *Y*. The predictor variable *X* can be either a single score, as in a single-selection method (e.g., a psychometric test), or a composite score derived from several selection methods (e.g., a psychometric test and a quantitative interview). In the case of a DRR scenario, the predictor variable itself is the selection variable, which is of interest in evaluating the predictive validity of a selection method or a composite score. For example, in higher education in Austria, prospective students are selected for various study programs solely on the basis of entrance examinations (e.g., Medical University of Vienna, 2015). In the case of DRR, values of *X* are available for all applicants whereas values of *Y* are only available for selected applicants.

The indirect range restriction scenario (IRR) occurs when applicants are selected on another variable *Z*, which is usually correlated with *X*, *Y*, or both (Sackett & Yang, 2000). Suppose a selection procedure consists of a psychometric test *X* (predictor of interest) and a quantitative interview. For example, if we use the composite score as selection variable *Z*, and we want to evaluate the predictive validity of the psychometric test *X*, then we have an IRR scenario for *X*. Organizations often use a composite score for selection but would still like to know the predictive validity of each individual selection method in order to increase the predictive validity of the whole selection procedure, e.g., by removing or giving more weight to a particular selection method. In the case of IRR, values of *X* and *Z* are available for all applicants, whereas values of *Y* are available for selected applicants only. In the Appendix, we present a numerical example of a selection scenario, in which prospective students completed an aptitude test and an interview.

In both scenarios, we have missing values in the criterion variable *Y* for non-selected applicants. The amount of data loss depends on the selection ratio (SR), which is defined as the ratio of available places to the number of applicants. The SR ranges between 0 and 1, or between 0% and 100%. For example, if 200 study places are available and 500 applicants apply for them, the SR is 200 divided by 500 or 40%. The top 40% of applicants will be selected and 60% will not be selected. Hence, we have missing values of *Y* for 60% of the applicants. Figure 1 shows the data matrix observed under a DRR scenario and an IRR scenario (Chan & Chan, 2004; Li, Chan, & Cui, 2011). X_r, Y_r, and Z_r are the values of *X*, *Y*, and *Z* obtained from the selected (restricted) sample, X_u and Z_u are the values of *X* and *Z* obtained from the unselected sample. Values of the criterion *Y* are not available for the unselected sample.

$$\text{a)}\quad \begin{array}{c|c} X_r & Y_r \\ X_u & - \end{array} \qquad\qquad \text{b)}\quad \begin{array}{c|c|c} Z_r & X_r & Y_r \\ Z_u & X_u & - \end{array}$$

Figure 1. Structure of the data matrix observed under a) a DRR scenario, and b) an IRR scenario.

Due to the fact of selection, the observed correlation coefficient r_{XY} underestimates the population correlation. The reduction of the correlation r_{XY} is given by the reduction of the covariance (the numerator in Equation 1) relative to the reduction of the product of the sample standard deviations s_X and s_Y (the denominator in Equation 1).

$$r_{XY} = \frac{cov(X,Y)}{s_X \cdot s_Y} \qquad (1)$$

For example, if we select the top 40% of applicants in a DRR scenario, the predictor *X* is restricted in range in the selected sample. If we look only at the standard deviation of *X*, we will see that the standard deviation of *X* in the selected sample (the top 40%) is smaller than for all applicants. After all, the reduction of the Pearson correlation increases as the SR decreases, assuming the correlation between *X* and *Y* does not equal zero.

The most famous and widely used formulas to correct the biased correlation coefficient were presented by Thorndike (1949). The formula for the DRR scenario is (Sackett & Yang, 2000, p. 114):

$$\hat{r}_{XY} = \frac{(S_X/s_X)r_{XY}}{\sqrt{1+r_{XY}^2(S_X^2/s_X^2-1)}} \qquad (2)$$

where $\hat{r}_{XY}$ is the point estimate of the population correlation, r_{XY} is the uncorrected Pearson correlation coefficient obtained from the restricted sample, s_X is the standard deviation of *X* for the restricted sample, and S_X is the standard deviation of *X* for the unrestricted population. The core term for correcting r_{XY} is the ratio S_X/s_X. The correction formula works because $\hat{r}_{XY} > r_{XY}$ if $S_X > s_X$.

In the case of an IRR scenario, the correction formula is (Sackett & Yang, 2000, p. 115):

$$\hat{r}_{XY} = \frac{r_{XY}+r_{ZX}\cdot r_{ZY}\left(S_Z^2/s_Z^2-1\right)}{\sqrt{1+r_{ZX}^2\left(S_X^2/s_X^2-1\right)}\cdot\sqrt{1+r_{ZY}^2\left(S_X^2/s_X^2-1\right)}} \qquad (3)$$

where r_{XY}, r_{ZX}, and r_{ZY} are the Pearson correlation coefficients obtained from the restricted sample, and s_z and S_z are the standard deviations of variable *Z* for the restricted sample and the unrestricted population. Both correction formulas require linearity between *X* and *Y*, and homoscedasticity (the probability distribution of the error term is the same in the restricted sample and in the population).

Range Restriction as a Missing Data Mechanism

As an inherent effect of the selection, we have missing values in the criterion variable *Y*, as shown in Figure 1. Therefore, it seems reasonable to view the range restriction problem as a missing data mechanism (Mendoza, 1993; Wiberg & Sundström, 2009). First, we give a brief overview of the three established missing data mechanisms in order to locate the range restriction problem in this line of research. After that, we introduce two state-of-the-art techniques for dealing with missing data.

Rubin (1976) outlined a theoretical classification scheme for missing data problems that is widely used in the scientific literature today. His so-called *missing data mechanisms* are theoretical assumptions necessary for analyzing missing data (Enders, 2010). Three mechanisms describe the relationship between the probability of missing values and measured variables (Little & Rubin, 2002): (1) MCAR means missing completely at random, i.e. the probability of missing data on a variable *Y* is unrelated to other measured variables and is unrelated to the values of *Y* itself. (2) MAR means missing at random, i.e. the probability of missing data on a variable *Y* is related to some other measured variable (or variables) in the analysis model but not to the values of *Y* itself. MAR is more general and often more realistic than MCAR. Modern missing data methods generally assume the MAR mechanism. (3) MNAR means missing not at random, i.e. the probability of missing data on a variable *Y* is related to the values of *Y* itself, even after controlling for other variables.

We consider the two selection scenarios discussed here (DRR and IRR) to be MAR, because there is no relationship between the probability of missing values for *Y* and the values of *Y* after partialling out other variables. The probability of missing data for *Y* depends on *X* (in a DRR scenario), or on *Z* (in an IRR scenario), but not on the values of *Y* itself.

Over the past few decades, methodologists have suggested various techniques for dealing with missing data, but several of them (e.g., listwise or pairwise deletion, and single imputation) are no longer considered state-of-the-art because they have potentially serious drawbacks (Enders, 2010). For example, single regression imputation overestimates correlations and attenuates variances and covariances even when the data are MCAR (Enders, 2010; Schafer & Graham, 2002). The problem is that all imputed values fall directly on the regression line and therefore lack variability. Single imputation techniques are not suitable for many reasons, especially with regard to estimating correlation coefficients. There are two approaches that methodologists currently regard as state-of-the-art (Schafer & Graham, 2002): (1) Full information maximum likelihood (FIML), and (2) multiple imputation (MI). Both missing data approaches make the same assumptions with regard to the missing data mechanism (MAR), have similar statistical properties, and frequently produce equivalent results (Enders, 2010; Graham, Olchowski, & Gilreath, 2007).

Full information maximum likelihood is a technique for estimating the most plausible parameters that produce the best fit to the data by maximizing the log-likelihood function. In other words, the goal is to identify those population parameter values that have the highest probability of producing the data of a certain sample. The basic estimation process in the case of missing data is largely the same as in the context of complete data. The first step is to specify the distribution of the population data, which in the social and behavioral sciences is commonly assumed to be multivariate normally distributed (Enders, 2010).

Finding those parameters that maximize the log-likelihood function is possible with iterative optimization algorithms, e.g. the expectation maximization (EM) algorithm, the Newton-Raphson method, or Bayesian simulation. The EM algorithm, or more broadly the generalized expectation maximization algorithm (GEM), is most important for missing data analyses. For readers interested in the mathematical details of EM-based maximum likelihood estimation, we refer to Dempster, Laird, and Rubin (1977), and Meng and Rubin (1993). An extension to non-normal data and missing values in covariates is possible under the broad class of generalized linear models (Ibrahim, Chen, Lipsitz, & Herring, 2005). For an overview of likelihood-based techniques with mathematical descriptions, see the book by Little & Rubin (2002).

The second state-of-the-art approach is multiple imputation, which has emerged as a flexible alternative to the likelihood-based approach for a wide variety of missing-data problems (Schafer & Graham, 2002; van Buuren, 2012). A multiple imputation analysis consists of three distinct phases: the imputation phase, the analysis phase, and the pooling phase. The imputation phase generates *m* complete datasets with plausible estimates of the missing values based on one dataset with missing values. Each of the complete datasets contains different estimates of the missing values, but identical values for the observed data. In contrast to single imputation, multiple imputation builds the uncertainty with regard to parameter estimates into the imputation model, meaning that the estimates of the missing values vary among the *m* complete datasets. In the analysis phase, conventional statistical methods can be applied to each complete dataset with each statistical method performed *m* times, once for each complete dataset. The pooling phase combines the *m* parameter estimates into a single set of parameter estimates. A pooled parameter estimate is typically the arithmetic average of the *m* estimates from the analysis phase (Rubin, 2004).

Multiple imputation is typically (but not necessarily) performed within a Bayesian framework, in which the parameters are drawn from their respective posterior distributions. In the case of incomplete multivariate normal data, calculating the posterior distribution is possible with the data augmentation algorithm (Schafer, 1997; Tanner & Wong, 1987). A general approach that can also handle non-normal data with missing values in the covariates is multivariate imputation by chained equations (MICE), also known as fully conditional specification (FCS) (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren, 2007, 2012). The imputation model is specified as a regression model for each incomplete variable involving the other variables as predictors. For example, the MICE algorithm is implemented in the R software package `mice` (van Buuren & Groothuis-Oudshoorn, 2011). Imputation techniques for numerous types of missing data problems receive excellent treatment in the book by van Buuren (2012).

To sum up, in both range restriction scenarios, we consider the missing data mechanism to be missing at random (MAR). In the case of MAR, the population parameters can be estimated based on the available data. Full information maximum likelihood estimation as well as multiple imputation meet the assumptions for handling the missing values in the criterion variable. Hence, the two approaches seem to be effective at providing unbiased estimates for the population correlation, and therefore good alternatives to Thorndike's correction formulas.

Aim of this Study

The aim of this study was to compare the accuracy of the corrections made using three approaches – (1) Thorndike's well known and most commonly applied correction formulas for DRR (Equation 2) and IRR (Equation 3), (2) full information maximum likelihood estimation, and (3) multiple imputation by chained equations – for direct and indirect range restriction scenarios depending on the selection ratio and the true population correlation.

Method

Procedure

We conducted two Monte Carlo simulations (DRR and IRR scenarios) using the program R-Statistics (R Core Team, 2014) to investigate the accuracy of the corrections made using the three approaches: (1) Thorndike's correction formulas for DRR and IRR, (2) full information maximum likelihood estimation, and (3) multiple imputation by chained equations. The Monte Carlo simulations were conducted with 5,000 trials for each of the two scenarios. The simulation procedure consisted of the following four steps.

Step 1 – Data simulation: We generated 5,000 unrestricted data sets (sample size $N = 500$) drawn from a multivariate normal distribution by varying the Pearson correlation coefficient between X and Y from .10 to .90. Additionally, in the case of IRR we varied not only the correlation coefficient between X and Y but also the correlations between Z and X, and Z and Y from .10 to .90.

Step 2 – Selection: We simulated the selection for nine selection ratios ranging from 10% to 90% with step width 10%, which corresponded to a proportion of missing values in Y from 90% to 10%. We selected those cases with the highest values in X (DRR) or with the highest values in Z (IRR) respectively. The percentage of selected cases depended on the selection ratio. Values in Y for non-selected cases were converted into missing values. The restricted sample created in this way was saved into a new data set and was used in applying the correction.

Step 3 – Correction: The three approaches were applied to the data set of the restricted sample (missing values in Y).

Step 4 – Analysis of parameter estimates: We compared the estimated correlation of the three approaches with the correlation obtained from the unrestricted population. In order to investigate the accuracy of the correction, we calculated the residuum of the population correlation estimate $\hat{r}_{XY} - \rho_{XY}$.

Correction

In order to correct for direct and indirect range restriction scenarios, the three approaches were applied to the restricted sample. In the first approach, we used Thorndike's correction formulas for DRR (Equation 2) and IRR (Equation 3). The results of these formulas are the estimates of the population correlation. Second, we used full information maximum likelihood estimation using the R package `mvnmle` (Gross & Bates, 2012), which provides a ML estimation for multivariate normal data with missing values. Third, we used multiple imputation by chained equations to replace the missing values of the criterion variable before estimating the population correlation. We used the R package mice (van Buuren & Groothuis-Oudshoorn, 2011) with the default specifications for the prior distributions and the Markov Chain Monte Carlo simulation, but we changed the number of imputations *m* from 5 (default) to 20. Conventional wisdom suggests that multiple imputation analysis requires about *m* = 5 imputations (Rubin, 2004; Schafer, 1997). This number of imputations was derived solely by considering the relative efficiency (Enders, 2010; Rubin, 2004). Contrary to this conventional wisdom, simulations studies show that only analyses based on *m* = 20 imputations yield comparable power to a maximum likelihood analysis and are therefore sufficient for many situations (Graham et al., 2007).

Analysis

In order to investigate and to compare the accuracy of the three correction methods, we analyzed the residual density of the population correlation estimates. Accuracy is defined as the closeness of the estimated value to the true value of the parameter being estimated (Ayyub & McCuen, 2011). The concept of accuracy encompasses both trueness and precision, and therefore provides important quantitative information about the goodness of the correction. The trueness is also known as bias or systematic error, and the precision as random error. If the residual value $\hat{r}_{XY} - \rho_{XY}$ is close to zero, then a correction method provides a very good estimation of the population correlation. We used the arithmetic mean of the residuals (over the 5000 Monte Carlo trials) as a measure of trueness, and the standard deviation of the residuals as a measure of precision. Figure 2 shows a graphical illustration of trueness and precision. A positive mean of the residuals represents an overestimation of the population correlation, while a negative mean of the residuals represents an underestimation. A smaller value for the standard deviation of the residuals represents a lower shape of the density, which means the estimate of the population correlation is more precise.

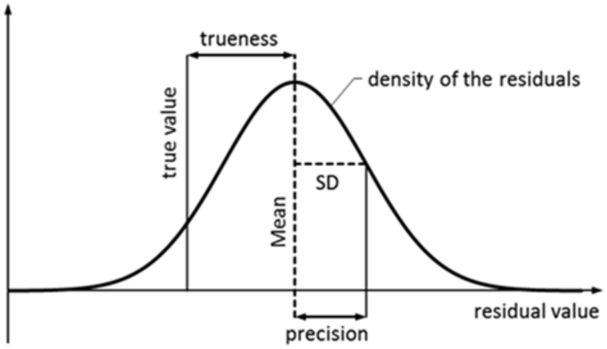

Figure 2. Graphical illustration of the concept of accuracy (trueness and precision).

In order to investigate the effect of the population correlation between predictor *X* and criterion *Y* on the accuracy of the correction, we partitioned the true population correlation coefficients into three levels: a weak correlation (from .10 to <.40), a moderate correlation (from .40 to <.70), and a strong correlation (from .70 to .90). With regard to the comparison of the three approaches, it is primarily of interest, whether the strength of the population correlation has a differential effect on the accuracy of the three approaches. In other words, is there an interaction between population correlation and approach? If the effect is not differentiated, we should observe the same changes in the accuracy of the estimates depending on the population correlation for each approach.

Results

Figure 3 shows 12 examples of histograms of the residuals of the population correlation estimate $\hat{r}_{XY}$. The histograms are arranged as follows: In the vertical direction, the three approaches Thorndike, MICE and FIML; in the horizontal direction, the two scenarios DRR and IRR for two selection ratios of 30% and 50%. In both scenarios, the residuals $\hat{r}_{XY} - \rho_{XY}$ are symmetrically distributed around zero, and the standard deviations of the residuals are smaller for a selection ratio of 50% than for a selection ratio of 30%. Thus, the trueness of $\hat{r}_{XY}$ for the three approaches is very high, and the precision increases as the selection ratio increases. In the DRR scenario, there are no significant differences between the standard deviations of the residuals of the three approaches (Bartlett's test for equal variances: all *p*'s > .05). In the IRR scenario, the standard deviations of the residuals of the three approaches are lower in comparison to the standard deviations of the residuals in the DRR scenario. Thorndike's correction formula for an IRR scenario is less precise than the correction with MICE or FIML (all *p*'s < .001), but there are no significant differences in the standard deviations between MICE and FIML (for more detailed information, Table 1 shows the mean values and the standard deviations of the residuals for all nine selection ratios).

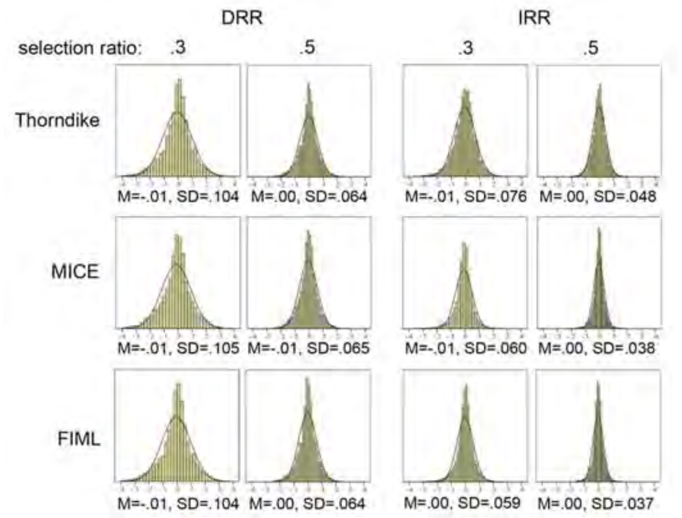

Figure 3. Distribution of the residuals for the population correlation estimates for the three approaches (Thorndike, MICE, FIML), for DRR and IRR, and for selection ratios of 30% and 50%.

As seen in Figure 3, the precision of the population correlation estimate decreases as the selection ratio increases. In order to take a closer look at this relationship, we examined the type of relationship between the standard deviation of the residuals and the selection ratio. Figure 4 shows that the standard deviation of the residuals experiences positive acceleration as the selection ratio decreases. For an IRR scenario (Figure 4b), the standard deviation of the residuals increases faster for Thorndike's correction formula than for the two missing data approaches MICE and FIML. For both scenarios, this relationship can be statistically modeled by an exponential function ($R^2 \geq .983$, $p < .001$, see Table 2). The results show that the precision of the population correlation estimates decreases exponentially as the selection ratio decreases (i.e., as the selection ratio becomes smaller).

Table 1. Accuracy of the population correlation estimates depending on the selection ratio for direct and indirect range restriction scenarios.

			DRR			IRR	
SR	Accuracy	Thorndike	MICE	FIML	Thorndike	MICE	FIML
	M	-.032	-.062	-.029	-.029	-.040	-.017
.1	SD	.237	.227	.238	.168	.126	.131
	M	-.013	-.027	-.012	-.011	-.016	-.006
.2	SD	.142	.141	.142	.103	.081	.080
	M	-.006	-.014	-.005	-.006	-.009	-.002
.3	SD	.104	.105	.104	.076	.060	.059
	M	-.004	-.009	-.004	-.004	-.005	-.002
.4	SD	.080	.081	.080	.060	.047	.046
	M	-.003	-.006	-.002	-.003	-.004	-.001
.5	SD	.064	.065	.064	.048	.038	.037
	M	-.001	-.003	-.001	-.002	-.003	-.001
.6	SD	.052	.053	.052	.039	.031	.030
	M	-.001	-.002	.000	-.001	-.001	.000
.7	SD	.042	.043	.042	.031	.025	.024
	M	.000	-.001	.000	-.001	-.001	.000
.8	SD	.033	.034	.033	.024	.019	.018
	M	-.001	-.001	.000	.000	.000	.000
.9	SD	.023	.023	.023	.017	.013	.013

Note. DRR = direct range restriction, IRR = indirect range restriction, SR = selection ratio, M = mean of the residuals of the population correlation estimate (trueness), SD = standard deviation of the residuals of the population correlation estimate (precision), Thorndike = Thorndike's correction formulas (Equation 2 and Equation 3), MICE = multiple imputation by chained equations, FIML = full information maximum likelihood estimation.

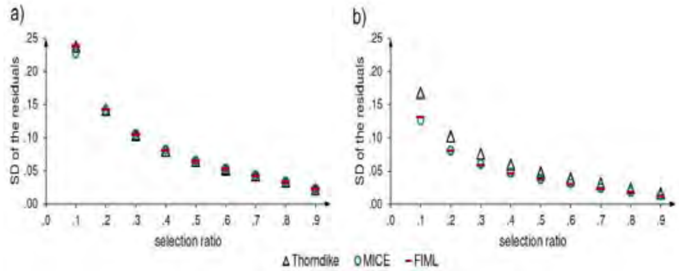

Figure 4. Exponential relationship between the selection ratio and the standard deviation of the residuals of the population correlation estimates for the three approaches (Thorndike, MICE, and FIML), for a) a DRR scenario, and b) an IRR scenario.

Table 2. Nonlinear regression analysis of the standard deviation of the residuals on the selection ratio.

	DRR			IRR		
	b_0	b_1	R^2	b_0	b_1	R^2
Thorndike	0.257	-2.668	.984	0.184	-2.620	.985
MICE	0.251	-2.603	.986	0.145	-2.633	.990
FIML	0.258	-2.673	.983	0.146	-2.684	.987

Note. Nonlinear regression analysis of the model function: $\mathrm{SD} = b_0 e^{b_1 \mathrm{SR}}$, SD = standard deviation of the residuals of the population correlation estimate (precision), SR = selection ratio, b_0 and b_1 = regression coefficients, DRR = direct range restriction, IRR = indirect range restriction, Thorndike = Thorndike's correction formulas (Equation 2 and Equation 3), MICE = multiple imputation by chained equations, FIML = full information maximum likelihood estimation.

In order to investigate the effect of the true population correlation between predictor and criterion on the accuracy of the population correlation estimates, we compared the means and standard deviations depending on three levels of the true population correlation. For both scenarios, there is no relevant effect of the true population correlation on the trueness of the correlation estimates, but there is an effect on the precision. Figure 5 and Figure 6 show the standard deviations of the residuals in dependence of the selection ratio, the true population correlation, and the three approaches. In addition to the effect of the selection ratio, the precision of the population correlation estimates increases as the true population correlation increases: for a DRR scenario $F(80, 2) = 9.603, p < .001, \eta_p^2 = .21$, and for an IRR scenario $F(80, 2) = 7.254, p = .001, \eta_p^2 = .16$.

With regard to the comparison of the three approaches, the true population correlation has no differential effect on trueness and precision. In other words, there is no significant interaction between population correlation and approach, F's(80, 4) < .01, p's > .99. For a DRR scenario, the precision of the three estimates is equal for weak, moderate, and strong true population correlations (see Figure 5). For an IRR scenario, as shown in Figure 6, the higher standard deviations of Thorndike's correction result from the fact that Thorndike's correction is less precise (compare with Figure 4b), but these differences are not affected by the true population correlation. As shown in Figure 6, the precision of Thorndike's estimate in the case of a moderate correlation corresponds to the precision of the estimates of MICE and FIML in the case of a weak correlation. However, this effect is small for selection ratios beyond 30%.

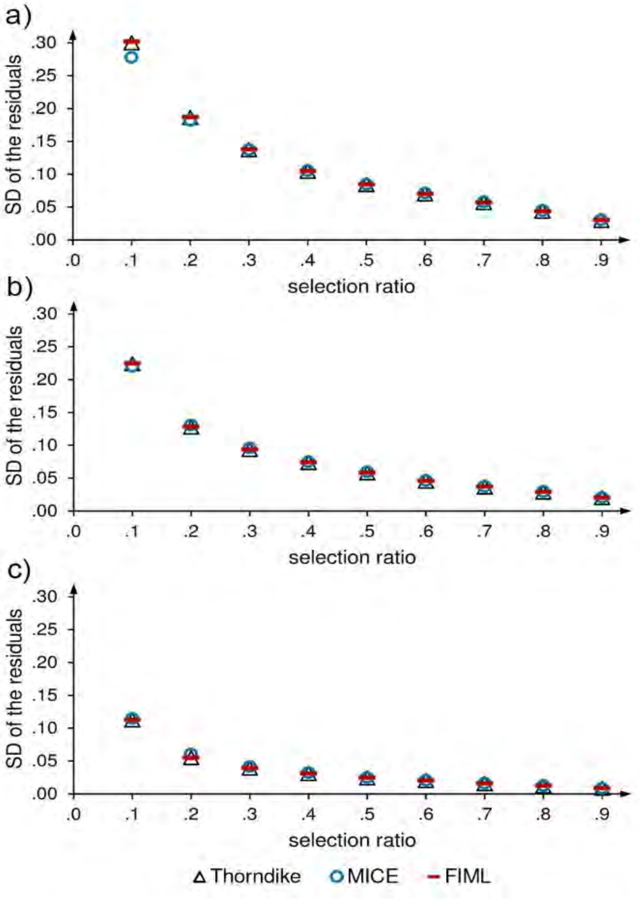

Figure 5. Effect of a a) weak, b) moderate and c) strong true population correlation on the precision of the population correlation estimates of the three approaches (Thorndike, MICE, FIML) for a DRR scenario.

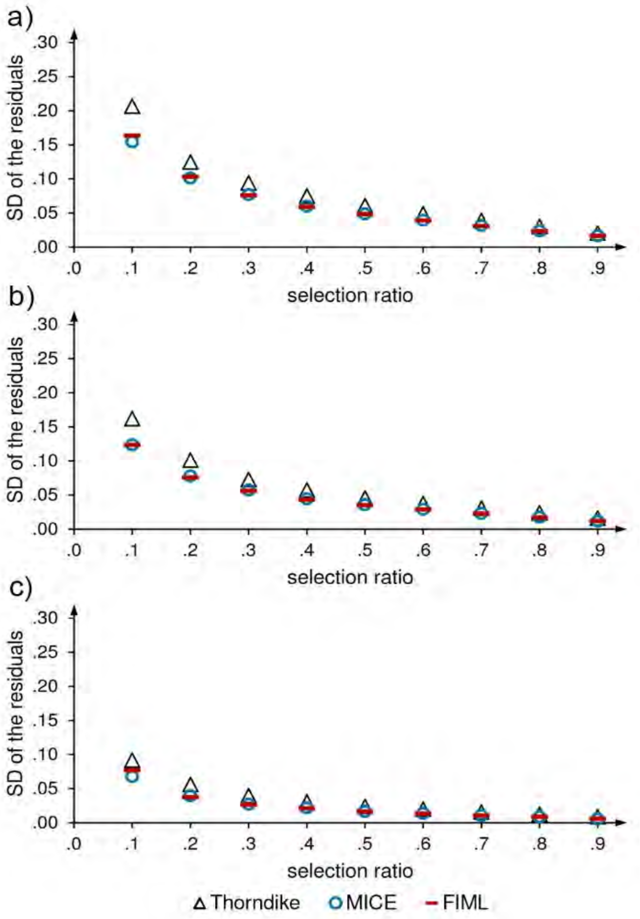

Figure 6. Effect of a a) weak, b) moderate and c) strong true population correlation on the precision of the population correlation estimates of the three approaches (Thorndike, MICE, FIML) for an IRR scenario.

Discussion

Range restriction is a common methodological problem in the evaluation of the predictive validity of a selection method. The correlation obtained from the selected sample is a biased estimate of the population correlation. An alternative approach to Thorndike's correction formulas is to view the selection mechanism as a missing data mechanism. The aim of this study was to compare the accuracy of the estimates of the population correlation for three approaches: 1) Thorndike's (1949) correction formulas, 2) multiple imputation by chained equations (MICE), and 3) full information maximum likelihood estimation (FIML) for direct (DRR) and indirect (IRR) range restriction scenarios.

The results show that the two missing data approaches perform effectively and provide unbiased estimates for both scenarios, though the correction for an IRR scenario is more precise than for a DRR scenario. For a DRR scenario, the three approaches are equallyaccurate. However, for an IRR scenario the correction using MICE or FIML is more precise than the correction using Thorndike's formula. An important finding is that the precision of the population correlation estimates decreases exponentially as the selection ratio decreases. Consequently, the confidence intervals of the point estimates are very wide for small selection ratios. This effect is of particular importance in the evaluation of the predictive validity in highly selective selection scenarios. In addition, if the population correlation between predictor and criterion is weak, then the prediction is less precise than in the case of a moderate or a strong population correlation. On the basis of our findings, we do not recommend corrections for range restriction for selection ratios lower than 30%, which translates into more than 70% missing values. The confidence interval of the population correlation estimate should be considered in evaluating the predictive validity. On the one hand, a cautious interpretation of correlations corrected for range restriction is necessary to avoid invalid conclusions about the predictive validity of a selection method. On the other hand, no range restriction correction is more likely to result in an invalid conclusion.

Our findings show that MICE and FIML provide similar results, and both approaches make the same assumptions with regard to the missing data mechanism. However, the two approaches differ in dealing with missing values, which may be relevant to the decision on their use in evaluation studies. In contrast to maximum likelihood estimation, multiple imputation generates several complete datasets with plausible estimates of the missing values. After the imputation phase, conventional statistical methods can be used on each complete dataset. This makes it easier to apply subsequent statistical analyses even when a user does not have profound knowledge about the handling of missing values. In addition, the imputation model may differ from subsequent analysis models. Typically, the imputation model includes many variables of the data set, whereas the analysis model includes a subset of these variables. In contrast, FIML generates the population estimates based only on the variables of interest from the analysis model. However, including some additional variables relevant to missing data to improve the estimation of the missing values is not an inherent advantage of multiple imputation, because these additional variables can be also included in the maximum likelihood model (Graham, 2003). If the imputation model includes variables that are not part of maximum likelihood analysis, then the two approaches can yield different estimates. The decision of which approach to use should depend on the user's knowledge and experience in dealing with missing values.

Some limitations of our study need to be considered. We investigated the accuracy of the estimates for one total sample size. As is known from previous studies of Thorndike's correction formulas (Dunbar & Linn, 1991), the sample size of the selected sample, which results from the total sample size in combination with the selection ratio, affects the precision of the population correlation estimate. Therefore, one important research question is how small the total sample size as well as the size of the selected sample can be while still allowing for unbiased and precise corrections for direct and indirect range restrictions. In our simulation study, we assumed that the variables are multivariate normally distributed, which is routinely the assumption in social and behavioral sciences (Enders, 2010). Multiple imputation assumes multivariate normality, but this missing data approach can provide valid estimates even when this assumption is violated (Demirtas, Freels, & Yucel, 2008). However, this assumption is robust for a large sample size and a low percentage of missing values. Further studies should investigate violations of the assumption of normality (e.g., skewness) in combination with the total sample size.

In summary, this simulation study shows that multiple imputation by chained equations and full information maximum likelihood estimation are accurate approaches correcting for DRR and IRR scenarios. Therefore, both approaches seem to be promising alternatives to Thorndike's correction formulas, especially in the case of indirect range restriction scenarios.

References

Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, 27(2), 187–189.

Ayyub, B. M., & McCuen, R. H. (2011). Probability, statistics, and reliability for engineers and scientists. Boca Raton, FL: CRC press.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. Journal of Applied Psychology, 68(4), 584–589. http://doi.org/10.1037/0021-9010.68.4.584

Chan, W., & Chan, D. W.-L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods*, 9(3), 369–385. http://doi.org/10.1037/1082-989X.9.3.369

Chernyshenko, O. S., & Ones, D. S. (1999). How selective are psychology graduate programs? The effect of the selection ratio on GRE score validity. *Educational and Psychological Measurement*, 59(6), 951–961. http://doi.org/10.1177/00131649921970279

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1), 69–84. http://doi.org/10.1080/10629360600903866

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Duan, B., & Dunlap, W. P. (1997). The accuracy of different methods for estimating the standard error of correlations corrected for range restriction. *Educational and Psychological Measurement*, 57(2), 254–265. http://doi.org/10.1177/0013164497057002005

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments in the prediction of military job performance. In Performance Assessment for the Workplace, II (pp. 127–157). Washington, DC: National Academy Press.

Enders, C. K. (2010). Applied missing data analysis. New York, NY: Guilford Press.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3), 206–213. http://doi.org/10.1007/s11121-007-0070-9

Gross, K., & Bates, D. (2012). mvnmle: ML estimation for multivariate normal data with missing values. R package version 0.1–11. Retrieved from http://cran.r-project.org/package=mvnmle

Holmes, D. J. (1990). The robustness of the usual correction for restriction in range due to explicit selection. *Psychometrika*, 55(1), 19–32. http://doi.org/10.1007/BF02294740

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, 91(3), 594–612. http://doi.org/10.1037/0021-9010.91.3.594

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-Data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469), 332–346. http://doi.org/10.1198/016214504000001844

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, Section: A Mathematics*, 62(A), 28–30. http://dx.doi.org/10.1017/S0080454100006385

Li, J. C., Chan, W., & Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology*, 64(3), 367–387. http://doi.org/10.1348/2044-8317.002007

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, 20(1), 1–15. http://doi.org/10.1111/j.1745-3984.1983.tb00185.x

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology*, 66(6), 655–663. http://dx.doi.org/10.1037/0021-9010.66.6.655

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, NJ: John Wiley & Sons.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2nd ed.). CRC Press.

Medical University of Vienna. (2015). MedAT - Aufnahmeverfahren Medizin [MedAT - admission test medicine]. Retrieved April 15, 2015, from http://www.medizinstudieren.at/

Mendoza, J. L. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika*, 58(4), 601–615.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2), 267–278.

Murphy, K. R. (2003). Validity generalization: A critical review. Taylor & Francis.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Royal Society of London. Retrieved from http://archive.org/details/philtrans02398796

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PLoS ONE*, 21, e0152330. http://dx.plos.org/10.1371/journal.pone.0152330

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85–96.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, 27(1), 52–71. http://doi.org/10.1177/0146621602239476

R Core Team. (2014). A language and environment for statistical computing (Version 3.1.2) [64-bit]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology*, 79(2), 298–301. http://doi.org/10.1037/0021-9010.79.2.298

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. http://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489. http://doi.org/10.1080/01621459.1996.10476908

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). New York: John Wiley & Sons.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1), 112–118. http://doi.org/10.1037/0021-9010.85.1.112

Schafer, J. L. (1997). Analysis of incomplete multivariate data (1st ed.). New York: Chapman and Hall/CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. http://doi.org/10.1037/1082-989X.7.2.147

Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission test scores. *Educational and Psychological Measurement*, 66(2), 305–317. http://doi.org/10.1177/0013164405282455

SjöBerg, S., SjöBerg, A., Näswall, K., & Sverke, M. (2012). Using individual differences to predict job performance: Correcting for direct and indirect restriction of range. Scandinavian *Journal of Psychology*, 53(4), 368–373. http://doi.org/10.1111/j.1467-9450.2012.00956.x

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540. http://doi.org/10.2307/2289457

Thorndike, R. L. (1949). Personnel selection: Test and measurement techniques. New York: Wiley.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219–242. http://doi.org/10.1177/0962280206074463

Van Buuren, S. (2012). Flexible Imputation of Missing Data. CRC Press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3). Retrieved from http://doc.utwente.nl/78938/

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, 14, 1–9.

Appendix

The following example illustrates the steps for estimating the predictive validity with full information maximum likelihood estimation (FIML) and multiple imputation by chained equations (MICE) using the R packages mvnmle and mice. We designed a small dataset (*N* = 50) to mimic a student selection scenario in which prospective students completed an aptitude test and an interview. The criterion measure is an achievement score after two semesters (e.g. average of grades). The college admitted those students who scored at least 100 in the aptitude test. The new interview was presented to the prospective students, but was not used for selection. After the two semesters, the college wants to evaluate the predictive validity of both selection methods. Thus, we have a direct range restriction scenario on the test scores and an indirect range restriction scenario on the interview scores. We assume that this sample is drawn from a multivariate normal distribution.

Without any correction, we observe a Pearson correlation coefficient between test scores and achievement scores of *r* = .28, and between interview scores and achievement scores of *r* = .34. We know that these correlations are biased. Next, we present the steps that need to be taken in R Statistics to estimate the unbiased population correlation with FIML and with MICE. After installing the R packages `mvnmle` (https://cran.r-project.org/web/packages/mvnmle/index.html) and `mice` (https://cran.r-project.org/web/packages/mice/index.html) from the Comprehensive R Archive Network (CRAN), load the packages:

```
R> library(mvnmle)
R> library(mice)
```

The data frame dataset contains three variables: `test` (aptitude test scores), `interview` (interview scores), and `achievement` (criterion scores). Missing values are labeled as NA.

```
R> dataset <- data.frame(
R> "test"=c(99,109,104,104,98,77,96,107,90,…),
R> "interview"=c(19,19,13,18,14,13,16,12,11,…),
R> "achievement"=c(NA,4.0,2.7,3.1,NA,NA,NA,2.4,NA,…))

R> dataset

      test   interview   achievement
1       99          19            NA
2      109          19           4.0
3      104          13           2.7
4      104          18           3.1
5       98          14            NA
6       77          13            NA
   …
```

The number of the missing values can be counted and visualized with the `md.pattern`() function of the `mice` package as follows:

```
R> md.pattern(dataset)

      test   interview   achievement
25       1           1             1    0
25       1           1             0    1
         0           0            25   25
```

There are 25 (out of 50) rows that are complete (last column), and all missing values are in the variable achievement. Estimating the correlation matrix of the dataset using FIML can be done with a call to `mlest()` and by converting the estimated covariance matrix in the correlation matrix as follows:

```
R> fiml <- mlest(dataset)
R> cov2cor(FIML$sigmahat)

            [,1]      [,2]      [,3]
[1,]  1.0000000 0.2557806 0.5097006
[2,]  0.2557806 1.0000000 0.4315233
[3,]  0.5097006 0.4315233 1.0000000
```

The symmetric correlations matrix shows correlations between test and achievement [1,3] = .51, between interview and achievement [2,3] = .43, and between test and interview [1,2] = .26. Creating complete datasets with MICE can be done with a call to `mice()` as follows:

```
miceimp <- mice(dataset, meth=c(”norm”,”norm”,”norm”), m = 20, seed = 6000)
```

where the multiple imputed dataset is stored in the object `miceimp` of class `mids`. Imputations are generated according to the method “norm” (normal distribution), which is specified for each column. The number of multiple imputations is equal to *m* = 20. Note that we used a fixed seed value in this example, so that the exact values can be reproduced. The `complete()` function extracts the 20 complete datasets of the `miceimp` object. Next, we calculate the correlation matrix for each of the complete datasets using the `cor()` function. The pooled correlation matrix is the arithmetic mean of the 20 correlation matrices. Van Buuren (2012) suggests a Fisher-*z* transformation when pooling correlation coefficients (for transforming and re-transforming the correlation matrix, we used the functions `fisherz()` and `fiherz2r()` from the `psych` package).

```
R> for(k in 1:20){
R>      corMatrix = corMatrix + fisherz(cor(complete(miceimp,k)))
R> }
R> fisherz2r(corMatrix/20)

            test      interview   achievement
test        NaN       0.2557759   0.4995916
interview   0.2557759 NaN         0.4343833
achievement 0.4995916 0.4343833   NaN
```

The correlation matrix shows a correlation estimate between test and achievement of .50, and between interview and achievement of .43. Table A1 summarizes the uncorrected and corrected correlations. Subsequently, you will find the final R script for this example including all data for copy and paste.

Table A1. Correlations of the student selection data.

	$\hat{\rho}_{\text{test, achievement}}$	$\hat{\rho}_{\text{interview, achievement}}$
uncorrected	.28	.34
FIML	.51	.43
MICE	.50	.43

```
# run
# Load packages
library(mvnmle)
library(mice)
library(psych)

# Dataset
dataset <- 
data.frame(“test”=c(99,109,104,104,98,77,96,107,90,107,120,98,92,118,101,81,10
0,109,106,103,101,97,119,95,98,107,110,90,107,108,93,110,99,100,106,89,91,98,1
11,84,111,115,92,95,76,102,96,98,98,86),
“interview”=c(19,19,13,18,14,13,16,12,11,16,15,12,16,13,16,14,20,18,16,20,20,2
0,19,20,19,16,17,18,16,18,18,19,11,13,13,10,15,14,15,19,16,20,14,13,14,13,17,1
6,16,12),
“achievement”=c(NA,4.0,2.7,3.1,NA,NA,NA,2.4,NA,3.9,3.3,NA,NA,4.0,2.6,NA,4.0,3.
8,2.8,3.5,2.5,NA,3.5,NA,NA,2.0,4.0,NA,3.7,4.0,NA,4.0,NA,4.0,3.0,NA,NA,NA,2.9,N
A,3.5,4.0,NA,NA,NA,3.1,NA,NA,NA,NA))

dataset # Print dataset

# Show missing data pattern
md.pattern(dataset)

# Correlation matrix without correction (biased estimates)
cor(na.omit(dataset))

# Full information maximum likelihood (FIML)
fiml <- mlest(dataset)
cov2cor(FIML$sigmahat)

# Multiple imputations by chained equations (MICE)
miceimp <- mice(dataset, meth=c(”norm”,”norm”,”norm”), m = 20, seed = 6000)
for(k in 1:20){
   corMatrix = corMatrix + fisherz(cor(complete(miceimp,k)))
}
fisherz2r(corMatrix/20)
```

Citation:

Pfaffel, Andreas, Schober, Barbara, & Spiel, Christiane. (2016). A Comparison of Three Approaches to Correct for Direct and Indirect Range Restrictions: A Simulation Study. *Practical Assessment, Research & Evaluation*, 21(6). Available online: http://pareonline.net/getvn.asp?v=21&n=6

Corresponding Author

Andreas Pfaffel
Faculty of Psychology
Department of Applied Psychology: Work, Education, Economy
University of Vienna, Austria
Universitaetsstrasse 7
1010 Vienna, Austria

andreas.pfaffel [at] univie.ac.at