universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Predictive Validity of Selection Procedures. Correcting for Range Restriction with a Dichotomous Criterion“

verfasst von / submitted by

Ing. Mag. rer. nat. Andreas Pfaffel

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Sozialwissenschaften (Dr. phil.)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 784 298
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Psychologie / Psychology
Betreut von / Supervisor:	Univ.-Prof. Dr. Dr. Christiane Spiel

I thank all my friends and colleagues who believed in me and supported me in carrying out my doctoral thesis. Especially I thank Prof. Christiane Spiel for her extensive support, for the many insightful discussions we had together, and for her valuable advices on the improvement of the quality of my research articles.

Table of contents

List of publications

1. Pfaffel, A., Schober, B., & Spiel, C. (2016). A comparison of three approaches to correct for direct and indirect rang restrictions: A simulation study. *Practical Assessment, Research & Evaluation, 21*(6), 1-15. Available online: http://pareonline.net/getvn.asp?v=21&n=6

2. Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect rang restrictions with a dichotomous criterion: A simulation study. *PLoS ONE, 11*(3), e0152330. Available online: http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0152330

3. Pfaffel, A. & Spiel, C. (2016). *Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study*. Manuscript submitted for publication.

1. Theoretical background

Assessing the predictive validity of selection methods

An effective selection of suitable applicants is of paramount importance for companies and universities in today's competitive job market and in higher education. The goal of a selection method (e.g., aptitude test, assessment center, or interview) or an entire selection procedure is to identify the most suitable applicants regarding a previously defined criterion of success such as job or academic performance, or graduation status. Effective selection methods help to reduce the degree of error in making hiring or admission decisions, and thus represent a win-win situation for the organization and for the applicants. An effective personnel selection can lead to higher employee performance, increased productivity, increased monetary value of output, and increased learning of job-related skills (Hunter, Schmidt, & Judiesch, 1990; Schmidt & Hunter, 1998; Schmidt, Hunter, McKenzie, & Muldrow, 1979). In higher education, it is well documented that the selection of students based on admission tests such as the SAT (former Scholastic Aptitude Test) has positive effects on academic achievement and on retention (Burton & Ramist, 2001; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Mattern & Patterson, 2009), and thus substantiate its use for university admission. In contrast, selection errors may have organization-wide consequences such as poor work quality, low academic achievement, or high dropout rates, and imply a substantial loss of time for both applicants and organizations.

With a view to the practical and economic value, the predictive validity is the most important property of a selection method (Brogden, 1949; Schmidt & Hunter, 1998) as it measures a selection method's effectiveness. The predictive validity is defined as the correlation between test scores collected at one point in time and criterion scores obtained at a later time (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The stronger the correlation between predictor and criterion, the better the prediction of the criterion values, and therefore the higher the predictive validity. As jobs and university places are limited, and specific demands are made on the applicants, companies and universities typically have to make selections and to decide, which selection method should be used. Therefore, companies and universities need to assess the predictive validity of selection methods.

A common methodological problem in the evaluation of the predictive validity of a selection method is the so-called *range restriction problem*, which is an inherent effect of the

selection. Due to the selection, values of the criterion variable are only available for selected applicants. Thus, the correlation coefficient between a predictor *X* and a criterion *Y* obtained from the selected sample is a biased estimate of the correlation between *X* and *Y* in the applicant population, and therefore a biased estimate of the predictive validity of a selection method. The range restriction problem arises because the selection process results in an observed sample (the selected sample), which is not random and therefore not representative of the population of interest (Sackett & Yang, 2000). Typically the sample correlation coefficient obtained from the selected sample underestimates the true population correlation (Levin, 1972; Linn, Harnisch, & Dunbar, 1981). The effect of range restriction on correlation coefficients is well documented in the psychometric literature (Alexander, Barrett, Alliger, & Carson, 1986; Duan & Dunlap, 1997; Greener & Osburn, 1979; Mendoza, Hart, & Powell, 1991; Thorndike, 1949).

The two most common range restriction scenarios in personnel selection and higher education are the direct and the indirect range restriction scenarios (Hunter, Schmidt, & Le, 2006). In a direct or explicit range restriction scenario, the selection of applicants is based directly on the predictor *X* top down (e.g., on test scores), with *X* being either a score from a single selection method or a composite score derived from several selection methods. In a direct range restriction scenario, the predictor *X* is restricted in range by complete truncation below a specific cutoff in *X* – therefore, it is called range restriction problem. In contrast, an indirect range restriction scenario occurs, if applicants have been selected on a third variable *Z*, which is usually correlated with *X*, *Y*, or both. Although selection is based on *Z*, the correlation coefficient between *X* and *Y* remains of interest. The variable *Z* can be based on a single selection method or on a combination of selection methods, possibly but not necessarily including *X* (Linn et al., 1981). Organizations often use a composite score (e.g., based on an aptitude test and a quantitative interview) for the selection of applicants, but need to know the predictive validity of each selection method, to be able to decide about removing or weighting a particular selection method. The degree to which correlation coefficients are affected is typically weaker in an indirect range restriction scenarios than in a direct range restriction scenario (Sackett & Yang, 2000). However, the estimate of the population correlation is biased in both scenarios and has to be corrected to provide a more valid estimate of the predictive validity.

Correcting for range restriction with a continuous criterion

Correcting correlation coefficients for range restriction is one of the oldest problems in statistics and researchers have studied this problem for more than hundred years. At the beginning of the 20th century, the biostatistician Pearson (1903) noticed that his product-moment correlation coefficient is biased in the case of natural selection and developed correction formulas to overcome this range restriction problem. The formulas have been developed only for continuous variables under the assumptions of linearity between *X* and *Y* and multivariate normality. Based on Pearson's work, Aitken (1935) and Lawley (1943) developed a generalized correction formula and provided a proof for the assumptions of linearity and homoscedasticity (same probability distribution of the error term) in the selected sample and in the population. Correcting for range restriction has received a considerable amount of attention as Thorndike (1949) classified range restriction problems into direct and indirect range restriction scenarios and presented correction formulas for both scenarios. One reason for Thorndike's impact in research and practice might be that his correction formulas were easy to apply in contrast to Lawley's generalized solution. Thorndike's formulas simply use the ratio of the standard deviation obtained from the unrestricted (applicant) dataset to the standard deviation obtained from the dataset of the selected sample. Lawley's (1943) generalized formula leads to the problem of estimating the variance-covariance matrix of incomplete multivariate data, which was not possible at that time (today we can handle this problem with full information maximum likelihood estimation, see Pfaffel, Schober, & Spiel, 2016).

Thorndike's formulas are widely studied and it is well-known in the psychometric literature that corrected correlation coefficients are less biased than uncorrected correlation coefficients, even over a wide range of assumption violations (Greener & Osburn, 1979; Gross & Fleischman, 1983; Holmes, 1990; Linn, 1983; Linn et al., 1981; Ree, Carretta, Earles, & Albert, 1994; Rydberg, 1963; Zimmerman & Williams, 2000). Although corrected correlation coefficients tend to be more accurate than uncorrected correlation coefficients, some researchers questioned the accuracy of the correction under extreme conditions, especially under conditions of extreme selection (Greener & Osburn, 1979; Lord & Novick, 1968; Novick & Thayer, 1969). Not surprisingly, the selection ratio, which is the number of selected applicants divided by the number of all applicants, has been found to be the most important factor affecting the accuracy of the corrected correlation coefficients.

Since the beginning of the 1980s, the application of formulas to correct for range restriction has more and more become professional practice (see the "Principles for the Validation and Use of Personnel Selection Procedures", Society for Industrial and Organizational Psychology & American Psychological Association, 1980/2003). Thorndike's formulas have often been applied in predictive validity studies in different fields; for example, in large-scale testing programs in higher education such as the SAT (Kobrin et al., 2008; Weitzman, 2005), the Graduate Record Examination (GRE) (Chernyshenko & Ones, 1999; Powers, 2004), and the Graduate Management Admission Test (GMAT) (Sireci & Talento-Miller, 2006). The formulas have also been applied in other fields, for example, in predicting job performance in personnel selection (Salgado, 2003; SjöBerg, SjöBerg, Näswall, & Sverke, 2012; Thorndike, 1949), in assessing the correlation between scores from a theoretical and a practical driving-license test (Wiberg & Sundström, 2009), and in assessing the predictive validity of selection procedures for pilot candidates in the U.S. Air Force (Ree et al., 1994). Correcting correlation coefficients is also an important issue in meta-analytic studies (Hunter et al., 2006; Murphy, 2003).

Beside the accuracy of the correction, considerable attention has been given to the estimation of the standard error and the confidence interval of the corrected correlation coefficient, because solely reporting a point estimate was no longer regarded as sufficient. In statistics, the standard error is a common measure of the variability of the sampling distribution of a population parameter estimate. In 1974, the APA 'Standards for Educational and Psychological Tests' required that 'full information' should be presented where correlation coefficients are corrected for range restriction (American Psychological Association, 1974). Thus, reporting the standard error constitutes an important step towards full information (Bobko & Rieck, 1980). Monte Carlo studies of the effectiveness of Thorndike's formulas showed that applying these formulas increases the variability of the population correlation estimate, i.e. the standard error of the corrected correlation coefficient is larger than the one of the uncorrected correlation coefficient (Forsyth, 1971). Bobko and Rieck (1980) derived a formula for calculating the standard error of the corrected correlation coefficient for direct range restriction scenarios under the assumption of linearity and bivariate normality. Different approaches to estimate the standard error of the corrected correlation coefficient have been proposed and investigated; for example, standard error for complete cases under range restriction (Millsap, 1989), Fisher's z-transformation (Mendoza, 1993), and bootstrapping (Chan & Chan, 2004; Li, Chan, & Cui, 2011).

In addition, scientists presented corrections for situations in which the assumptions for applying Thorndike's formulas are violated or not all necessary data are available (for an expanded typology of range restriction scenarios, see Sackett & Yang, 2000). All these approaches have in common that the trustworthiness of the correction decreases the less information is available. For example, Thorndike's formulas require the unrestricted variance to be known. For a direct range restriction scenario, Schmidt, Hunter, and Urry (1976) presented a function, which describes the relationship between the selection ratio and the ratio of the restricted to the unrestricted standard deviation for large samples. Another approach was to provide tables in which the observed cutoff for truncation (which is related to Cohen's ratio, Cohen, 1959) is entered and the proportional reduction in variance is obtained (Alexander, Alliger, & Hanges, 1984; Dobson, 1988). These tables can also be applied for an indirect range restriction scenario if Z is measured for the selected sample but the unrestricted variance is unknown (Sackett & Yang, 2000). The most problematic case in terms of estimating population parameters under selection arises when the selection variable is unknown or not measured, which is known as sample selection bias or non-ignorable selection (Gross & McGanney, 1987; Rubin, 1976). For example, this is the case when selection is based on unquantified subjective judgements (Hunter et al., 2006), or in self-selected groups (Heckman, 1976, 1979; Linn, 1968). This problem, however, is not addressed in this doctoral thesis.

Correcting for Range Restriction with a Dichotomous Criterion

Dichotomous criterion variables can be found in a variety of fields of sociological, medical, and psychological research (Cleary & Angel, 1984; Peng, Lee, & Ingersoll, 2002). For example, in higher education the graduation status is dichotomous, graduated versus non-graduated. In the U.S., graduation rates are very popular in evaluating institutional performance and accountability of colleges and universities (Gold & Albert, 2006; Zhang, 2008). In 2002, 41 U.S. states used "graduation rate data for state- and/or system-level accountability and performance reporting purposes", and 7 states established this indicator in performance funding systems (American Association of State Colleges and Universities, 2002, p. 2). Therefore, valid predictions of the graduation status (or dropout) are important for universities to support their decisions on the use of admission tests. Graduation versus non-graduation is a dichotomous criterion, which can be used for almost any course of training, whether academic, industrial, or military (Thorndike, 1949). Consequently, the idea behind applying a selection method is that

applicants with a higher score in *X* will complete their studies with a higher probability than those with a lower score.

A dichotomous variable is characterized by a division of the individuals of a sample or population into two distinct groups. Two kinds of dichotomous variables can be distinguished (Ulrich & Wirtz, 2004): naturally and artificially dichotomous variable. A dichotomous variable is labeled as natural when the division is based on a qualitative characteristic, e.g. 'graduated' versus 'non-graduated'. In contrast, an artificially dichotomous variable has a continuous underlying scale, which has been divided into two groups; for example, when student performance is measured on a continuous scale, but students are divided into 'low performers' and 'high performers' based on their performance score. Assuming the predictor variable is measured on a continuous scale and the criterion variable is dichotomous, two distinct correlation coefficients are available as measures of the predictive validity: In the case of a naturally dichotomous variable, the point-biserial correlation coefficient (Tate, 1954) is an appropriate measure describing the strength of the relationship between predictor and criterion. In the case of an artificially dichotomous variable, the appropriate measure is the biserial correlation coefficient, because the point-biserial correlation coefficient systematically underestimates the Pearson correlation coefficient, which would have been obtained before dichotomization (Ulrich & Wirtz, 2004). For example, the upper limit of the point-biserial correlation coefficient between a normally distributed variable and a dichotomous one is .798 when the arithmetic average of the normal variable was used as cutoff point (Gradstein, 1986).

Although the range restriction problem has been studied for a long time, few researchers have considered dichotomous criterion variables. One important factor to be considered when dealing with a dichotomous criterion variable is the base rate of success, which is calculated by dividing the number of successful individuals by the number of applicants (Taylor & Russell, 1939). The base rate of success represents the proportion of individuals who would be successful on the criterion if there were no selection, i.e. hundred percent of the applicants were selected. Unfortunately, the base rate of success is unknown in case of selection, as we cannot obtain the proportion of applicants able to succeed on the criterion on the basis of the selected sample. Similar to the correlation coefficient, the success rate obtained from the selected sample is a biased estimate of the base rate of success. The success rate is the proportion of successful individuals among those selected. Consequently, when correcting for range restriction with a dichotomous criterion, both the correlation coefficient as well as the base rate of success have to be considered.

The few suggested approaches to correcting correlation coefficients for range restriction are often inapplicable because they require the base rate of success to be known (e.g., from the literature) or assumed (Abrahams, Alf, & Wolfe, 1971; Bobko, Roth, & Bobko, 2001; Taylor & Russell, 1939). Taylor and Russell (1939) developed tables for an artificially dichotomous criterion variable under the assumptions of linearity between predictor and criterion, and bivariate normality. The tables show the relation between four important factors: the true Pearson correlation coefficient, the selection ratio, the base rate of success, and the success rate. For example, the tables can be used to assess how changing the selection ratio affects the change of the success rate. In order to use these tables, three of these four factors must be known. Indeed, the correlation coefficient as well as the base rate of success are typically unknown. In such cases, Taylor and Russell recommended that a value of the base rate of success has to be assumed. However, assuming a value is an arbitrary approach as different values of the base rate of success result in different values of the correlation coefficient, i.e. in different assessments of the predictive validity. Abrahams and colleagues (1971) presented Taylor-Russell tables for a naturally dichotomous criterion variable in which the validity coefficient is the point-biserial correlation coefficient. However, the problem remains. So far, no method has been proposed to estimate the base rate of success based on the selected sample.

Bobko and colleagues (2001) focused on the correction of the effect size Cohen's *d* and converted *d* in *r* and converted *r* back in *d* for direct and indirect range restriction scenarios. However, the final formulas (p. 53 and p. 55) require a value for the base rate of success. A different approach presented by Raju, Steinhaus, Edwards, and DeLessio (1991) focuses on the odds ratio rather than on correlations. The researchers proposed applying a logistic regression model based on data of the selected sample. Applying a logistic regression is a valuable approach, but the proposed method has two significant disadvantages: On the on hand, it does not focus on correcting the correlation coefficient. On the other hand, applying a single regression model in case of missing data is not state-of-the-art from a modern methodological perspective because the regression model does not incorporate the uncertainty of the model parameters (Enders, 2010; van Buuren, 2012). Thorndike (1949) already discussed settings in which the criterion measure is artificially or naturally dichotomous. He presented no solution for this problem, but discussed the effects of the base rate of success, more precisely the unequal division of the two groups, on the accuracy of the correction formulas. So far, the accuracy of Thorndike's (1949) correction formulas in the case of a dichotomous criterion variable have never been investigated empirically. Hence, the scientific literature does not provide any

statistical method to correct correlation coefficients for range restriction in settings in which the base rate of success is unknown.

2. Goals of the doctoral thesis

As shown in the previous section, assessing the predictive validity of selection methods is problematic due to the so-called range restriction problem. So far, approaches to correcting correlation coefficients for range restriction have been widely studied for continuous criterion variables but not for dichotomous ones. The few suggested approaches to overcoming the range restriction problem in the case of dichotomous criterion variables are often inapplicable because they require a value for the base rate of success, which is typically unknown. Hence, there is a deficiency of scientific research on suitable correction methods to overcome the range restriction problem when the criterion variable is dichotomous. To overcome the range restriction problem with a focus on dichotomous criterion variables, this doctoral thesis pursues the following goals:

- The first goal is to question, whether the existing approaches to correcting correlation coefficients for range restriction are state-of-the-art from a modern methodological and statistical perspective.

- The second goal is to propose a state-of-the-art approach to correct correlation coefficients for direct and indirect range restriction scenarios with a dichotomous criterion variable.

- The third goal is to investigate the accuracy of the proposed approach for a number of conditions, which have an effect on the accuracy of the population estimates.

3. Synopsis of publications

1st Publication

Pfaffel, A., Schober, B., & Spiel, C. (2016). A comparison of three approaches to correct for direct and indirect rang restrictions: A simulation study. *Practical Assessment, Research & Evaluation, 21*(6), 1-15. Available online: http://pareonline.net/getvn.asp?v=21&n=6

The first paper of the present doctoral thesis (Pfaffel, Schober, & Spiel, 2016) shows that range restriction scenarios can be viewed as a missing data mechanism. The paper shows that the loss of criterion data in direct and indirect range restriction scenarios is missing at random regarding Rubin's (1976) theoretical classification scheme for missing data problems. Missing at random means the probability of missing data on a criterion variable is related to other measured variables in the analysis model (i.e. the variable, which were used for selection), but not to values of the criterion variable itself (Enders, 2010). In case of missing at random, state-of-the-art missing data techniques such as full information maximum likelihood and multiple imputation can be applied to estimate the predictive validity of a selection method in the case of direct and indirect range restriction scenarios. Although Mendoza (1993) already mentioned this missing data approach as an opportunity to overcome the range restriction problem, no study has been presented that investigates this approach empirically.

The purpose of the first paper was to compare the accuracy of the correction of two state-of the-art missing data techniques (full information maximum likelihood, and multiple imputation by chained equations) with Thorndike's (1949) well known correction formulas for direct and indirect range restriction scenarios. Monte Carlo simulations were conducted to investigate the accuracy of the population correlation estimates in dependence of the selection ratio and the true Pearson correlation coefficient between predictor and criterion in an experimental design.

The results show that the two missing data techniques are effective to estimate the population correlation between predictor and criterion, and therefore suitable to estimate the predictive validity of a selection method. In the case of a direct range restriction scenario, the two missing data techniques are equally accurate as Thorndike's correction formula. In the case of an indirect range restriction scenario, the two missing data techniques are more precise than Thorndike's formula when the selection ratio is small and the true relationship between predictor and criterion is strong (a higher precision means that the random error of the

correlation corrected for range restriction is smaller). In general, the precision of the population correlation estimates decreases as the selection ratio and the population correlation decrease.

In summery, the two state-of-the-art missing data techniques are suitable to estimate the predictive validity of selection methods for direct and indirect range restriction scenarios with the same or higher precision than Thorndike's formulas. Thus, the missing data approach proposed in this study represents an alternative to Thorndike's correction formulas.

2nd Publication

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect rang restrictions with a dichotomous criterion: A simulation study. *PLoS ONE, 11*(3), e0152330. Available online: http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0152330

The second paper of my doctoral thesis (Pfaffel, Kollmayer, Schober, & Spiel, 2016) shows that the proposed missing data approach can also be applied when the criterion variable is artificially or naturally dichotomous. Handling missing values in dichotomous criterion variables is possible with the state-of-the-art missing data technique multiple imputation by chained equations based on a logistic regression model (van Buuren & Groothuis-Oudshoorn, 2011). The procedure is as follows: First, multiple imputation replaces the missing values of the dichotomous the criterion variable and generates several complete datasets. Then, the unbiased correlation coefficient between predictor and criterion as well as the base rate of success can be calculated based on these complete datasets. As clearly depicted above, proposed approaches dealing with a dichotomous criterion require the base rate of success to be known or assumed. In contrast, the proposed missing data approach provides an empirical estimation for both the predictive validity and the base rate of success.

The main purpose of this study was to compare the accuracy of the population correlation estimates of multiple imputation by chained equations with Thorndike's (1949) formulas for direct and indirect range restriction scenarios when the criterion variable is artificially or naturally dichotomous. Another purpose was to investigate the accuracy of the estimate of the base rate of success. The population correlation between predictor and criterion, the base rate of success, and the selection ratio were systematically varied over a wide range in order to investigate the accuracy of the estimates for a variety of possible datasets.

The comparison of the two approaches show that the accuracy of the population correlation estimates is higher for the missing data approach than for Thorndike's formulas for direct and indirect range restriction scenarios and for both kinds of dichotomous criterion variables. In addition, the missing data approach provides an accurate estimate of the base rate of success. The accuracy of all population estimates decrease as the selection ratio decreases. For selection ratios smaller than 30%, the estimates of the population correlation and the estimate of the base rate of success become less precise and tend to be biased.

In summary, the proposed missing data approach first time provides an empirical estimation for both the predictive validity of a selection method and the base rate of success when the criterion variable is artificially or naturally dichotomous.

3rd Publication

Pfaffel, A. & Spiel, C. (2016). *Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study*. Manuscript submitted for publication.

The third paper of my doctoral thesis (Pfaffel & Spiel, submitted) investigates the accuracy of population correlation estimates using multiple imputation by chained equations for a continuous and a naturally dichotomous criterion when the number of applicants is small or moderate. The proposed missing data approach to correcting correlations for range restriction has thus far only been investigated with relatively large samples (Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, et al., 2016). Therefore, it is questionable whether missing data techniques such as multiple imputation by chained equations are able to correct correlations for range restriction when the sample size is small or moderate. Additionally, no empirical study has investigated so far the accuracy of the multiple imputation standard error of the population correlation estimates in the case of direct and indirect range restriction scenarios. The standard error is a measure of the sampling deviation of the population correlation estimate. The findings of this study provide empirical evidence of the accuracy of the correction and support researchers and evaluators in their assessment of conditions (i.e. combinations of sample size, selection ratio, kind of criterion variable, etc.) under which corrected correlation coefficients can be trusted.

The purpose of this study was to investigate the accuracy of the population correlation estimates and their associated multiple imputation standard errors for direct and indirect range

restriction scenarios in terms of a continuous and a naturally dichotomous criterion variable, small and moderate sample sizes, population correlation, selection ratio, and base rate of success.

The results show that in the case of a continuous criterion variable the estimate of the population correlation is negatively biased for both direct and indirect range restriction scenarios, whereby the bias increases as the sample size and the selection ratio decrease, and the population correlation increases. This bias is known from simulation studies of the accuracy of Thorndike's formulas (Chan & Chan, 2004), which means the underestimation of the population correlation due to range restriction cannot be fully corrected in either approach. The multiple imputation standard error underestimates the empirical sampling deviation of the sample correlation coefficient, especially in a direct range restriction scenario when the sample size and the selection ratio are small.

In the case of a naturally dichotomous criterion variable, the correction could not be applied for a large number of selected samples because the criterion variable was constant. This was often the case, when the sample size and the selection ratio were small, and the base rate of success was high. The estimate of the population correlation is strongly biased for both direct and indirect range restriction scenarios. The estimates are negatively biased for base rates of success of 20% and 50%, but positively biased for a base rate of success of 80%. The results suggests a non-linear relationship between the base rate of success and accuracy of the population estimates. In contrast to the population correlation estimates, the multiple imputation standard error is accurate for a variety of conditions for direct and indirect range restriction scenarios.

In summary, a small or moderate number of applicants yield in biased estimates of the predictive validity of the selection method, especially in a direct range restriction scenario with a small selection ratio. In the case of a naturally dichotomous variable, the estimate of the predictive validity is biased for a variety of conditions when the number of applicants is small. The findings of this study provide evidence about conditions under which estimates of the predictive validity of selection methods can be trusted.

4. Summary and discussion

Selection methods in personnel selection or in selection in higher education are usually applied to identify the most suitable applicants. With a view to the practical and economic value, the predictive validity is the most important property of a selection method because it is a useful measure for it's effectiveness. Therefore, companies and universities need to assess the predictive validity of their selection methods. However, assessing the predictive validity of a selection method is related to the range restriction problem, which causes biased population estimates of the predictive validity. This bias has to be corrected to provide more valid estimates of the predictive validity. To overcome the range restriction problem, researchers have proposed formulas to correct the biased estimates of the predictive validity. In the long history of investigating the range restriction problem, researchers focused almost exclusively on settings in which the predictor variable and the criterion variable are continuous. However, dichotomous criterion variables can be often found in a variety of fields of sociological, medical, and psychological research and practice. The few existing approaches dealing with a dichotomous criterion variable are often inapplicable because they require information about the base rate of success. However, this information is typically not available. The psychometric literature so far does not provide any method to estimate the predictive validity of a selection method for situations in which the base rate of success is unknown. Hence, there is a lack of scientific research on suitable methods able to correct for range restriction in the case of a continuous predictor and a dichotomous criterion.

Therefore, the aim of the present doctoral thesis was to overcome this deficiency by providing an approach, which is able to handle direct and indirect range restriction scenarios with a dichotomous criterion. In order to achieve this, the proposed approach is to view range restriction scenarios as missing data mechanisms (Mendoza, 1993; Pfaffel, Schober, et al., 2016). This offers some advantages, for example, that we can draw on modern and methodologically well-founded techniques in handling missing data (Enders, 2010). Both direct and indirect range restriction scenarios can be considered as missing at random (Pfaffel, Schober, et al., 2016). Consequently, the same missing data techniques such as full information maximum likelihood and multiple imputation can be applied to both scenarios in order to estimate the relationship between predictor and criterion of the applicant population. However, the first goal of this doctoral thesis was to show that the proposed missing data approach is appropriate to overcome the range restriction problem with continuous criterion variables, and

to compare the accuracy of the population estimates with Thorndike's well-known correction formulas. The results show the two missing data techniques are equally or more accurate than Thorndike's (1949) formulas and effective to estimate the predictive validity of a selection method.

The second paper (Pfaffel, Kollmayer, et al., 2016) shows that the missing data approach is also able to estimate the predictive validity of selection methods in the case of direct an indirect range restriction scenarios with an artificially or a naturally dichotomous criterion. In addition, the missing data approach allows an empirical estimation of the base rate of success based on the observable selected sample. Moreover, the study examined the accuracy of Thorndike's formulas in the case of an artificially or naturally dichotomous criterion, but the formulas are less precise than the proposed missing data technique multiple imputation by chained equations. Thus for the first time an approach is presented, which is able to estimate the predictive validity of selection methods as well as the base rate of success. For example, the estimate of the base rate of success can then be used for the application of the Taylor-Russell tables (Taylor & Russell, 1939) and the Taylor-Russell tables for a dichotomous criterion (Abrahams et al., 1971).

Approaches to overcoming the range restriction problem have been developed under the framework of large sample theory. However, researcher and evaluators are often faced with a small or moderate number of applicants but must still attempt to estimate the predictive validity. The first and the second paper of this doctoral thesis dealt with relatively large sample sizes. Therefore, the third paper (Pfaffel & Spiel, submitted) investigated the missing data technique multiple imputation by chained equations for small and moderate samples in the case of a continuous and a naturally dichotomous criterion variable. The results show that small samples lead to biased estimates of the predictive validity, especially in the case of a dichotomous criterion variable. In addition, the paper investigated the accuracy the standard error of the population estimate. In contrast to the point estimate, the associated standard error is accurate for a variety of conditions, which have an effect on the accuracy of the population estimates. Although each of the three studies examining the accuracy of the population estimates under various conditions (third goal of the present doctoral thesis), especially the third study (Pfaffel & Spiel, submitted) shows limits of the empirical assessment of the predictive validity of selection methods. The population estimates of the predictive validity are not trustworthy in small and moderate samples in combination with a rigorous selection. Graham and Schafer (1999, p. 26) pointed out "…the limitations of analysis with small sample size lie in the small sample size itself… Multiple imputation cannot compensate for having collected too little

data.”. therefore, the findings in paper 3 (Pfaffel & Spiel, submitted) provide empirical evidence about the accuracy of the population estimates, and thus support researchers and evaluators in assessing the predictive validity of selection methods.

Methodological strengths and limitations

Full information maximum likelihood and multiple imputation under the assumption of missing at random are currently regarded as state-of-the-art techniques in dealing with missing data (Enders, 2010; Schafer & Graham, 2002). State of the art means that both techniques do not suffer from problems (e.g., under- and overestimation of correlation coefficients) of other missing data techniques as, for example, listwise or pairwise deletion, mean imputation, and single regression imputation (Enders, 2010). Although Thorndike’s (1949) correction formulas represent likelihood estimates, the missing data approach has several advantages over the formulas. First, the missing data approach can be applied to all scenarios that meet the assumption of missing at random. Hence, no distinction between direct and indirect range restriction is to be made in applying full information maximum likelihood or multiple imputation. Second, the missing data techniques can handle incomplete multivariate datasets with several covariates. In contrast, Thorndike’s formulas do not consider other variables of the dataset, which can improve the accuracy of the correction. Third, the missing data techniques can handle incomplete multivariate datasets with non-identically distributed variables. This makes it possible to correct correlation coefficients for range restriction and to estimate the base rate of success in the case of a dichotomous criterion variable (this is possible because both techniques provide estimates of the mean vector and the variance-covariance matrix).

The loss of criterion data is an inherent effect of the selection. Consequently, real datasets with complete data are typically not available. Therefore, simulation studies are essential in investigating the accuracy of methods to correct biased population correlation estimates. Monte Carlo simulations have clear advantages over real datasets, as they are real experiments in which influencing factors can be systematically varied and investigated. This is not possible with real datasets because each of them has its certain combination of factors that cannot be modified (e.g., the population correlation between predictor and criterion, or the base rate of success). Therefore, Monte Carlo simulations have been applied for a wide range of factor combinations necessary to provide detailed empirical evidence of the accuracy of the estimates.

Whenever a new correction method is proposed, it is necessary to examine the accuracy of the statistical estimates. Accuracy is defined as the closeness of the value of an estimate to the true value of the parameter being estimated (Ayyub & McCuen, 2011). The concept of accuracy encompasses trueness, which is also know as bias or systematic error, and precision, which is also known as random error. In the psychometric literature, it is well documented that corrected correlation coefficients are less biased than uncorrected correlation coefficients. However, trueness is a necessary but not sufficient property of an accurate estimate. When the precision is low, the variation of the estimate around the true value is high. Therefore, the three papers also focused on the precision of the estimates. This information supports researchers and evaluators in their assessment and interpretation of the predictive validity of selection methods.

However, beside the methodological strengths of the studies there are also methodological limitations. The main assumption, which has been made in these studies, is that the missing mechanism is missing at random. This assumption is valid for direct and indirect range restriction scenarios. When the missing data mechanism is considered missing not at random (Rubin, 1976), the two proposed missing data techniques – as well as Thorndike's correction formulas – cannot be applied without additional assumptions about the probability of missingness (Enders, 2010). For example, this is the case when the selection variable is not measured or the probability of missing values is related to unmeasured self-selection mechanisms (Alexander, 1990; Linn, 1968). Another assumption that have been made in the studies is linearity between predictor and criterion, and between the selection variable and all other variables in the case of indirect range restriction scenarios. If the relationship is non-linear, the proposed techniques lead to inaccurate estimates without additional model specifications. However, multiple imputation is flexible enough to fit non-linear relationships (van Buuren, 2012).

Furthermore, the accuracy of the proposed missing data techniques have been investigated only on minimum information data sets with two or three variables in order to make them comparable with Thorndike's formulas. Therefore, the accuracy of the population estimates are not directly comparable with real datasets with many variables because additional variables in the dataset correlated with the criterion variable will increase the accuracy of the population estimates. In the case of a dichotomous criterion variable, the simulated multivariate data assumed a mixture distribution of the predictor based on two univariate normal distributions for each of the two criterion groups. Thus, the studies are consistent with the literature on assessing the predictive validity of a selection method when the criterion variable is naturally dichotomous (Abrahams et al., 1971). Although many reasons speak in favor of

assuming normally distributed predictor values for the two criterion groups, other distributions are quite conceivable. At present, it is not clear whether and to what extent, skewed distributions affect the accuracy of the population estimates.

Recommendations for further research

As mentioned above, the proposed missing data approach offers some advantages over the correction formulas, for example, handling more than one criterion variable, simultaneous estimation of predictive validity of several predictors (i.e. of several selection methods), including covariates in the analysis model (e.g., sociodemographic variables), and even handling covariates that have some missing values themselves. Therefore, further studies should investigate the accuracy of the population estimates in more complex and therefore more realistic datasets. However, generating multivariate datasets with dependent and non-identical distributed variables is often a big challenge in simulation studies. Additionally, the number of parameters to be specified in simulating multivariate distributions (especially the intercorrelations) increase as the number of variables increase. The difficulty lies mainly in the interpretation of the effect of various inter-correlation matrices, because the correlations between independent variables affect the accuracy of population estimates, especially between predictor and selection variable (Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, et al., 2016).

As clearly shown in the present doctoral thesis, multiple imputation is able to correct for range restriction with a dichotomous criterion variable. However, multiple imputation can also handle, for example, a trichotomous criterion variable. More specifically, multiple imputation can be used to impute ordered categorical (ordinal) and unordered categorical (nominal) criterion variables, which is possible under the broad class of generalized linear models (van Buuren, 2007, 2012). For example, achievement indicators such as grades are typically consist of five or six ordered categories. No study has been presented that investigate the accuracy of range restriction corrections with these kind of criterion variables. Therefore, further studies should investigate the accuracy of the population estimates when the criterion variable is ordered or unordered categorical. Especially, an unequal number of observations in each of the criterion categories could have a large effect on the accuracy of estimates similar as shown for the base rate of success with two criterion groups.

Conclusion

Although researchers have studied the range restriction problem for a long time, little is known of correcting correlation coefficients for range restriction in the case of a dichotomous criterion. The present doctoral thesis proposes a state-of-the-art missing data approach to overcome direct and indirect range restriction scenarios with continuous and dichotomous criterion variables. The findings show that missing data techniques are effective in producing accurate population estimates over a variety of conditions. Thus, the missing data approach provides an effective way to estimate the predictive validity of selection methods. However, the findings also show that some conditions lead to inaccurate population estimates. Such findings are very important because they support researchers and evaluations in their assessment of conditions under which population estimates can be trusted. The missing data approach offers a number of advantages over the traditional correction formulas. In summary, the present doctoral thesis represents an important basis for future research on the range restriction problem and offers researchers and evaluators methodologically well-founded and flexible techniques for the assessment of the predictive validity of selection methods.

5. References

Abrahams, N. M., Alf, E. F., & Wolfe, J. J. (1971). Taylor-Russell tables for dichotomous criterion variables. *Journal of Applied Psychology*, *55*(5), 449–457. http://doi.org/10.1037/h0031761

Aitken, A. C. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, *4*(2), 106–110. http://doi.org/10.1017/S0013091500008063

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

Alexander, R. A., Alliger, G. M., & Hanges, P. J. (1984). Correcting for range restriction when the population variance is unknown. *Applied Psychological Measurement*, *8*(4), 431–437. http://doi.org/10.1177/014662168400800407

Alexander, R. A., Barrett, G. V., Alliger, G. M., & Carson, K. P. (1986). Towards a general model of non-random sampling and the impact on population correlation: Generalizations of Berkson's Fallacy and restriction of range. *British Journal of Mathematical and Statistical Psychology*, *39*(1), 90–105. http://doi.org/10.1111/j.2044-8317.1986.tb00849.x

American Association of State Colleges and Universities. (2002). *Accountability and graduation rates: Seeing the forest and the trees.* Washington, DC: Author. Retrieved from http://eric.ed.gov/?id=ED474375

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

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

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement*, *4*(3), 385–398. http://doi.org/10.1177/014662168000400309

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, *4*(1), 46–61. http://doi.org/10.1177/109442810141003

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, *2*(2), 171–183. http://doi.org/10.1111/j.1744-6570.1949.tb01397.x

Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (No. 2001–2). New York: College Entrance Examination Board. Retrieved from http://research.collegeboard.org/publications/content/2012/05/predicting-success-college-sat-studies-classes-graduating-1980

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

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, 334–348.

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, *1*(3), 217–237. http://doi.org/10.2307/1266442

Dobson, P. (1988). The correction of correlation coefficients for restriction of range when restriction results from the truncation of a normally distributed variable. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 227–234. http://doi.org/10.1111/j.2044-8317.1988.tb00898.x

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

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

Forsyth, R. A. (1971). An empirical note on correlation coefficients corrected for restriction in range. *Educational and Psychological Measurement*, *31*(1), 115–123. http://doi.org/10.1177/001316447103100109

Gold, L., & Albert, L. (2006). Graduation rates as a measure of college accountability. *American Academic*, *2*(1), 89–106.

Gradstein, M. (1986). Maximal correlation between normal and dichotomous variables. *Journal of Educational Statistics*, *11*(4), 259–261. http://doi.org/10.2307/1164698

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: SAGE Publications, Inc.

Greener, J. M., & Osburn, H. G. (1979). An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement*, *3*(1), 31–41. http://doi.org/10.1177/014662167900300104

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement*, *7*(2), 227–237. http://doi.org/10.1177/014662168300700210

Gross, A. L., & McGanney, M. L. (1987). The restriction of range problem and nonignorable selection processes. *Journal of Applied Psychology*, *72*(4), 604–610. http://doi.org/10.1037/0021-9010.72.4.604

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, *47*(1), 153–161. http://doi.org/10.2307/1912352

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., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology*, *75*(1), 28–42. http://doi.org/10.1037/0021-9010.75.1.28

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

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for predicting First-Year College Grade Point Average* (Research Report No. 2008–5). New York: The College Board. Retrieved from https://professionals.collegeboard.com/profdownload/Validity_of_the_SAT_for_Predicting_First_Year_College_Grade_Point_Average.pdf

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://doi.org/10.1017/S0080454100006385

Levin, J. (1972). The occurrence of an increase in correlation by restriction of range. *Psychometrika*, *37*(1), 93–97. http://doi.org/10.1007/BF02291414

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. (1968). Range restriction problem in the use of self-selected groups for test validation. *Psychological Bulletin*, *69*(1), 69–73. http://doi.org/10.1037/h0025263

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://doi.org/http://dx.doi.org/10.1037/0021-9010.66.6.655

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Retrieved from http://doi.apa.org/psycinfo/1968-35040-000

Mattern, K. D., & Patterson, B. F. (2009). *Is performance on the SAT related to college retention?* (Research Report No. 2009–7). New York: The College Board. Retrieved from http://research.collegeboard.org/publications/content/2012/05/performance-sat-related-college-retention

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

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26*(2), 255–269. http://doi.org/10.1207/s15327906mbr2602_4

Millsap, R. E. (1989). Sampling variance in the correlation coefficient under range restriction: A Monte Carlo study. *Journal of Applied Psychology*, *74*(3), 456–461. http://doi.org/10.1037/0021-9010.74.3.456

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

Novick, M. R., & Thayer, D. T. (1969). *An investigation of the accuracy of the Pearson selection formulas* (Reserach Memorandum RM-69-22). Princeton, NJ: Educational Testing Service. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=AD0695411

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

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14. http://doi.org/10.1080/00220670209598786

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*, *11*(3), e0152330. http://doi.org/10.1371/journal.pone.0152330

Pfaffel, A., Schober, B., & Spiel, C. (2016). A comparison of three approaches to correct for direct and indirect range restrictions: A simulation study. *Practical Assessment, Research & Evaluation*, *21*(6), 1–15.

Pfaffel, A., & Spiel, C. (submitted). Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study, 37.

Powers, D. E. (2004). Validity of Graduate Record Examinations (GRE) general test scores for admissions to colleges of veterinary medicine. *Journal of Applied Psychology*, *89*(2), 208. http://doi.org/http://dx.doi.org/10.1037/0021-9010.89.2.208

Raju, N. S., Steinhaus, S. D., Edwards, J. E., & DeLessio, J. (1991). A logistic regression model for personnel selection. *Applied Psychological Measurement*, *15*(2), 139–152. http://doi.org/10.1177/014662169101500204

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. http://doi.org/10.2307/2335739

Rydberg, S. (1963). *Bias in prediction on correction methods*. Stockholm: Almqvist & Wiksell. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A567526&dswid=9465

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

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, *76*(3), 323–346. http://doi.org/10.1348/096317903769647201

Schafer, J. L. 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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, *64*(6), 609.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*(4), 473–485. http://doi.org/10.1037/0021-9010.61.4.473

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

Society for Industrial and Organizational Psychology, & American Psychological Association. (1980). *Principles for the validation and use of personnel selection procedures*. Bowling Green OH: Society for Industrial and Organizational Psychology.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603–607. http://doi.org/10.1214/aoms/1177728730

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, *23*(5), 565–578. http://doi.org/10.1037/h0057079

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

Ulrich, R., & Wirtz, M. (2004). On the correlation of a naturally and an artificially dichotomized variable. *British Journal of Mathematical and Statistical Psychology*, *57*(2), 235–251. http://doi.org/10.1348/0007110042307203

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/

Weitzman, R. A. (2005). The prediction of college achivement by the scholastic aptitude test and the high school record. *Journal of Educational Measurement*, *19*(3), 179–191. http://doi.org/10.1111/j.1745-3984.1982.tb00126.x

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.

Zhang, L. (2008). Does state funding affect graduation rates at public four-year colleges and universities? *Educational Policy*, *23*(5), 714–731. http://doi.org/10.1177/0895904808321270

Zimmerman, D. W., & Williams, R. H. (2000). Restriction of range and correlation in outlier-prone distributions. *Applied Psychological Measurement*, *24*(3), 267–280. http://doi.org/10.1177/01466210022031741

1st Publication

Pfaffel, A., Schober, B., & Spiel, C. (2016). A comparison of three approaches to correct for direct and indirect rang restrictions: A simulation study. *Practical Assessment, Research & Evaluation, 21*(6), 1-15. Available online: http://pareonline.net/getvn.asp?v=21&n=6

Contribution by the individual authors:

Pfaffel, A.	Conception and design, R programming for the MC simulations, statistical analysis and interpretation of data, creation of graphics, drafting the paper, responsible for the revision process
Schober, B.	Recommendations for improvement of intellectual content, supervision of manuscript preparation
Spiel, C.	Revising the paper critically for important intellectual content, supervision of manuscript preparation

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.

a)

X_r	Y_r
X_u	–

b)

Z_r	X_r	Y_r
Z_u	X_u	–

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).

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

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):

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

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):

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

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.

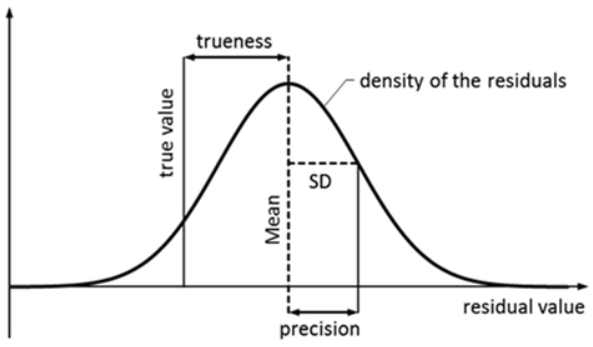

Figure 2. Graphical illustration of the concept of accuracy (trueness and precision).

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).

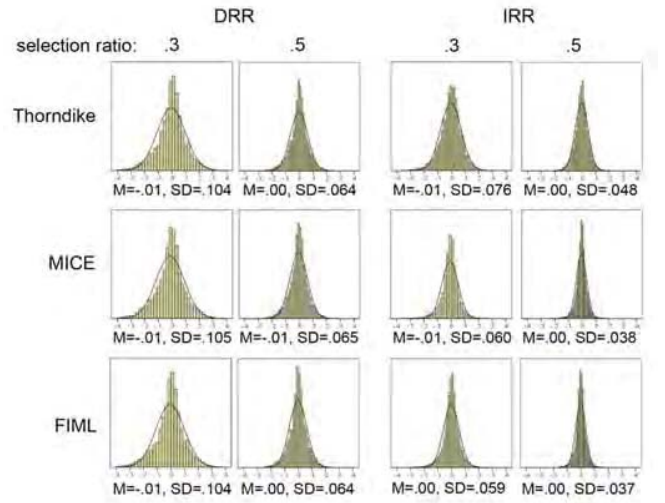

Figure 3. Distribution of the residuals for the population correlation estimates for the three approaches (Thorndike, MICE, FIML), for DRR and IRR, and for selection ratios of 30% and 50%.

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.

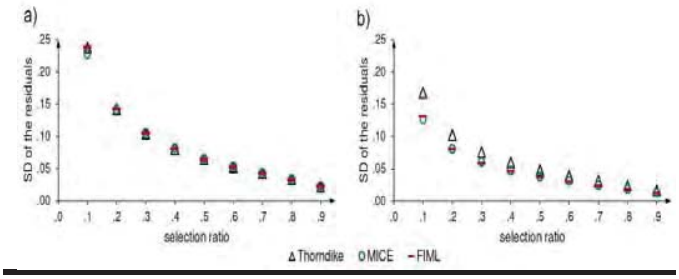

Figure 4. Exponential relationship between the selection ratio and the standard deviation of the residuals of the population correlation estimates for the three approaches (Thorndike, MICE, and FIML), for a) a DRR scenario, and b) an IRR scenario.

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: $SD = b_0 e^{b_1 SR}$, 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%.

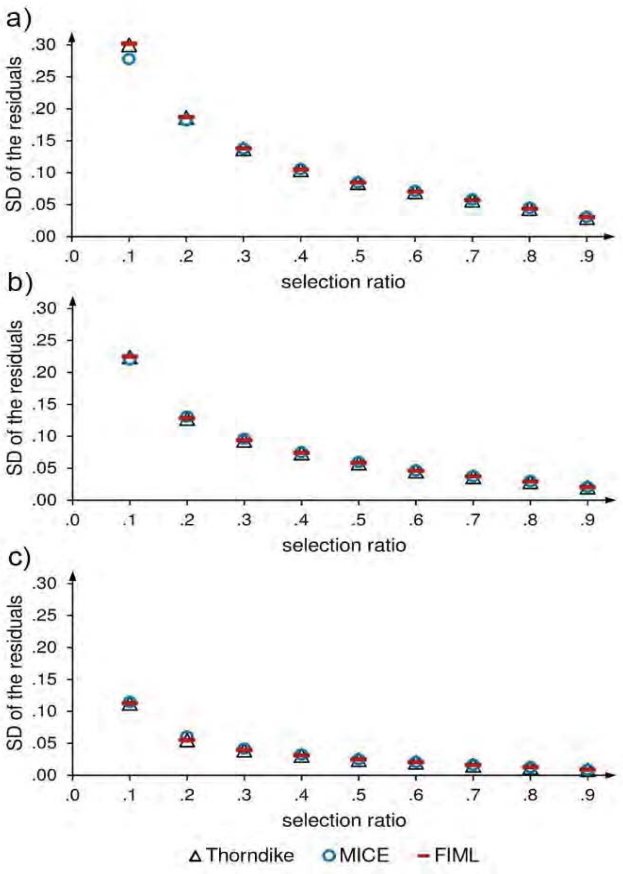

Figure 5. Effect of a a) weak, b) moderate and c) strong true population correlation on the precision of the population correlation estimates of the three approaches (Thorndike, MICE, FIML) for a DRR scenario.

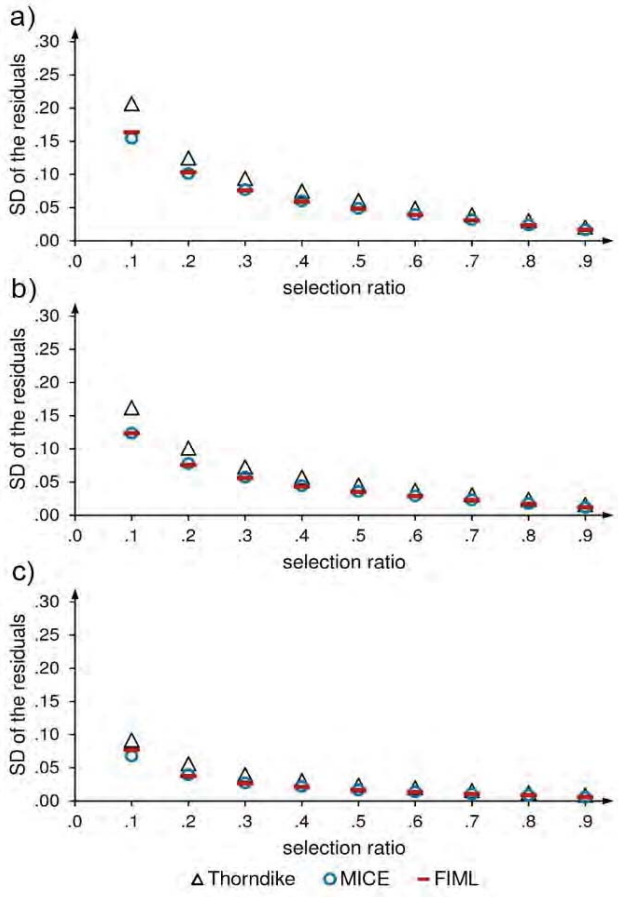

Figure 6. Effect of a a) weak, b) moderate and c) strong true population correlation on the precision of the population correlation estimates of the three approaches (Thorndike, MICE, FIML) for an IRR scenario.

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

2nd Publication

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect rang restrictions with a dichotomous criterion: A simulation study. *PLoS ONE, 11*(3), e0152330. Available online: http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0152330

Contribution by the individual authors:

Pfaffel, A.	Conception and design, R programming for the MC simulations, statistical analysis and interpretation of data, creation of graphics, drafting the paper, responsible for the revision process
Kollmayer, M.	Interpretation of data, improving the understanding and readability of the paper, assistance in the implementation of the reviewers comments
Schober, B.	Recommendations for improvement of intellectual content, supervision of manuscript preparation
Spiel, C.	Revising the paper critically for important intellectual content, supervision of manuscript preparation

PLOS ONE

RESEARCH ARTICLE

A Missing Data Approach to Correct for Direct and Indirect Range Restrictions with a Dichotomous Criterion: A Simulation Study

Andreas Pfaffel*, Marlene Kollmayer, Barbara Schober, Christiane Spiel

Department of Applied Psychology: Work, Education, Economy, Faculty of Psychology, University of Vienna, Vienna, Austria

* andreas.pfaffel@univie.ac.at

Abstract

A recurring methodological problem in the evaluation of the predictive validity of selection methods is that the values of the criterion variable are available for selected applicants only. This so-called range restriction problem causes biased population estimates. Correction methods for direct and indirect range restriction scenarios have widely studied for continuous criterion variables but not for dichotomous ones. The few existing approaches are inapplicable because they do not consider the unknown base rate of success. Hence, there is a lack of scientific research on suitable correction methods and the systematic analysis of their accuracies in the cases of a naturally or artificially dichotomous criterion. We aim to overcome this deficiency by viewing the range restriction problem as a missing data mechanism. We used multiple imputation by chained equations to generate complete criterion data before estimating the predictive validity and the base rate of success. Monte Carlo simulations were conducted to investigate the accuracy of the proposed correction in dependence of selection ratio, predictive validity, and base rate of success in an experimental design. In addition, we compared our proposed missing data approach with Thorndike's well-known correction formulas that have only been used in the case of continuous criterion variables so far. The results show that the missing data approach is more accurate in estimating the predictive validity than Thorndike's correction formulas. The accuracy of our proposed correction increases as the selection ratio and the correlation between predictor and criterion increase. Furthermore, the missing data approach provides a valid estimate of the unknown base rate of success. On the basis of our findings, we argue for the use of multiple imputation by chained equations in the evaluation of the predictive validity of selection methods when the criterion is dichotomous.

Citation: 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 11(3): e0152330. doi:10.1371/journal.pone.0152330

Editor: Lise Lotte Gluud, Hvidovre Hospital, DENMARK

Received: July 30, 2015

Accepted: March 11, 2016

Published: March 28, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This article was supported by the Open Access Publishing Fund of the University of Vienna.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

A recurring methodological problem in the evaluation of the predictive validity of selection methods is that the values of the criterion variable are available only for selected applicants. This loss of criterion data for non-selected applicants is an inherent effect of selection and is

known as the *range restriction problem* [1–5]. The problem occurs, for example, in the evaluation of an admission test in higher education, because data on academic success are only available for applicants who are admitted to the program. As an effect of the selection, the sample of selected applicants is not random and therefore not representative of the applicant population. Consequently, the observed sample correlation is a biased estimate of the population correlation, i.e. of the predictive validity. The correlation between a predictor X and a criterion Y obtained from the (available) range restricted dataset (i.e., the selected sample) underestimates the correlation we would obtain from the (not available) unrestricted dataset. Hence, this biased sample correlation has to be corrected to provide a more valid population estimate.

Correction methods for the range restriction problem have been widely studied for continuous criterion variables [5–17]. However, sociological, medical, and psychological research often deal with dichotomous criterion variables [18,19]. Dichotomous variables are characterized by a division of the individuals of a sample or population into two groups. The division can be based on either a qualitative or a quantitative characteristic. In the former case, the dichotomous variable is labelled as natural, and in the latter case as artificial [20]. For example, in higher education, the graduation status of a student is naturally dichotomous ('graduated' versus 'not graduated'). An artificially dichotomous variable is one that has a continuous underlying scale, but has been dichotomized (e.g., 'high performers' versus 'low performers'). The few existing approaches [21,22] to correct the biased correlation in the case of a dichotomous criterion are inapplicable because they require information about the base rate of success, i.e. the proportion of successful individuals in the unrestricted dataset. However, this information is typically not available. Thus, there is a lack in scientific research on suitable correction methods and their accuracies when the criterion variable is dichotomous.

In the present paper, we aim to overcome this deficiency by viewing the range restriction problem as a missing data mechanism [14,23]. As there is comprehensive literature on dealing with missing data, we can draw on a variety of techniques and research results. This potential is a great advantage of this approach, which has not yet been used to correct for range restriction with a dichotomous criterion variable [24–26]. We apply this approach to the two most common selection scenarios in personnel selection and higher education [27]: the *direct range restriction* (DRR) scenario and the *indirect range restriction* (IRR) scenario. In a DRR scenario, selection is based directly on the predictor variable X, whereas in an IRR scenario, selection is based on another variable Z.

First of all, we describe the loss of data in the two selection scenarios DRR and IRR, show which data are used for the correction, and give a brief introduction to Thorndike's well-known and widely used correction formulas in the case of a continuous criterion variable. Next, we provide a brief overview of methods for handling missing data and demonstrate that the proposed approach, *multiple imputation by chained equations*, is suitable for correcting for range restriction in both scenarios involving a dichotomous criterion. Then, we emphasize the critical role of the base rate of success, which has not been taken into account in previous approaches. Our proposed missing data approach generates complete data from which the base rate of success as well as the unbiased predictive validity can be obtained. Finally, we investigate the accuracy of the proposed correction by conducting Monte Carlo simulations, which allow for a comparison of the corrected parameters and the unbiased parameters (predictive validity and base rate of success) in an experimental design.

Direct and indirect range restriction

The most straightforward selection scenario is the direct range restriction (DRR) scenario, or explicit selection, which is commonly referred to as Thorndike's Case 2 [4,5]. In a DRR

scenario, selection is based directly on the predictor variable X and occurs top down. The idea is that applicants with a higher value of X are more suitable, and thus more likely to have a higher value in Y. As selection is based on values of X, the range of X is restricted in the selected sample. For this reason, this methodological problem is called the range restriction problem. The variable X can be either a score 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). For example, in higher education in Austria, prospective students of medicine are selected based solely on an entrance examination [28–30]. 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 the basis of another variable Z, which is usually correlated with X, Y, or both. The IRR scenario or incidental selection is commonly referred to as Thorndike's Case 3 [4,5]. Although selection is based on Z, the predictive validity of X remains of interest. Suppose a selection procedure consists of a psychometric test and a quantitative interview, and we want to assess the predictive validity of the psychometric test, the predictor X. For selection, we use the composite score Z derived from both selection methods. Organizations often use a composite score for selection and need to know the predictive validity of a single selection method in order to increase the predictive validity of the whole selection procedure (e.g., by removing or weighting 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.

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 be unselected. Hence, in this case we have missing values in the criterion variable Y for 60% of the applicants, but no missing values in X or Z. Fig 1 shows the missing data pattern for a SR of 40% in the cases of DRR and IRR.

In both scenarios, the observed sample correlation between X and Y is smaller than the correlation we would obtain from the unrestricted dataset, i.e. the predictive validity of the selection method is underestimated. To overcome this problem in the case of a continuous criterion variable, Thorndike [5] presented formulas to correct the Pearson correlation coefficient for DRR and IRR. The goal of these correction formulas is to estimate the correlation in the unrestricted dataset, which is the best estimate available of the population correlation, based on the correlation obtained from the restricted dataset. Correction formulas are commonly applied in predictive validity studies of large-scale testing programs, such as the Graduate Record Examination (GRE) [31,32], the Scholastic Aptitude Test (SAT) [33,34], and the Graduate Management Admission Test (GMAT) [35]. Correction formulas are also applied in other fields, e.g. in predicting job performance [36], and in evaluating the selection of pilot candidates in the US Air Force [37].

The formula for correcting for direct range restriction (DRR) presented by Thorndike is:

$$\rho_{XY} = \frac{(S_X/s_X)r_{XY}}{\sqrt{1 + r_{XY}^2(S_X^2/s_X^2 - 1)}}, \quad (1)$$

where ρ_{XY} is the true or unrestricted correlation coefficient, r_{XY} is the biased Pearson correlation coefficient obtained from the restricted dataset, and s_X and S_X are the standard deviations of X for the restricted and the unrestricted datasets [4]. The formula for correcting for indirect

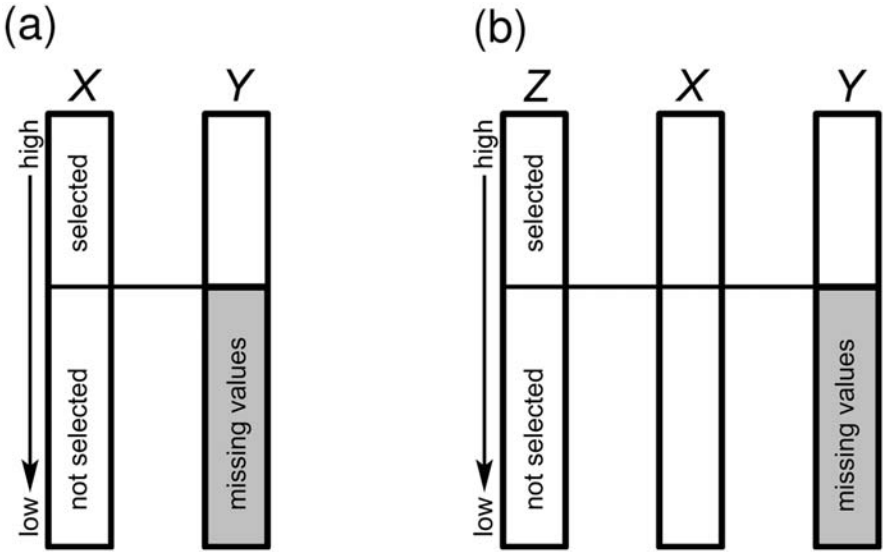

Fig 1. Missing data patterns under range restriction. (a) Direct range restriction scenario (selection on *X*), and **(b)** indirect range restriction scenario (selection on *Z*). The shaded areas in *Y* represent the location of the missing values in the dataset.

doi:10.1371/journal.pone.0152330.g001

range restriction (IRR) is:

$$\rho_{XY} = \frac{r_{XY} + r_{ZX} \cdot r_{ZY}(S_Z^2/s_Z^2 - 1)}{\sqrt{1 + r_{ZX}^2(S_Z^2/s_Z^2 - 1)} \cdot \sqrt{1 + r_{ZY}^2(S_Z^2/s_Z^2 - 1)}}, \quad (2)$$

where r_{XY}, r_{ZX}, and r_{ZY} are the uncorrected Pearson correlation coefficients obtained from the restricted dataset, and s_Z and S_Z are the standard deviations of variable Z for the restricted and the unrestricted dataset [4]. The core term in both correction formulas is the ratio of the standard deviations of the selection variable (X or Z).

The two formulas require that the assumption of linearity between X and Y as well as the assumption of homoscedasticity hold. In psychometric literature, it is well documented that corrected correlations are less biased than uncorrected correlations, even over a wide range of assumption violations [9,17,38,39]. Correcting for range restriction is recognized as professional practice because the corrected correlation coefficient is generally the best estimate of the population validity coefficient [40]. In general, the accuracy of the correction increases as the selection ratio increases and as the predictive validity increases [41]. Whereas Thorndike's correction formulas have been widely studied for continuous criterion variables, they have not been studied for dichotomous ones. Therefore, we investigate how usable Thorndike's correction formulas are in the case of a dichotomous criterion variable. Furthermore, we propose an approach based on state of the art methods for dealing with missing values that has not yet been applied in predictive validity studies [23].

Range restriction as a missing data mechanism

First, we give a brief overview of missing data mechanisms to locate the range restriction problem in this line of research. Afterwards, we introduce different methods of handling missing data and propose an approach for handling missing values in dichotomous dependent variables. One advantage of viewing the range restriction as a missing data mechanism is that we

can draw on a variety of techniques and research results in dealing with missing data [24,25,42]. However, this approach is seldom used with range restriction problems [14].

Rubin [43] describes three mechanisms essential as assumptions in dealing with missing values. These three mechanisms describe how the probability of a missing value relates to the data [24]: (1) Missing completely at random (MCAR) means the probability of missing data on Y is unrelated to other measured variables and is unrelated to the values of Y itself; (2) Missing at random (MAR) means the probability of missing data on Y is related to some other measured variable (or variables) in the analysis model but not to the values of Y itself; and (3) Missing not at random (MNAR) means the probability of missing data on Y is related to the values of Y itself, even after controlling for other variables. We consider both selection scenarios to be MAR, because the probability of missing values on Y depends either on X in the case of DRR, or on Z in the case of IRR, and not on values of Y itself. In other words, there is no relationship between the probability of missing values on Y and the values of Y after partialling out other variables. In the case of a MAR mechanism, we can estimate the missing values based on the observed values [24].

Over the past few decades, methodologists have proposed different techniques for dealing with missing data. Many of these approaches have enjoyed widespread use, but several of them, like listwise or pairwise deletion and single imputation techniques (e.g., arithmetic mean imputation, single regression imputation, and single EM imputation) are no longer considered to be state of the art, because they have potentially serious drawbacks [24]. Listwise and pairwise deletion require an MCAR mechanism, and produce biased parameter estimates with MAR and MNAR data. Deletion of incomplete cases can reduce the statistical power dramatically, even when the data are MCAR. Single imputation techniques also produce biased parameter estimates with MAR data and attenuate standard errors. Single regression imputation and single EM imputation overestimate correlations and attenuate variances and covariances, even when the data are MCAR, because they impute the data with perfectly correlated scores [24,44]. In a single regression imputation, all imputed values fall directly on the regression line and therefore lack variability. In contrast, arithmetic mean imputation attenuates correlations. Consequently, single imputation techniques are not suitable for many reasons, especially with regard to estimating correlation coefficients.

The two approaches that methodologists currently regard as state of the art [25,26] are (1) full information maximum likelihood (FIML), and (2) multiple imputation (MI). Neither of these approaches suffers from the problems mentioned for deletion of incomplete cases and single imputation techniques. The FIML approach estimates the most plausible parameters of a statistical model given the data. In other words, the goal is to identify the population parameter values with the highest probability of producing the data of a certain sample. The population parameter values are estimated with iterative optimization algorithms (e.g., expectation maximization algorithm). For a detailed description of likelihood-based approaches, see Little and Rubin [25], or for a less technical description see Enders [24].

The second state of the art approach to handle missing data problems is multiple imputation (MI) [24,25,45]. A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. The imputation phase creates several complete datasets (e.g., $m = 20$ imputations) based on one dataset with missing values. Each of these complete datasets contains different plausible estimates of the missing values, but identical values for the observed data. In the analysis phase, data can be analyzed with conventional statistical methods, but the analysis has to be performed m times, once for each complete dataset. The goal of the pooling phase is to combine the m parameter estimates into a single set of parameter estimates. The pooled parameter estimate is simply the arithmetic average of the m estimates from the analysis phase [46]. Analyzing multiple datasets and pooling the results sounds laborious, but modern MI software packages automate this procedure.

FIML und MI make the same assumptions (MAR and multivariate normality), have similar statistical properties, and frequently produce equivalent results [24,42]. Despite making the same assumptions, the two approaches differ in their mathematical background: The mathematical background of FIML is maximum likelihood estimation, whereas MI is based on Bayesian estimation. Therefore, there are important differences between the two approaches. While FIML maximizes the likelihood function to estimate the parameters without replacing missing values, MI replaces the missing values before estimating the parameters from the complete datasets. In contrast to FIML, MI effectively separates the imputation and the analysis phase. This may yield to different parameter estimates between the two approaches. Typically, in MI the imputation model includes many variables of the dataset, whereas the analysis model includes a subset of these variables.

Real data often do not conform to the modeling assumption of multivariate normality. Real data might be skewed, not negative, or bimodal, to name just a few deviations from normality. This mismatch between the distribution of the observed and imputed data may adversely affect the estimates of interest. MI generally tends to be robust against violations of normality [45,47,48]. Deviations from the normal distribution have a small effect on estimates that rely on the center of the distribution, like mean or regression coefficients, but may have significant effects on variances. Demirtas et al. [48] found that MI performs accurately with regard to the mean structure of skewed or multimodal distributions in large samples (n = 400), even for 75% missing values.

In the present study, we have to handle missing values in a dichotomous dependent variable. In light of this, we want to give a conceptual overview of Bayesian multiple imputation using logistic regression, which is considered to handle dichotomous variables most efficiently. Imputation of incomplete dichotomous variables is possible under the broad class of generalized linear models (GLM) [49]. The logistic regression models the probability that $Y_i = 1$ given X_i and model parameter vector β as [45]:

$$\Pr(Y_i = 1|X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} \quad (3)$$

The general idea is to estimate the probability model on the subset of the observed data (i.e., the restricted sample), and to impute the missing values with plausible values according to the fitted probabilities. For example, a probability of .80 means that Y_i has a chance of 80% of becoming 1 and a 20% chance of becoming 0. For a large number of imputations, the percentage of datasets in which $Y_i = 1$ tends towards 80%. The Bayesian method draws β from its respective posterior distributions. The posterior distribution contains the variability of β that needs to be incorporated into the imputations. In Bayesian statistics, Markov chain Monte Carlo (MCMC) methods are used to find the posterior distribution of the parameters. MCMC algorithms draw samples from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. The state of the chain after a very large number of steps is then used as a sample of the desired distribution. The quality of this sample increases with the number of steps. For a mathematical description of the Bayesian logistic regression imputation model, see Rubin [46].

Currently, MI is generally accepted as the best method for dealing with incomplete data in many fields [45]. Therefore, we suggest using multiple imputation by chained equations (MICE) to correct for range restriction in cases of DRR and IRR when the criterion variable is dichotomous. The proposed missing data approach first replaces the missing values of the criterion variable Y and generates several complete (unrestricted) datasets. Then, the correlation coefficient can be calculated based on these complete datasets.

The critical role of the base rate of success

A very important factor to be considered when correcting for range restriction with a dichotomous criterion is the base rate of success (BR) [21]. The BR is the percentage of individuals who would be successful on the criterion if there were no selection. The BR is calculated by dividing the number of successful individuals by the number of applicants, and ranges between 0 and 1, or between 0% and 100%. The BR contains unbiased information about the proportions of cases in the categories p ($Y = 0$; not successful) and q ($Y = 1$; successful) of a dichotomous criterion variable. For example, if all applicants are admitted to a study program and 60% of them complete this study program successfully, the BR is .6, or 60%.

Unfortunately, in the case of selection, the BR is unknown, as we cannot obtain the percentage of unselected applicants able to succeed on the criterion. We can only obtain the *success rate* of the selected sample, which is the number of successful individuals divided by the number of selected applicants. The success rate, however, is a biased estimator for the BR. Assuming a selection method determines the most suitable applicants, we will obtain a success rate which is higher than the BR.

Next, we will show how the two proportions p and q constituting the BR affect the magnitude of the correlation coefficient between a continuous variable X and a dichotomous variable Y [50,51]. Two correlation coefficients can be distinguished depending on whether the dichotomous variable is based on a qualitative or on a quantitative characteristic. In the former case, the dichotomous variable is labelled as natural, and in the latter case as artificial [20]. For a naturally dichotomous variable, the point-biserial correlation coefficient ρ_{pb} is used [50,51]:

$$\rho_{pb} = \frac{(M_1 - M_0)\sqrt{pq}}{\sigma_X}, \tag{4}$$

where M_1 and M_0 are the mean values of the continuous variable X for the group 'not successful' ($Y = 0$) and the group 'successful' ($Y = 1$), and σ_X is the standard deviation of X. p and $q = 1 - p$ represent the proportions of the two groups 'not successful' and 'successful'. ρ_{pb} ranges between -1 and +1. Normality of X is an assumption for significance testing, but not for calculating ρ_{pb}.

An artificially dichotomous variable is created whenever the values of a continuous variable are divided into two groups at a specific cut-off point. For example, student performance is measured on a continuous scale, and students are divided into 'low' and 'high' performers on the basis of their performance. In this case, a biserial correlation coefficient ρ_b is the more appropriate calculation. In the case of an artificially dichotomous variable, ρ_{pb} systematically underestimates the Pearson correlation coefficient which would have been obtained before dichotomization [41]. ρ_b is related to ρ_{pb} as shown in Formula 5:

$$\rho_b = \rho_{pb}\frac{\sqrt{pq}}{h}, \tag{5}$$

where h is the ordinate of the standard normal distribution at the point at which the cut for the dichotomization was made.

Both ρ_{pb} and ρ_b depend on the proportions p and q. As p and q become more extreme (e.g., .1 and .9), the correlation coefficient becomes smaller. Different values of the BR and the success rate result in different values of the correlation coefficients. Hence, p and q as obtained from the restricted dataset are different from the p and q we would obtain from the unrestricted dataset. Therefore, using the success rate to estimate the predictive validity results in biased correlation coefficients.

Two approaches have been proposed so far to assess the predictive validity of a selection method when the criterion variable is dichotomous. However, both approaches assume that the BR is known (e.g., from the literature), or should be assumed. The first approach is to apply the Taylor-Russell tables for a dichotomous criterion variable [21]. These tables indicate values of ρ_{pb} for the combination of the SR, the success rate, and the BR. The value for ρ_{pb} can only be taken from the tables if the values for the other three parameters are known. While the SR and the success rate are typically known, the BR is not and must be assumed. The second approach focuses on the effect size Cohen's *d* as a measure of the predictive validity in the case of a naturally dichotomous variable, and offers correction formulas for DRR and IRR [22]. The formulas correct the biased effect size *d* (obtained from the selected sample) into an unbiased *d* using the ratio of the unrestricted standard deviation to the restricted standard deviation, as known from Thorndike's correction formulas. Both formulas to calculate the unbiased *d* require the BR, which must be known or assumed. However, assuming the BR is an arbitrary approach, and different assumptions of the BR result in different values of the predictive validity, i.e. different values of the correlation coefficients.

So far, the scientific literature does not provide any correction method for situations in which the BR is unknown. However, when correcting for range restriction with a dichotomous criterion, both the biased success rate and the biased correlation have to be considered. The proposed missing data approach allows for this, as it generates complete datasets from which the BR as well as the unrestricted correlation can be obtained. Therefore, the proposed approach provides an empirical estimation for both the correlation coefficients and the BR based on the selected sample.

Purposes of this Study

Correction methods for range restriction have been studied almost exclusively for continuous criterion variables. Therefore, the aim of the present study was to give empirical evidence in an experimental design on correcting for direct range restriction (DRR) and indirect range restriction (IRR) when the criterion variable is dichotomous.

The first purpose is to compare the two approaches (1) multiple imputation by chained equations (MICE), and (2) Thorndike's correction formulas (Formulas 1 & 2) with regard to the accuracy of the correction of the biased sample correlations.

The second purpose is to investigate the effect of a weak, moderate, and strong relationship between predictor and criterion on the accuracy of the correction of the biased sample correlations with multiple imputation by chained equations. Studies investigating Thorndike's correction formulas have shown that the accuracy of the correction increases as the correlation between predictor and criterion increases.

The third purpose is to investigate the accuracy of the correction of the biased BR with multiple imputation by chained equations. Previous approaches are less useful when the criterion variable is dichotomous because they do not consider that the success rate is a biased estimate for the unknown BR. However, the proposed missing data approach allows us to estimate the BR.

The fourth purpose of this study is to investigate the effect of the strength of the correlation between *Z* and *X* on the accuracy of the correction with multiple imputation by chained equations in an IRR scenario.

Method

Procedure

We conducted Monte Carlo simulations to investigate the two correction approaches: a) Thorndike's correction formulas, and b) multiple imputation by chained equations (MICE) in

an experimental design using the program R Statistics [52]. We wrote four R scripts (see S1–S4 Rscripts) in order to conduct the Monte Carlo simulations for the following four conditions: 1. DRR with an artificially dichotomous criterion variable; 2. DRR with a naturally dichotomous criterion variable; 3. IRR with an artificially dichotomous criterion variable; and 4. IRR with a naturally dichotomous criterion variable. The Monte Carlo simulations were conducted with 5,000 iterations for each condition. The procedure for the Monte Carlo simulation consisted of the following steps.

Step 1—Data simulation. We generated 5,000 unrestricted multivariate datasets (sample size $N = 500$) for each condition by varying the correlation coefficient between X and Y from .10 to .90 and the base rate of success (BR) from 10% to 90%. In the case of IRR, there was a third variable Z, meaning that 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 levels of the selection ratio (SR) ranging from 10% to 90% with step width 10% (which corresponded to missing values in Y from 90% to 10%). This yielded $5000 * 9 = 45000$ restricted datasets. In the case of DRR, datasets were sorted in descending order by X; in the case of IRR, in descending order by Z. We selected those cases with the highest values in X (DRR), and with the highest values in Z (IRR). The percentage of selected cases depended on the SR. Values of Y for non-selected cases were deleted (i.e., converted into missing values). The range restricted or selected samples created in this way were saved into new datasets and were used for applying the correction.

Step 3—Correction. Both approaches were applied to the range restricted datasets. In the first approach, Thorndike's correction formulas for DRR (Eq 1) and IRR (Eq 2) were used to calculate the estimate of the correlation coefficient between predictor X and criterion Y. In the second approach, we used multiple imputation by chained equations to generate $m = 20$ imputed datasets (see subsection Imputation of the missing values). For each imputed dataset, we calculated the correlation coefficient between X and Y, and the BR. The MI analysis pools the $m = 20$ estimates into a single point estimate. Rubin [46] showed that the multiple imputation point estimate is the arithmetic mean of the m estimates.

Step 4—Analysis of parameter estimates. In order to analyse the accuracy of the correction, we compared the parameter estimates of both approaches with the true parameters obtained from the unrestricted dataset. All estimates of the parameters are denoted with the accent symbol hat, where $\hat{r}_b$ is the estimate of the biserial correlation coefficient and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient. We calculated the residual of each parameter estimate. For example, the residual for the point-biserial correlation coefficient was $\hat{r}_{pb} - \rho_{pb}$, where ρ_{pb} was the true unrestricted correlation coefficient.

When running Monte Carlo simulations, extreme conditions typically cause problems in statistical analysis. Consequently, marginal conditions have to be defined. The logistic regression could not be applied when Y was constant. This was particularly likely to be the case when the BR and the correlation between X and Y were high and the SR was small. Therefore, a minimum variance in Y, or a minimum number of observations with $Y = 1$ (or $Y = 0$) was a necessary precondition for a valid estimate. We determined that a minimum of five observations in the two categories ('successful' and 'not successful') was a sufficient number of observations in the restricted dataset. This minimum number of observations was based on the rule of thumb in chi-square statistics for contingency tables. Therefore, we excluded samples that did not meet this prerequisite.

Data simulation

We simulated multivariate data for two kinds of dichotomous criterion variables: a) an artificially dichotomous variable, and b) a naturally dichotomous variable. Both kinds of dichotomous criterion variables were simulated for a DRR and an IRR scenario. Purpose 1 was to compare the accuracy of the two approaches (Thorndike and MICE) for all possible combinations influencing the accuracy (ρ_{XY}, ρ_{ZX}, ρ_{ZY}, BR, SR). Therefore, we generated the data using uniform random values for the correlation coefficients, and for the BR, both varied continuously from .1 to .9. The continuous variation of the factors facilitated the subsequent calculation of estimates aggregated over all factors and factor levels. On the basis of these aggregated estimates, the comparison of the two approaches can be displayed more clearly than based a large number of factor combinations.

a) We generated a bivariate standard normal distribution (DRR) or a trivariate standard normal distribution (IRR) using the `mvrnorm()` function of the MASS package [53]. Table 1 shows the design of the intercorrelation matrix for the DRR and IRR scenarios. In order to create an artificially dichotomous criterion variable *Y*, we dichotomized one of the standard normally distributed variables at a specific cut-off point. The cut-off point corresponded to the BR, which represented the number of 'successful' and 'not successful' individuals. Values higher than the cut-off point were coded as 1 ('successful'); all other values were coded as 0 ('not successful'). For example, a cut-off point at zero (= mean of the standard normal distribution) represented a BR of 50% ($p = q = .50$).

b) In the case of a naturally dichotomous criterion, we simulated multivariate data based on a dichotomous variable (*Y*) and (for *X* and *Z*) a mixture of two univariate normal distributions, one normal distribution for each of the two criterion groups. This kind of data was used to develop the Taylor-Russell tables for a dichotomous criterion variable [21]. We followed this approach to be consistent with the literature on evaluating the predictive validity of a selection method when the criterion is dichotomous. First, we generated *Y* with the proportions of 'successful' and 'not successful' individuals based on the BR. Second, we generated two normally distributed variables (X_0 and X_1; one for each criterion group) with standard deviations of 1, and a mean difference $M_1 - M_0$. The mixture of X_0 and X_1 was the distribution of the continuous variable *X*. The mean difference is related to the amount of the point-biserial correlation coefficient ρ_{pb}. The higher the mean difference, the higher ρ_{pb} (for constant BR, and constant standard deviations of X_0 and X_1). For example, when X_0 and X_1 are normally distributed with standard deviations of 1, the mean difference is 1.5, and the BR is 50%. In this example, the standard deviation of *X* is 1.25, resulting in a point-biserial correlation coefficient $\rho_{pb} = 1.5 \cdot \sqrt{.25}/1.25 = .60$ (see Eq 4). For details on how to calculate σ_X for a mixture of two

Table 1. Design of the intercorrelation matrix of the correlation coefficients for direct range restriction (DRR) and indirect range restriction (IRR).

	DRR			IRR		
	X	*Y*		*X*	*Y*	*Z*
X	1	r_b/r_{pb}	*X*	1	r_b/r_{pb}	*r*
Y	r_b/r_{pb}	1	*Y*	r_b/r_{pb}	1	r_b/r_{pb}
			Z	*r*	r_b/r_{pb}	1

X is the predictor variable; *Y* is the dichotomous criterion variable; *Z* is the selection variable in the case of IRR; r_b is the biserial correlation coefficient; r_{pb} is the point-biserial correlation coefficient; *r* is the Pearson correlation coefficient.

doi:10.1371/journal.pone.0152330.t001

normal distributions in an analytical way, see Cohen [54]. With this procedure, the data simulation was completed for the DRR scenario. For the IRR scenario, we added the third variable Z in the same way as X, where Z was correlated with X and with Y. Li and colleagues also used this three-variable design in their Monte Carlo simulations to estimate the bootstrapped standard error of the Pearson correlation coefficient of Thorndike's correction formula for IRR [7].

Imputation of the missing values

We used the R package *MICE* (multivariate imputation by chained equations; version 2.22 [55]) to implement multiple imputation using fully conditional specification. The MICE package supports multivariate imputations of continuous data, binary data, unordered categorical data, and ordered categorical data. The algorithm imputed an incomplete variable by generating plausible values given other variables in the dataset. For the imputation of the dichotomous criterion variable, we used a Bayesian logistic regression implemented using the elementary imputation method `logreg()` of the MICE package. The imputation method `logreg()` was used with default specifications for the prior distributions and the Markov Chain Monte Carlo simulation (MCMC). Conventional wisdom suggests that multiple imputation analysis requires about $m = 5$ imputations [46,47]. This number of imputations was derived solely by considering the relative efficiency [24,46]. Contrary to this conventional wisdom, simulation 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 [24,42].

In our simulation study, we investigated samples with a rate of missing values up to 90% (corresponding to a SR of 10%). Therefore, we conducted a preliminary study to investigate the impact of the number of imputations on the accuracy of the parameter estimation dependent on the rate of missing values. We conducted Monte Carlo simulations using $m = 5$, 20, and 50 imputations for samples with rates of missing values of 70%, 50%, and 30% ($N = 500$). Table 2 shows the results of the preliminary study that $m = 20$ imputations provided a more accurate estimate than only $m = 5$ imputations. However, increasing the number of imputations beyond 20 provided no relevant improvement in the accuracy of the estimates. Therefore, we used $m = 20$ imputations in each simulation.

Analysis of parameter estimates

For our purpose of investigating the accuracy of the correction methods, we calculated the residual of each parameter estimate (Step 4 of the procedure). Accuracy is defined as the closeness of the estimated value to the true value of the parameter being estimated [56]. If the residual of a parameter estimate is close to zero, a correction method provides a very good estimation of the true parameter obtained from the unrestricted dataset. The concept of accuracy encompasses both precision (random error) and trueness (bias or systematic error), and therefore provides important quantitative information about the goodness of the correction. We used the root mean square error (RMSE) as a measure of precision, and the mean error (ME) as a measure of trueness. Let $\hat{\theta}$ be the parameter estimate and θ the true parameter, then the

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2}, \tag{6}$$

and the

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta). \tag{7}$$

Table 2. Results of the preliminary study: Root mean square errors of the correlation estimates using *m* = 5, 20, and 50 imputations for 70%, 50%, and 30% missing values (DRR scenario, *N* = 500, 1000 iterations).

	$\hat{r}_b$			$\hat{r}_{pb}$		
m	70%	50%	30%	70%	50%	30%
5	.163	.118	.075	.112	.082	.049
20	.161 (-.002)	.106 (-.012)	.067 (-.008)	.105 (-.007)	.073 (-.009)	.044 (-.005)
50	.158 (-.003)	.104 (-.002)	.070 (.003)	.106 (.001)	.070 (-.003)	.042 (-.002)

$\hat{r}_b$ is the estimate of the biserial correlation coefficient, $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient, *m* is the number of imputations. Values in brackets show the change in the RMSE as a result of the additional imputations.

doi:10.1371/journal.pone.0152330.t002

The RMSE provides information about the probability that a correction is close to the true value. A small RMSE represents a small random error, i.e. a correction with high precision. The ME is the sample arithmetic mean of the residuals. An estimate is biased if the ME is different from zero. A positive ME represents an overestimation, and a negative ME represents an underestimation of the true parameter value.

We used *F*-ratio tests to compare Thorndike's correction formulas with our proposed missing data approach in terms of the precision of the two estimates $\hat{r}_{pb}$ and $\hat{r}_b$. The *F*-ratio compares the mean square errors (MSEs), i.e. the variances of the residuals of the two approaches. For example, an *F*-ratio of 1 means that both correction methods have equal precision, while an *F*-ratio of 2 means that one correction method is twice as precise as the other one.

In order to investigate the effect of the strength of the relationship between predictor *X* and criterion *Y* (Purpose 2) and between selection variable *Z* and predictor *X* (Purpose 4) on the accuracy of the correction with multiple imputation by chained equations, we partitioned the true correlation coefficients obtained from the unrestricted dataset into three levels: a weak relationship (from .10 to < .40), a moderate relationship (from .40 to < .70), and a strong relationship (from .70 to .90). We compared the RMSEs of these three levels in order to demonstrate how the strength of the relationship between predictor and criterion affected the precision of the estimation.

Results

Figs 2–6 show the root mean square errors (RMSEs) of the estimated parameters in dependence of the selection ratio (SR). As an overall effect, the accuracy (trueness and precision) of all estimates gradually improved as the SR increased from .1 to .9, i.e. as the loss of criterion data decreased from 90% to 10%. For each purpose, except Purpose 4 that explicitly refers to an IRR scenario, we first display the results for the DRR scenario and then the results for the IRR scenario.

Purpose 1—Comparison of the two approaches

The first purpose was to compare the correction with multiple imputation by chained equations (MICE) with Thorndike's correction formulas with regard to the accuracy of the correlation estimates.

DRR scenario: Table 3 summarizes the mean errors (MEs) as a measure of the trueness of the predictive validity. The results show that both approaches underestimate the unrestricted correlation at SR of .1 (90% missing values). The underestimation at SR = .1 is larger when correcting with MICE than when correcting with Thorndike's formula (MICE about -.10;

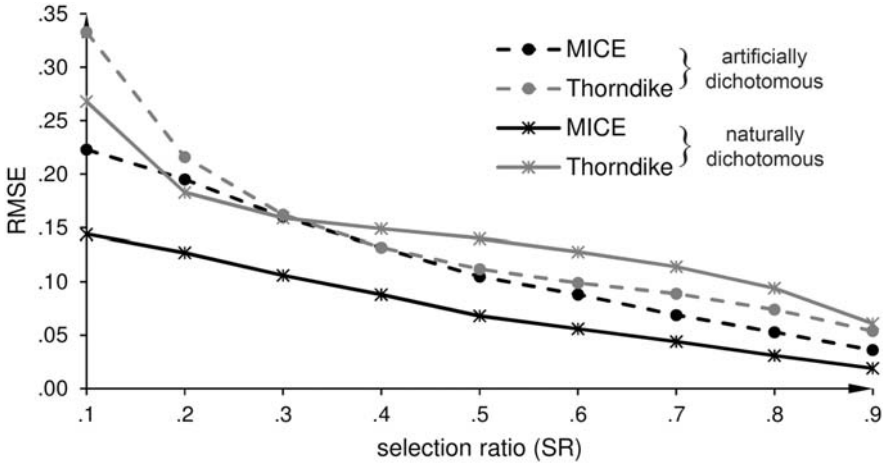

Fig 2. Direct range restriction (DRR): Root mean square error (RMSE) of the estimates of the predictive validity ($\hat{r}_b$ and $\hat{r}_{pb}$). $\hat{r}_b$ is the estimate of the biserial correlation coefficient for an artificially dichotomous criterion variable, and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient for a naturally dichotomous criterion variable.

doi:10.1371/journal.pone.0152330.g002

Thorndike about -.05). For SRs beyond .2, the estimates for both kinds of criterion variables are less biased when correcting with MICE. Next, we compared the RMSEs of $\hat{r}_{pb}$ (naturally dichotomous criterion variable) and $\hat{r}_b$ (artificially dichotomous criterion variable) when correcting with MICE and with Thorndike's correction formula for DRR (Eq 1). In the case of $\hat{r}_b$, the *F*-ratios range from 1.14 to 2.25 (all $ps < .001$), except at SR = .3 and SR = .4 (*F*-ratios 1.03 and 1.00). In the case of $\hat{r}_{pb}$, the *F*-ratios range from 2.08 to 10.3 (all $ps < .001$), as shown in Table 4. Thus, the correction with MICE is more precise than the correction with Thorndike's formula (Eq 1) for both kinds of dichotomous criterion variables. The difference in the extent

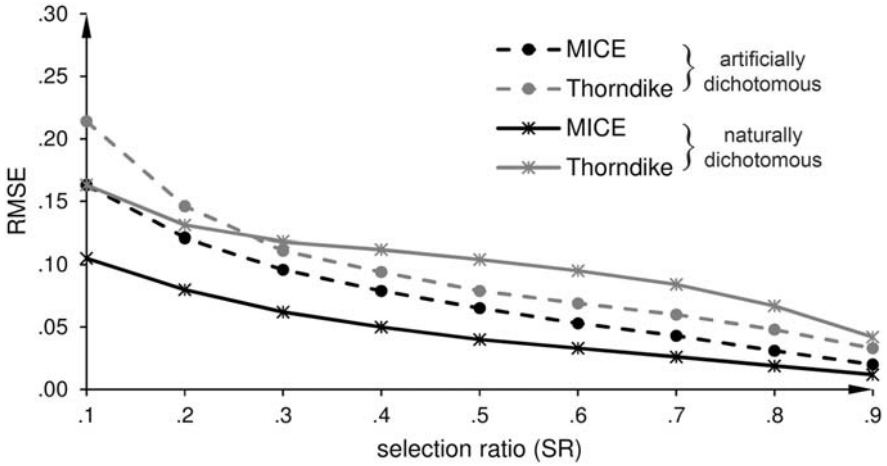

Fig 3. Indirect range restriction (IRR): Root mean square error (RMSE) of the estimates of the predictive validity ($\hat{r}_b$ and $\hat{r}_{pb}$). $\hat{r}_b$ is the estimate of the biserial correlation coefficient for an artificially dichotomous criterion variable, and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient for a naturally dichotomous criterion variable.

doi:10.1371/journal.pone.0152330.g003

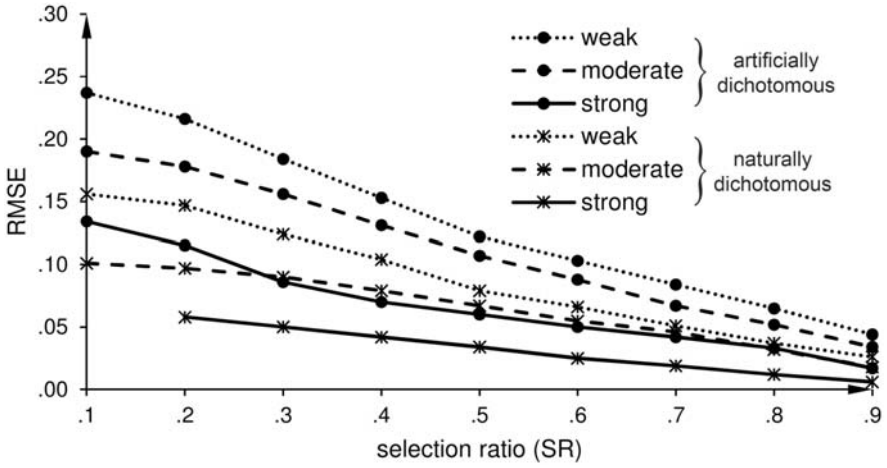

Fig 4. Direct range restriction (DRR): Effects of a weak, moderate, and strong predictive validity on the root mean square error (RMSE) of the estimates of the predictive validity ($\hat{r}_b$ and $\hat{r}_{pb}$). $\hat{r}_b$ is the estimate of the biserial correlation coefficient for an artificially dichotomous criterion variable, and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient for a naturally dichotomous criterion variable.

doi:10.1371/journal.pone.0152330.g004

of precision between the two approaches is higher for a naturally dichotomous criterion variable than for an artificially dichotomous criterion variable. Fig 2 shows the RMSEs of both correlation estimates ($\hat{r}_b$ and $\hat{r}_{pb}$) for our proposed correction with multiple imputation by chained equations (MICE) and for the correction with Thorndike's formula (Eq 1).

IRR scenario: As shown in Table 3, MICE underestimates the unrestricted correlation for both kinds of criterion variables at SR = .1. However, this bias tends to be smaller for the IRR scenario than for the DRR scenario. With regard to the precision of the estimates, Fig 3 shows that the correlation estimates are more precise for our proposed correction with MICE than for the correction with Thorndike's formula. The course of the lines is similar to the DRR scenario.

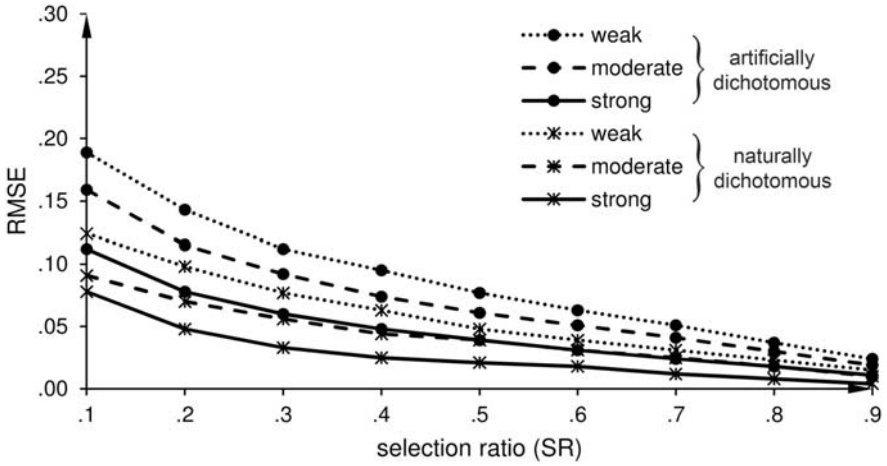

Fig 5. Indirect range restriction (IRR): Effects of a weak, moderate, and strong predictive validity on the root mean square error (RMSE) of the estimates of the predictive validity ($\hat{r}_b$ and $\hat{r}_{pb}$). $\hat{r}_b$ is the estimate of the biserial correlation coefficient for an artificially dichotomous criterion variable, and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient for a naturally dichotomous criterion variable.

doi:10.1371/journal.pone.0152330.g005

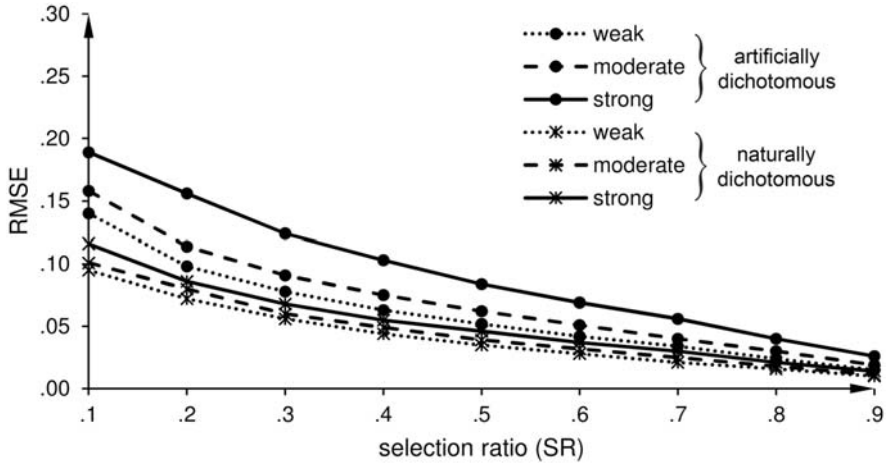

Fig 6. Indirect range restriction (IRR): Effects of a weak, moderate, and strong relationship between predictor *X* and selection variable *Z* on the root mean square error (RMSE) of the estimates of the predictive validity ($\hat{r}_b$ and $\hat{r}_{pb}$). $\hat{r}_b$ is the estimate of the biserial correlation coefficient for an artificially dichotomous criterion variable, and $\hat{r}_{pb}$ is the estimate of the point-biserial correlation coefficient for a naturally dichotomous criterion variable.

doi:10.1371/journal.pone.0152330.g006

Table 3. Mean errors (ME) of the correlation estimates.

		Selection ratio (SR)								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
DRR, artificially dichotomous	MICE	-.12	-.05	-.02	<.01	<.01	<.01	<.01	<.01	<.01
	Thorndike	-.06	-.05	-.05	-.04	-.04	-.03	-.03	-.02	<.01
DRR, naturally dichotomous	MICE	-.09	-.04	-.02	-.01	<.01	<.01	<.01	<.01	<.01
	Thorndike	-.05	-.05	-.05	-.05	-.04	-.03	-.02	-.01	<.01
IRR, artificially dichotomous	MICE	-.08	-.03	-.01	<.01	<.01	<.01	<.01	<.01	<.01
	Thorndike	-.04	-.02	-.02	-.01	-.01	-.01	<.01	<.01	<.01
IRR, naturally dichotomous	MICE	-.07	-.03	-.02	-.01	<.01	<.01	<.01	<.01	<.01
	Thorndike	-.01	-.03	-.04	-.04	-.03	-.03	-.02	-.01	<.01

$\hat{r}_b$ is the estimate of the biserial correlation coefficient, $\hat{r}_{pb}$ is the estimate point-biserial correlation coefficient.

doi:10.1371/journal.pone.0152330.t003

Table 4. *F*-ratio of the correlation estimates when correcting with multiple imputation by chained equations and Thorndike's formulas.

	Selection ratio (SR)								
	.1	.2	.3	.4	.5	.6	.7	.8	.9
DRR, artificially dichotomous	2.23**	1.23**	1.03	1.00	1.14**	1.27**	1.66**	1.95**	2.25**
DRR, naturally dichotomous	3.46**	2.08**	2.25**	2.87**	4.24**	5.22**	6.71**	9.19**	10.3**
IRR, artificially dichotomous	1.72**	1.46**	1.34**	1.42**	1.48**	1.69**	1.95**	2.40**	2.72**
IRR, naturally dichotomous	2.41**	2.68**	3.62**	5.02**	6.76**	8.29**	10.4**	12.4**	12.3**

$\hat{r}_b$ is the estimate of the biserial correlation coefficient, $\hat{r}_{pb}$ is the estimate point-biserial correlation coefficient, *F*-ratio is calculated by the mean square error (MSE) of the estimate using Thorndike's formula divided by the MSE of the estimate using MICE, ** $p < .001$.

doi:10.1371/journal.pone.0152330.t004

The differences in the RMSEs between MICE and Thorndike's formula are larger for $\hat{r}_{pb}$ than for $\hat{r}_{b}$, as shown in Table 4. The *F*-ratios for $\hat{r}_{b}$ range from 1.34 to 2.72 (all *p*s $< .001$), and for $\hat{r}_{pb}$ from 2.41 to 12.4 (all *p*s $< .001$).

Purpose 2—The effect of the strength of the relationship (X, Y)

The second purpose was to investigate the effect of the strength of the relationship between predictor *X* and criterion *Y* on the accuracy of the correction with MICE. Therefore, we investigated the effect of a weak, moderate, and strong relationship between *X* and *Y* on the precision of the correction with MICE.

DRR scenario: Fig 4 shows that the precision of $\hat{r}_{pb}$ and $\hat{r}_{b}$ increases (RMSEs decrease) when the strength of the correlation in the unrestricted dataset increases. In Fig 4, we excluded the value of $\hat{r}_{pb}$ for the condition of SR = .1 combined with a strong relationship between *X* and *Y*, because for this case only three restricted datasets met the prerequisite.

IRR scenario: Similar to the DRR scenario, the precision of the estimated correlation coefficient for naturally and artificially dichotomous criterion variables increases when the strength of the relationship between *X* and *Y* increases, as shown in Fig 5.

Purpose 3—Correcting the biased base rate of success (BR)

The third purpose was to investigate the accuracy of the correction of the biased BR with MICE. Table 5 summarizes the MEs and the RMSEs of the estimate of the base rate of success ($\hat{BR}$) for the two scenarios and both kinds of criterion variables.

DRR scenario: The mean errors in Table 5 show an overestimation of the base rate of success (+.07) at an SR of .1 for both kinds of criterion variables. This effect is contrary to the correlation estimates, which underestimate the unrestricted correlation. For SRs beyond .2, the estimates are not biased. In the same manner as for the estimation of the correlation coefficients, the RMSEs of $\hat{BR}$ decreases as the selection ratio increases (from .157 to .007 for an artificially dichotomous criterion variable and from .154 to .005 for a naturally dichotomous one).

IRR scenario: The results for the IRR scenario are similar to the DRR scenario. The MEs show an overestimation of the base rate of success only at an SR of .1 and the precision of $\hat{BR}$ increases as the selection ratio decreases.

Table 5. Accuracy of the estimate of the base rate of success when correcting via multiple imputation by chained equations (MICE).

		Selection ratio (SR)								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
DRR, artificially dichotomous	ME	.07	.01	<.01	<.01	<.01	<.01	<.01	<.01	<.01
	RMSE	.157	.120	.090	.066	.047	.034	.022	.014	.007
DRR, naturally dichotomous	ME	.07	.02	<.01	<.01	<.01	<.01	<.01	<.01	<.01
	RMSE	.154	.111	.081	.059	.040	.028	.018	.011	.005
IRR, artificially dichotomous	ME	.07	.01	<.01	<.01	<.01	<.01	<.01	<.01	<.01
	RMSE	.151	.113	.084	.061	.045	.031	.020	.012	.006
IRR, naturally dichotomous	ME	.08	.02	<.01	<.01	<.01	<.01	<.01	<.01	<.01
	RMSE	.142	.108	.078	.054	.037	.026	.016	.010	.005

DRR is the direct range restriction, IRR is the indirect range restriction, ME is the mean error, and RMSE is the root mean square error.

doi:10.1371/journal.pone.0152330.t005

Purpose 4—The effect of the strength of the relationship (*Z*, *X*)

The fourth purpose of this study was to investigate the effect of the strength of the relationship between *Z* and *X* on the accuracy of the correction with MICE in an IRR scenario. Therefore, we investigated the effect of a weak, moderate, and strong relationship between the selection variable *Z* and the predictor variable *X* on the precision of the correction with MICE. The results in Fig 6 show that the precision of the estimates increases (RMSEs decrease) when the relationship between *Z* and *X* decreases.

Discussion

A recurring methodological problem in the evaluation of the predictive validity of selection methods is the loss of data for the criterion variable. This so-called range restriction problem results in biased population estimates because the observed sample (the selected sample) is not representative of the population of interest (the applicant population). Hence, these biased estimates have to be corrected. However, researchers have almost exclusively focused on correction in the case of a continuous criterion variable. Therefore, our aim was to propose an approach for correcting for range restriction when the criterion variable is dichotomous. We applied this approach to the two most common selection scenarios in personnel selection and higher education: a direct range restriction scenario (DRR) and an indirect range restriction scenario (IRR). We investigated two kinds of dichotomous criterion variables: artificially and naturally dichotomous criterion variables.

The proposed approach correcting for range restriction is to view the selection as a missing data mechanism. We used multiple imputation by chained equations (MICE), which is a state-of-the-art method for dealing with missing data. We pointed out the importance of the unknown base rate of success, which has to be considered when correcting for range restriction in the case of a dichotomous criterion. The proposed approach corrects for range restriction by replacing the missing values of the criterion variable before estimating the predictive validity and the BR at the same time.

We investigated the accuracy of the proposed correction by conducting Monte Carlo simulations, which allowed us to compare the parameter estimates with the true parameters in an experimental design. In the present simulation study, we varied several factors (correlations, base rate of success, and selection ratio) over a wide range in order to examine the accuracy of the correction for a variety of possible datasets.

We compared our proposed missing data approach with Thorndike's formulas (established for a continuous criterion) in terms of the accuracy of the parameter estimates. The Monte Carlo simulations show that our proposed approach performs effectively in both the DRR scenario and the IRR scenario. The correction of the biased predictive validity with MICE is more precise than the correction with Thorndike's formulas. Furthermore, we were able to show that the missing data approach provides a valid estimate of the base rate of success that has not been considered in the scientific literature. On the basis of our findings, we argue for the use of multiple imputation by chained equations in the evaluation of the predictive validity of selection methods when the criterion is dichotomous. To our knowledge, the proposed correction for range restriction using multiple imputation by chained equations is the first approach that provides a proper correction for the biased predictive validity when the criterion variable is dichotomous. The missing data approach facilitates the correction of the biased correlation coefficient as well as of the unknown base rate of success.

Some limitations of our study should be mentioned that open the field for further research. In the simulation study, we used a small data matrix of two variables in the DRR scenario and of three variables in the IRR scenario [7]. As in simulation studies, it is often difficult to

generate multivariate random correlated datasets, especially for multivariate non-normal distributions. Further research should examine the effect of a data matrix with more variables on the accuracy of the correction. In the present study, we investigated the accuracy of the correction for one sample size. However, one important research question with regard to the multiple imputation by chained equations approach is how small the total sample size as well as the restricted sample size can be for a precise and unbiased correction. We recommend investigating these limitations in further studies.

The IRR scenario assumes that the selection variable Z is measured. In this case, the missing data mechanism is MAR (ignorable selection process, [3]), and therefore we can use a multiple imputation technique. However, in cases of incidental selections, Z is sometimes either partially measured or unmeasured. For example, this is the case when selection is based on an unquantified subjective judgment, or in the case of self-selection, when individuals remove themselves from a sample for reasons that are not measured. In such cases, the missing data mechanism is missing not at random (MNAR, non-ignorable selection process). In statistics, this methodological problem is known as sample selection bias [3]. Traditional range restriction corrections yield unsatisfactory estimates of r_{XY} when the selection process is non-ignorable [3,39]. A correction procedure for selection bias for a continuous dependent variable has been developed in the field of economics [57,58]. Muthén and Hsu [59] presented a latent variable model. In cases of non-ignorable selection, further studies should examine the accuracy of this latent variable model for a dichotomous criterion using weighted least squares means and variance adjusted (WLSMV). As another approach, MICE can also be used to correct for MNAR data [55].

Some recommendations for practitioners and organizations can be derived from our research. Sometimes, test data from applicants who were not selected are discarded are not available in later validity studies. However, discarding applicants' test data leads to a needless loss of information regarding the predictive validity of selection methods. Therefore, we recommend that organizations store the data in an anonymized form for future evaluations of the predictive validity. Although our approach is applicable for data with up to 90% missing values, we urge caution in the interpretation of the estimates when missing values exceed 70%.

In summary, the results from the simulation study show that the proposed correction with multiple imputation by chained equations is effective in correcting for DRR and IRR scenarios when the criterion variable is dichotomous. Therefore, the approach presented in this paper seems to be promising in terms of overcoming recurring range restriction problems in the evaluation of the predictive validity of selection methods.

Supporting Information

S1 Data. Direct range restriction—artificially dichotomous criterion variable.
(ZIP)

S2 Data. Direct range restriction—naturally dichotomous criterion variable.
(ZIP)

S3 Data. Indirect range restriction—artificially dichotomous criterion variable.
(ZIP)

S4 Data. Indirect range restriction—naturally dichotomous criterion variable.
(ZIP)

S1 Rscript. Direct range restriction—artificially dichotomous criterion variable.
(DOCX)

S2 Rscript. Direct range restriction—naturally dichotomous criterion variable.
(DOCX)

S3 Rscript. Indirect range restriction—artificially dichotomous criterion variable.
(DOCX)

S4 Rscript. Indirect range restriction—naturally dichotomous criterion variable.
(DOCX)

Author Contributions

Conceived and designed the experiments: AP MK BS CS. Performed the experiments: AP MK BS CS. Analyzed the data: AP MK BS CS. Wrote the paper: AP MK BS CS.

References

1. Alexander RA, Barrett GV, Alliger GM, Carson KP. Towards a general model of non-random sampling and the impact on population correlation: Generalizations of Berkson's Fallacy and restriction of range. Br J Math Stat Psychol. 1986; 39: 90–105. doi: 10.1111/j.2044-8317.1986.tb00849.x PMID: 3768267
2. Berk RA. An introduction to sample selection bias in sociological data. Am Sociol Rev. 1983; 48: 386–398. doi: 10.2307/2095230
3. Gross AL, McGanney ML. The restriction of range problem and nonignorable selection processes. J Appl Psychol. 1987; 72: 604–610. doi: 10.1037/0021-9010.72.4.604
4. Sackett PR, Yang H. Correction for range restriction: An expanded typology. J Appl Psychol. 2000; 85: 112–118. doi: 10.1037/0021-9010.85.1.112 PMID: 10740961
5. Thorndike RL. Personnel selection: Test and measurement techniques. New York: Wiley; 1949.
6. Chan W, Chan DW-L. Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. Psychol Methods. 2004; 9: 369–385. doi: 10.1037/1082-989X.9.3.369 PMID: 15355154
7. Li JC, Chan W, Cui Y. Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. Br J Math Stat Psychol. 2011; 64: 367–387. doi: 10.1348/2044-8317.002007 PMID: 21973092
8. Duan B, Dunlap WP. The accuracy of different methods for estimating the standard error of correlations corrected for range restriction. Educ Psychol Meas. 1997; 57: 254–265. doi: 10.1177/0013164497057002005
9. Linn RL, Harnisch DL, Dunbar SB. Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. J Appl Psychol. 1981; 66: 655–663. http://dx.doi.org/10.1037/0021-9010.66.6.655
10. Lord FM, Novick MR, Birnbaum A. Statistical theories of mental test scores. 1968; Available: http://doi.apa.org/psycinfo/1968-35040-000
11. Millsap RE. Sampling variance in the correlation coefficient under range restriction: A Monte Carlo study. J Appl Psychol. 1989; 74: 456–461. doi: 10.1037/0021-9010.74.3.456
12. Ree MJ, Carretta TR, Earles JA, Albert W. Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. J Appl Psychol. 1994; 79: 298–301. doi: 10.1037/0021-9010.79.2.298
13. Schmidt FL, Oh I-S, Le H. Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. Pers Psychol. 2006; 59: 281–305. doi: 10.1111/j.1744-6570.2006.00065.x
14. Wiberg M, Sundström A. A comparison of two approaches to correction of restriction of range in correlation analysis. Pract Assess Res Eval. 2009; 14: 1–9.
15. Wolfe JH, Held JD. Standard errors of multivariate range-corrected validities. Mil Psychol. 2010; 22: 356.
16. Zimmerman DW, Williams RH. Restriction of range and correlation in outlier-prone distributions. Appl Psychol Meas. 2000; 24: 267–280. doi: 10.1177/01466210022031741
17. Linn RL. Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. J Educ Meas. 1983; 20: 1–15. doi: 10.1111/j.1745-3984.1983.tb00185.x

18. Cleary PD, Angel R. The analysis of relationships involving dichotomous dependent variables. J Health Soc Behav. 1984; 334–348. PMID: 6501841
19. Peng C-YJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Educ Res. 2002; 96: 3–14. doi: 10.1080/00220670209598786
20. Ulrich R, Wirtz M. On the correlation of a naturally and an artificially dichotomized variable. Br J Math Stat Psychol. 2004; 57: 235–251. doi: 10.1348/0007110042307203 PMID: 15511306
21. Abrahams NM, Alf EF, Wolfe JJ. Taylor-Russell tables for dichotomous criterion variables. J Appl Psychol. 1971; 55: 449–457. doi: 10.1037/h0031761
22. Bobko P, Roth PL, Bobko C. Correcting the effect size of d for range restriction and unreliability. Organ Res Methods. 2001; 4: 46–61. doi: 10.1177/109442810141003
23. Mendoza JL. Fisher transformations for correlations corrected for selection and missing data. Psychometrika. 1993; 58: 601–615.
24. Enders CK. Applied missing data analysis. New York, NY: Guilford Press; 2010.
25. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.
26. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychol Methods. 2002; 7: 147–177. doi: 10.1037/1082-989X.7.2.147 PMID: 12090408
27. Hunter JE, Schmidt FL, Le H. Implications of direct and indirect range restriction for meta-analysis methods and findings. J Appl Psychol. 2006; 91: 594–612. doi: 10.1037/0021-9010.91.3.594 PMID: 16737357
28. University of Vienna. Aufnahmeverfahren [Admission procedures] [Internet]. 2015 Accessed: 15 April 2015. Available: https://aufnahmeverfahren.univie.ac.at/en/home/
29. Medical University of Vienna. MedAT—Aufnahmeverfahren Medizin [MedAT—admission test medicine] [Internet]. 2015 Accessed: 15 April 2015. Available: http://www.medizinstudieren.at/
30. Medical University of Vienna. Zulassung [Admission] [Internet]. 2015 Accessed: 5 April 2015. Available: http://www.meduniwien.ac.at/homepage/content/studium-lehre/zulassung-administratives/zulassung-zum-studium/diplomstudien-human-und-zahnmedizin/zulassung/
31. Chernyshenko OS, Ones DS. How selective are psychology graduate programs? The effect of the selection ratio on GRE score validity. Educ Psychol Meas. 1999; 59: 951–961. doi: 10.1177/00131649921970279
32. Powers DE. Validity of Graduate Record Examinations (GRE) general test scores for admissions to colleges of veterinary medicine. J Appl Psychol. 2004; 89: 208. http://dx.doi.org/10.1037/0021-9010.89.2.208 PMID: 15065970
33. Kobrin JL, Patterson BF, Shaw EJ, Mattern KD, Barbuti SM. Validity of the SAT® for predicting First-Year College Grade Point Average [Internet]. New York: The College Board; 2008 p. 16. Report No.: 2008–5. Available: https://professionals.collegeboard.com/profdownload/Validity_of_the_SAT_for_Predicting_First_Year_College_Grade_Point_Average.pdf
34. Weitzman RA. The prediction of college achivement by the scholastic aptitude test and the high school record. J Educ Meas. 2005; 19: 179–191. doi: 10.1111/j.1745-3984.1982.tb00126.x
35. Sireci SG, Talento-Miller E. Evaluating the predictive validity of Graduate Management Admission test scores. Educ Psychol Meas. 2006; 66: 305–317. doi: 10.1177/0013164405282455
36. SjöBerg S, SjöBerg A, Näswall K, Sverke M. Using individual differences to predict job performance: Correcting for direct and indirect restriction of range. Scand J Psychol. 2012; 53: 368–373. doi: 10.1111/j.1467-9450.2012.00956.x PMID: 22612634
37. Carretta TR. Pilot Candidate Selection Method. Aviat Psychol Appl Hum Factors. 2011; 1: 3–8. doi: 10.1027/2192-0923/a00002
38. Holmes DJ. The robustness of the usual correction for restriction in range due to explicit selection. Psychometrika. 1990; 55: 19–32. doi: 10.1007/BF02294740
39. Gross AL, Fleischman L. Restriction of range corrections when both distribution and selection assumptions are violated. Appl Psychol Meas. 1983; 7: 227–237. doi: 10.1177/014662168300700210
40. Society for Industrial Organizational Psychology (SIOP). Principles for the validation and use of personnel selection procedures [Internet]. 4th ed. Bowling Green, OH: Society for Industrial Organizational Psychology (SIOP); 2003. Available: http://www.siop.org/_Principles/principles.pdf
41. Raju NS, Brand PA. Determining the significance of correlations corrected for unreliability and range restriction. Appl Psychol Meas. 2003; 27: 52–71. doi: 10.1177/0146621602239476
42. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007; 8: 206–213. doi: 10.1007/s11121-007-0070-9 PMID: 17549635

43. Rubin DB. Inference and missing data. Biometrika. 1976; 63: 581–592. doi: 10.1093/biomet/63.3.581
44. Von Hippel PT. Biases in SPSS 12.0 Missing Value Analysis. Am Stat. 2004; 58: 160–164. doi: 10.1198/0003130043204
45. Van Buuren S. Flexible Imputation of Missing Data. CRC Press; 2012.
46. Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 2004.
47. Schafer JL. Analysis of incomplete multivariate data. 1st ed. New York: Chapman and Hall/CRC Press; 1997.
48. Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. J Stat Comput Simul. 2008; 78: 69–84. doi: 10.1080/10629360600903866
49. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. CRC Press; 1989.
50. Kemery ER, Dunlap WP, Griffeth RW. Correction for variance restriction in point-biserial correlations. J Appl Psychol. 1988; 73: 688–691. doi: 10.1037/0021-9010.73.4.688
51. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002; 7: 19–40. doi: 10.1037//1082-989X.7.1.19 PMID: 11928888
52. R Core Team. A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available: http://www.R-project.org/
53. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer Science & Business Media; 2002.
54. Cohen AC. Estimation in mixtures of two normal distributions. Technometrics. 1967; 9: 15–28. doi: 10.2307/1266315
55. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. J Stat Softw. 2011; 45. Available: http://doc.utwente.nl/78938/ PMID: 22289957
56. Ayyub BM, McCuen RH. Probability, statistics, and reliability for engineers and scientists. Boca Raton, FL: CRC press; 2011.
57. Heckman JJ. Sample selection bias as a specification error. Econom J Econom Soc. 1979; 47: 153–161. doi: 10.2307/1912352
58. Heckman JJ. The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Ann Econ Soc Meas. 1976; 5: 475–492.
59. Muthén BO, Hsu J-WY. Selection and predictive validity with latent variable structures†. Br J Math Stat Psychol. 1993; 46: 255–271.

3rd Publication

Pfaffel, A. & Spiel, C. (2016). *Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study*. Manuscript submitted for publication.

Contribution by the individual authors:

Pfaffel, A.	Conception and design, R programming for the MC simulations, statistical analysis and interpretation of data, creation of graphics, drafting the paper
Spiel, C.	Revising the paper critically for important intellectual content, supervision of manuscript preparation

Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study

Andreas Pfaffel, Christiane Spiel
University of Vienna

Author Note
Andreas Pfaffel and Christiane Spiel
Department of Applied Psychology: Work, Education, Economy, Faculty of Psychology, University of Vienna, Austria.

Correspondence concerning this article should be addressed to Andreas Pfaffel, Faculty of Psychology, Department of Applied Psychology: Work, Education, Economy, University of Vienna, Vienna, Austria. Universitaetsstrasse 7, 1010 Vienna, Austria.
E-mail: andreas.pfaffel@univie.ac.at

Abstract

Approaches to correcting correlation coefficients for range restriction have been developed under the framework of large sample theory. The accuracy of missing data techniques for correcting correlation coefficients for range restriction has thus far only been investigated with relatively large samples. However, researchers and evaluators are often faced with a small or moderate number of applicants but must still attempt to estimate the population correlation between predictor and criterion. Therefore, in the present study we investigated the accuracy of population correlation estimates and their associated standard error in terms of small and moderate sample sizes. We applied multiple imputation by chained equations for continuous and naturally dichotomous criterion variables. The results show that multiple imputation by chained equations is accurate for a continuous criterion variable, even for a small number of applicants when the selection ratio is not too small. In the case of a naturally dichotomous criterion variable, a small or moderate number of applicants leads to biased estimates when the selection ratio is small. In contrast, the standard error of the population correlation estimate is accurate over a wide range of conditions of sample size, selection ratio, true population correlation, for continuous and a naturally dichotomous criterion variables, and for direct and indirect range restriction scenarios. The findings of this study provide empirical evidence about the accuracy of the correction, and support researchers and evaluators in their assessment of conditions under which correlation coefficients corrected for range restriction can be trusted.

Keywords: correlation analysis; range restriction; predictive validity; missing data; multiple imputation; standard error

Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study

In psychometrics, it is well known that estimating predictive validity based on selected samples leads to biased population estimates, which is known as the range restriction problem. The correlation between a predictor (e.g., scores on an aptitude test, assessment center, or interview) and a criterion of success (grades, achievement scores, or graduation status) obtained from the selected sample typically underestimates the correlation in the applicant population, i.e. it underestimates the predictive validity. This problem arises because the selected sample is not random and therefore not representative of the applicant population (Sackett & Yang, 2000). Researchers and evaluators are often faced with a moderate or a small number of applicants but must still attempt to evaluate the predictive validity of a selection method. Such samples cause problems in terms of the accuracy of the population estimate and in examining its statistical significance because sample size is an important factor affecting the accuracy of a parameter estimate. This problem becomes worse in cases of selection because population estimates are based on only a subsample of applicants, i.e. on the available selected sample.

Researchers have proposed two approaches to correct correlation coefficients for range restriction. The traditional approach is to use the correction formulas presented by Thorndike (1949) based on earlier works by Pearson (1903), Aitkin (1935), and Lawley (1943). In the psychometric literature, it is well documented that the corrected Pearson product-moment correlation coefficients are less biased than uncorrected correlation coefficients even over a wide range of assumption violations (Greener & Osburn, 1979; Gross & Fleischman, 1983; Holmes, 1990; Linn, 1983; Linn, Harnisch, & Dunbar, 1981; Ree, Carretta, Earles, & Albert, 1994). The modern approach is to view the selection as a missing data mechanism (Pfaffel, Schober, & Spiel, 2016; Mendoza, 1993; Wiberg & Sundström, 2009). This approach offers some advantages over the correction formulas. Recent simulation studies show that state-of-the-art missing data techniques such as full information maximum likelihood estimation (FIML) and multiple imputation (MI) are equally or under some conditions more accurate than the traditional correction formulas (Pfaffel, Kollmayer, Schober, & Spiel, 2016; Pfaffel, Schober, et al., 2016).

Both approaches, the correction formulas and the missing data approach, have been derived and justified in terms of large sample theory, which is a generic framework for assessing the properties of statistical estimators as sample size grows indefinitely (Lehmann, 1999).

Although multiple imputation and full information maximum likelihood estimation make the same assumptions, simulation studies suggest that multiple imputation performs better than maximum likelihood estimation with small or moderate sample sizes (Graham & Schafer, 1999; Little & Rubin, 1989). The accuracy of the missing data techniques have been investigated so far only with relatively large samples (Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, et al., 2016). Investigations in small and moderate samples are missing. Therefore, it is questionable whether missing data techniques are able to correct correlation coefficients for range restriction in small or moderate samples. Additionally, correction methods have been widely studied for continuous criterion variables but little is known about range restriction correction when the criterion is dichotomous. In particular, there is a lack of studies considering the standard error. To the best of our knowledge, no empirical study has investigated so far the accuracy of the multiple imputation standard error of the population correlation estimate in the case of range restriction. Therefore, the purpose of the present study is to investigate both the accuracy of the range restriction correction and the accuracy of the associated standard error when the sample size is small or moderate. We apply a Bayesian multiple imputation technique for both continuous and naturally dichotomous criterion variables. 'Naturally' means the dichotomous criterion has no underlying continuous distribution (Ulrich & Wirtz, 2004).

We first describe the two most common range restriction scenarios (direct and indirect range restriction) for both a continuous criterion variable and a dichotomous one. We then give a brief overview of approaches to correcting for range restriction with a focus on missing data techniques. After that, we give a brief introduction to calculating the standard error in the case of missing values under the framework of maximum likelihood estimation and multiple imputation. Finally, we investigate the accuracy of multiple imputation by chained equations under various conditions with a focus on the sample size by conducting several Monte Carlo simulations.

Range restriction in the case of a continuous and a dichotomous criterion

Direct and indirect range restriction are the two most common scenarios in the selection of applicants. In a direct range restriction scenario (DRR), the selection is based directly on the predictor X, whereas X can be either a score from a single selection method or a composite score derived from several selection methods, e.g. an aptitude test, an assessment center, and an interview (Pfaffel, Schober, et al., 2016). In a DRR scenario, we are interested in the predictive validity of the variable used for the selection. For example, this is the case if we want to assess

the predictive validity of a selection method or of an entire selection procedure, which is based on several selection methods. In contrast, in an indirect range restriction scenario (IRR), selection is based on another variable *Z*, which is usually correlated with *X*, *Y*, or both. In an IRR scenario, we are interested in the predictive validity of a selection method *X* (the predictor of interest), which is not the selector *Z*. *Z* can either be a single selection method or a combination of selection methods, possibly but not necessarily including *X* (Linn et al., 1981). For example, this is the case if scores on another selection method or a composite score are used for the selection, but we want to assess the predictive validity of a certain selection method *X*.

The predictive validity of *X*, or more precisely the correlation between a predictor *X* and a criterion of success *Y* is a measure of the effectiveness of the selection. The higher the correlation between *X* and *Y*, the smaller the prediction error of the criterion values. However, the correlation between *X* and *Y* can only be obtained from the selected sample. Due to the selection itself, values of the criterion are not available for non-selected applicants. Figure 1 illustrates the loss of criterion data for DRR and IRR scenarios in the case of a continuous criterion variable. Figure 1a shows the complete data in which the (unrestricted) Pearson population correlation ρ_{XY} is .50. Figure 1b and 1c illustrates the effects of selection on *X* and *Z*, respectively. The blue data points are the available selected sample, the gray data points represents the non-selected sample in which the values for *Y* are missing. In both scenarios, the selection ratio is .40, which is the ratio of the number of selected individuals to the number of applicants. Figure 1b shows that the top 40% of applicants are selected while 60% are not selected. Applicants with scores below a specific value of *X* are thus excluded from the sample. It is clear that the scores of *X* in the selected sample are restricted in range. Consequently, the Pearson correlation coefficient obtained from the selected sample $r_{XY} = .23$ is significantly smaller than in the complete dataset. The correlation coefficient obtained from the selected sample underestimates the true population correlation.

Figure 1c shows an IRR scenario in which the loss of criterion data is based on another variable *Z*. In this example, *Z* is correlated with *X* and *Y* at .50, respectively, and the top 40% of applicants with respect to *Z* are selected. Consequently, the Pearson correlation coefficient obtained from the selected sample is $r_{XY} = .38$. The effect on correlations due to selection on *Z* is typically weaker than in the case of selection on *X* (Sackett & Yang, 2000). Levin (1972) showed that it is theoretically possible that selection on *Z* can increase rather than decrease the correlation coefficient when the correlations of *Z* with *X* and with *Y* become extreme. However, this effect is

rarely encountered in real datasets, meaning that selection on *Z* can be expected to reduce the magnitude of the correlation coefficient (Linn et al., 1981).

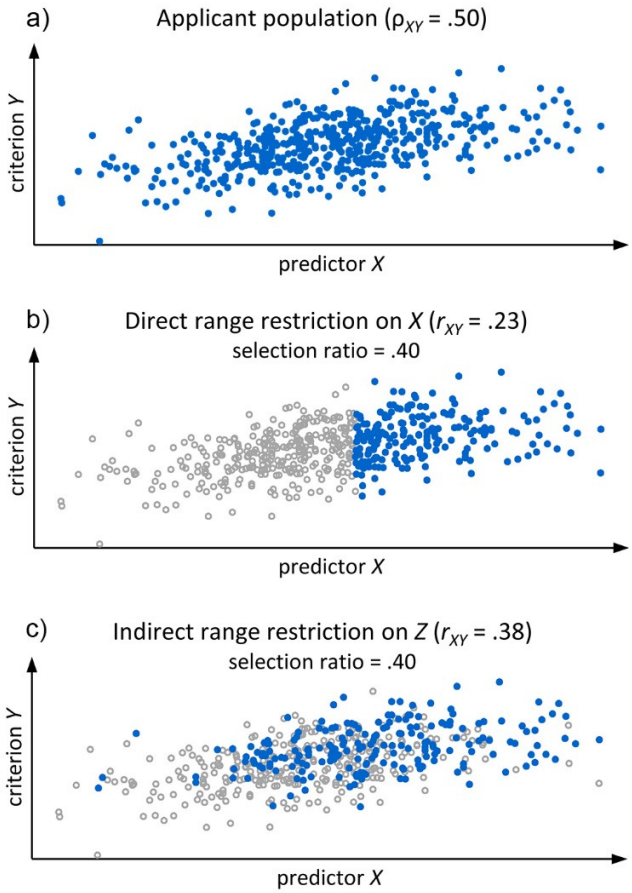

Figure 1. An illustration of the loss of criterion data for direct and indirect range restriction scenarios in the case of a continuous criterion variable.

A closer look at the problem shows that the effect on the Pearson correlation coefficient does not stem directly from the restriction in range of *X*, but as a result of the reduction of the sample variances of *X* and *Y* as well as by the reduction of the sample covariance between *X* and *Y* in the selected sample. The problem arises from the formula of the Pearson correlation coefficient (Equation 1). The reduction of r_{XY} is given as the reduction in the sample covariance (the numerator) relative to the reduction in the product of the sample standard deviations s_X and s_Y (the denominator).

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

Next, we look at direct and indirect range restriction scenarios and the loss of criterion data in the case of a dichotomous criterion variable. So far only a few studies have focused on range restriction correction in the case of a dichotomous criterion variable (Bobko, Roth, & Bobko, 2001; Pfaffel, Kollmayer, et al., 2016; Raju, Steinhaus, Edwards, & DeLessio, 1991). Figure 2 shows that the criterion *Y* is divided into two groups ('not successful' and 'successful'). The correlation coefficient used to express the relationship between a continuous and a naturally dichotomous variable is the point-biserial correlation coefficient r_{pb} (Ulrich & Wirtz, 2004), which is calculated by

$$r_{\mathrm{pb}XY} = \frac{(M_1 - M_0)\sqrt{pq}}{s_X} \tag{2}$$

where M_1 and M_0 are the mean values of the continuous variable *X* for the two groups *p* ('not successful', $Y = 0$) and *q* ('successful', $Y = 1$), and s_X is the standard deviation of the continuous variable *X*. Figure 2a shows the complete data in which the unrestricted point-biserial correlation coefficient $\rho_{\mathrm{pb}XY}$ is .50. Figures 2b and 2c illustrate the effects due to selection on *X* and *Z* on the point-biserial correlation coefficient. In both scenarios, the selection ratio is 40%. In a DRR scenario, as shown in Figure 2b, applicants with scores below a specific value of *X* have been excluded from the sample. Consequently, $r_{\mathrm{pb}XY}$ obtained from the selected sample is .27. Figure 2c shows an IRR scenario in which the top 40% applicants with respect to *Z* have been selected; *Z* is correlated with *X* and *Y* at .50, respectively. In the case of IRR, we obtain a value for $r_{\mathrm{pb}XY}$ of .40.

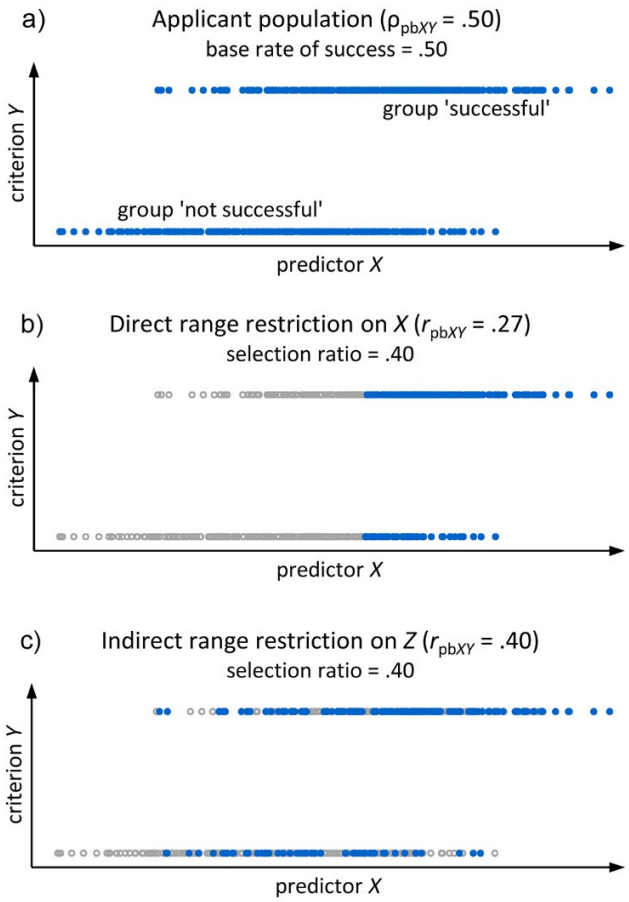

Figure 2. An illustration of the loss of criterion data for direct and indirect range restriction scenarios in the case of a dichotomous criterion variable.

Range restriction in the case of a dichotomous criterion variable is similar to range restriction scenarios in the case of a continuous criterion variable. However, a very important factor that has to be considered additionally is the base rate of success BR (Abrahams, Alf, & Wolfe, 1971; Pfaffel, Kollmayer, et al., 2016). The BR is the percentage of applicants who would be successful on the criterion if there were no selection, and is calculated by dividing the number of successful individuals by the number of applicants. The BR ranges from 0 to 1, or from 0% to 100%. For example, if all applicants were to be admitted to a study program and 50% percent of them complete this program, then the BR is 50%. In our examples in Figure 2, we used a BR of 50%. The BR is closely related to the effectiveness of the selection, because a selection is considered effective when the percentage of successful applicants (in the selected sample) is higher than the BR, i.e. when the selected applicants are more frequently successful than would be the case by random chance. It is not surprising that when the BR is high, the probability of

gaining an effective selection is low. In such a case, the incremental predictive validity of additional and resource-intensive selection methods should be examined. Thus, the BR also plays a role in assessing the efficiency of a selection method.

Unfortunately, the BR is unknown in the case of selection as is unbiased information about the proportion of successful individuals in the applicant population. We can only obtain the *success rate* from the selected sample, which is a biased estimator for the BR. The success rate is the number of successful individuals divided by the number of selected applicants. In Figures 2b and 2c, the success rate is 0.75. This success rate is higher than the BR, because the relationship between predictor and criterion is positive. Hence, more applicants who would be successful have been selected.

In addition to the range restriction effect, the magnitude of the observed (restricted) point-biserial correlation coefficient is also affected by variance-restriction due to unequal *p*-*q*-split (Kemery, Dunlap, & Griffeth, 1988). The variance of a dichotomous variable is the product of p and q with a maximum value of .25 at $p = q = 0.5$. If p does not equal q, the variance will be less than .25. Consequently, r_{pb} decreases as p and q move away from .50, and increases as p and q move towards .50. The two effects can sometimes act in opposite directions. For example, assume that ρ_{pb} is positive, the BR is .10, and after selection, the observed success rate is .60. Because .60 is closer to 0.50 than 0.10, the variance of the dichotomous variable in the selected sample is higher than in the population, and this consequently leads to an increase in r_{pb}. In such a case, r_{pb} decreases due to range restriction and increases due to the *p*-*q* split. Despite range restriction, it is conceivable that r_{pb} is not much smaller than ρ_{pb} because of the combination of the two effects. Therefore, correction methods for range restriction must take into account the effect of the *p*-*q* split in the case of a dichotomous criterion variable.

Approaches to correcting for direct and indirect range restriction scenarios

Researcher have proposed two approaches to correct correlations for direct and indirect range restriction scenarios. The traditional approach is to apply the correction formulas presented by Thorndike (1949) based on earlier works by Pearson (1903), Aitkin (1935), and Lawley (1943). The formulas correct the Pearson correlation coefficient for univariate direct and indirect range restriction scenarios for continuous variables. They were derived within the framework of maximum likelihood estimation under the assumptions of multivariate normality, linearity between X and Y, and homoscedasticity. In the psychometric literature, it is well documented that

corrected Pearson correlations are less biased than uncorrected correlations over a wide range of assumption violations (Greener & Osburn, 1979; Gross & Fleischman, 1983; Holmes, 1990; Linn, 1983; Linn et al., 1981; Ree et al., 1994). The corrected Pearson correlations are always higher than the uncorrected correlations. The formulas include only the variables *X* and *Y*, or *X*, *Y*, and *Z*, where *X* and *Z* must have no missing values. Covariates that could potentially contribute to the prediction of *Y* are not considered.

The modern approach is to view the selection mechanism as a missing data mechanism (Mendoza, 1993; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, et al., 2016; Wiberg & Sundström, 2009). Rubin (1976) identified three missing data mechanisms, according to the underlying cause of missing data. These mechanisms are important since they are necessary assumptions for the missing data methods: Missing completely at random (MCAR) means the probability of missing values of *Y* is unrelated to other measured variables and to the values of *Y* itself. Missing at random (MAR) means the probability of missing values of *Y* is related to other measured variables, but not related to the values of *Y* itself. Missing not at random (MNAR) means the probability of missing values of *Y* is related to the values of *Y* itself, even after controlling for other variables. The missing data mechanism in both range restriction scenarios (DRR and IRR) is missing at random (MAR) because the missing values depend either on *X* or *Z*, but not on the values of *Y* itself (Pfaffel, Schober, et al., 2016).

The missing data approach has several advantages over the correction formulas: 1) This approach no longer requires a distinction between DRR and IRR to be made in applying the correction because both scenarios are considered to be MAR and the same techniques can be used to correct for both range restriction scenarios. 2) State-of-the-art missing data techniques such as full information maximum likelihood and multiple imputation can handle multivariate datasets with multiple covariates and 3) can also handle different types of predictor and criterion variables (e.g., dichotomous, unordered and ordered categorical, continuous). Pfaffel, Kollmayer, and colleagues (2016) showed that the correction using multiple imputation by chained equations is more accurate than Thorndike's (1949) correction formulas when the criterion variable is dichotomous, especially in the case of IRR. 4) In contrast to Thorndike's formulas, covariates – but not the selection variable – may have some missing values (missing values in the selection variable is MNAR). However, no empirical studies have been presented, which investigate the effect of covariates with missing values on the accuracy of the correction.

Methodologists currently regard full information maximum likelihood and multiple imputation as state of the art when dealing with missing data. Techniques such as listwise or pairwise deletion, arithmetic mean imputation, single regression imputation, or single EM imputation are no longer considered state-of-the-art because they have potentially serious drawbacks (Enders, 2010). For example, arithmetic mean imputation imputes values that fall directly on a horizontal line. Consequently, the correlations between imputed values and other variables is zero for the subset of cases with imputed values. Arithmetic mean imputation attenuates correlations and covariances. In single regression imputation, the imputed values fall directly on the (straight) regression line, which overestimates correlations and covariances. This under- and overestimation of correlations is present under any missing data mechanism, including MCAR, and increases as the missing data rate increases (Enders, 2010). In addition, single imputation techniques attenuate standard errors. Neither state-of-the-art technique, full information maximum likelihood and multiple imputation, suffers from the problems mentioned for deletion of incomplete cases and single imputation techniques (Enders, 2010).

Full information maximum likelihood (FIML) is a technique of finding population parameters by maximizing the log-likelihood function that has the highest probability of producing the data of a certain sample. FIML requires the missing data mechanism to be either MAR or MCAR. Finding the parameter values that maximize the log-likelihood function is possible with iterative optimization algorithms such as expectation maximization (EM) algorithms (Dempster, Laird, & Rubin, 1977; Meng & Rubin, 1993). In the social and behavioral sciences, population data is commonly assumed to be multivariate normally distributed (Enders, 2010). Dealing with identically distributed variables is straightforward and many software packages can handle missing values under the condition of multivariate normality. FIML estimation with non-identically distributed variables in multivariate datasets is much more complicated, for example in logistic regression analysis. FIML with complex multivariate incomplete data is typically only possible with structural equation modeling (SEM) software, e.g. Mplus (Muthén & Muthén, 2015), or the lavaan package for R Statistics (Rosseel, 2012). For a detailed description of likelihood-based techniques, see Little and Rubin (2002), or for a less technical description see Enders (2010).

Multiple imputation (MI), proposed by Rubin (1978), is another state-of-the-art technique for handling missing values that allows the data analyst to use statistical methods designed for complete data. In contrast to FIML, MI creates plausible estimates for the missing values. MI and

FIML makes the same assumptions regarding the missing data mechanism (MAR or MCAR), their estimators have similar statistical properties (e.g., consistency, asymptotic normality), and they frequently produce equivalent results (Enders, 2010; Graham, Olchowski, & Gilreath, 2007). A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. The imputation phase creates *m* complete datasets (e.g., *m* = 20 imputations) based on one dataset with missing values. Each of these *m* complete datasets contains different plausible estimates of the missing values, but the observed values are identical. In contrast to a single imputation technique, the created *m* complete datasets reflect the uncertainty of the missing data. Thus, the imputed values do not fall on the regression line. Consequently, MI do not attenuate correlations and covariances. In the analysis phase, each complete dataset is analyzed with conventional statistical methods, e.g. *m* correlation analyses. Finally, the pooling phase combines the *m* parameter estimates into a single set of parameters, e.g. *m* correlation coefficients are combined into one pooled value. The pooled parameter values are typically the arithmetic average of the *m* estimates generated in the analysis phase (Rubin, 2004). Analyzing and pooling a large number of imputed datasets sound laborious, but modern MI software packages automate this procedure.

Handling incomplete multivariate normal data is possible with the data augmentation algorithm (Schafer, 1997; Tanner & Wong, 1987). A general multiple imputation technique, which can handle incomplete datasets with not necessarily normal or non-identically distributed variables 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 MICE algorithm, for example, is implemented in the R software package *mice* (van Buuren & Groothuis-Oudshoorn, 2011).

The multivariate imputation model is specified on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable (van Buuren & Groothuis-Oudshoorn, 2011). In the case of an incomplete dichotomous variable such as our example in Figure 2, multiple imputation is possible using a logistic regression model, which incorporates the parameter uncertainty (Pfaffel, Kollmayer, et al., 2016; van Buuren, 2012). Typically, all variables or many variables in the dataset are part of the imputation model used to generate the plausible estimates of the missing values. Because MI clearly separates the imputation and the analysis phase, the analysis model can differ from the imputation model. Therefore, the plausible estimates contain information on variables that might not be included in the analysis model.

However, the imputation model has to be more general then the analysis model. For example, a common source for incompatibility occurs when the analysis model contains interactions and non-linearities, but the imputation model did not. Recent simulation studies show that correction for DRR and IRR with full information maximum likelihood or multiple imputation by chained equations is equally or more accurate compared to Thorndike's (1949) correction formulas in the case of multivariate normality (Pfaffel, Schober, et al., 2016) and in the case of an artificially or naturally dichotomous criterion variable (Pfaffel, Kollmayer, et al., 2016). Especially in the case of IRR, correction with a missing data technique is more precise and therefore more efficient than the formulas. Full information maximum likelihood and multiple imputation by chained equations produce equal parameter estimates. Because of these empirical findings and the advantages mentioned above, we recommend the use of missing data techniques to correct for range restriction.

Standard error of correlations corrected for range restriction

Estimating the standard error and confidence intervals of correlation coefficients in the case of missing data is often much more complex than with complete datasets. In this section, we first give a brief overview of calculating the standard error of correlation coefficients in the case of complete samples and in the case of missing data. Next, we present some approaches for estimating the standard error of correlation coefficients corrected for direct and indirect range restriction scenarios. Finally, we show that multiple imputation allows for calculating the standard error and confidence intervals of correlation coefficients very similar to complete datasets.

In statistics, it is well known that an unbiased estimator converges in probability to the true quantity being estimated as the sample size goes to infinity (property of consistency). This means that the sampling error of a sample parameter becomes smaller and smaller as the sample size increases, and is zero when the sample size is infinitely large. Conversely, if the sample size decreases, then the sampling error of the estimate increases, making the estimation less precise. A common measure of the variability of the sampling distribution is the standard error *SE*. The standard error is used for calculating confidence intervals for a parameter estimate in hypothesis testing. A larger standard error is less likely to reject the null hypothesis. The sampling distribution of the Pearson correlation coefficient r_{XY} (Equation 1) is quite complex even under bivariate normality, and r_{XY} is a negatively biased estimator of ρ_{XY} (Olkin & Pratt, 1958).

However, this bias is small and decreases as the sample size increases. Thus, the true *SE* is not easy to calculate and only valid when the underlying assumptions are fully met. For complete data analysis, Kendall and Stuart (1977) proposed an approximation of *SE* of the sample Pearson correlation coefficient r_{XY} in samples with size *N*:

$$SE(r_{XY}) \approx \frac{1-r_{XY}^2}{\sqrt{N-1}} \quad (3)$$

Equation 3 shows that the *SE* of r_{XY} depends on the sample size and on the value of r_{XY} itself. Consequently, with the same sample size, a stronger correlation can be estimated more accurately than a weaker one. The procedure for examining the statistical significance and the asymmetric confidence interval of r_{XY} is to transform r_{XY} into a Fisher *z*-value with the associated standard error (Fisher, 1915):

$$SE_{Fisher\ z}(r_{XY}) = \frac{1}{\sqrt{N-3}} \quad (4)$$

As we can see, this standard error depends only on the sample size. Fisher's *z*-transformation is necessary because the sampling distribution of r_{XY} is skewed and correlation coefficients are only supported on the bounded interval [-1,1]. An asymmetric sampling distribution leads to asymmetric confidence intervals. Fisher's *z*-transformation can also be applied to the point-biserial correlation coefficient because r_{pb} is mathematically equivalent to the Pearson correlation coefficient.

Next, we want to show what happens to the *SE* in the case of range restriction, i.e. in the case of missing values. Many researchers have demonstrated that the *SE* of a corrected Pearson correlation coefficient is larger than for the uncorrected correlation coefficient (Bobko & Rieck, 1980; Mendoza, 1993; Millsap, 1989; Raju & Brand, 2003). The increase in the magnitude of standard error can be explained by considering two circumstances: First, the correlation coefficient is measured in a subsample with sample size $n \leq N$, where *n* is the size of the selected sample. As shown in Equation 3 and 4, the magnitude of the standard error is approximately inversely proportional to square root of the sample size. Second, the population correlation estimate must include the uncertainty caused by the proportion of missing values. Consequently, the magnitude of the standard error increases by applying corrections for range restriction.

Bobko and Rieck (1980) presented a large sample estimator for the standard error of correlation coefficients corrected for a direct range restriction scenario. The estimator is the product of the standard error of the correlation coefficient obtained from the selected sample and

a factor derived from Thorndike's (1949) formula for direct range restriction scenarios. In case of indirect range restriction scenarios, a large sample estimator has been presented by Allan and Dunbar (1990), but this formula is very long and complicated. As shown for complete datasets, Fisher's *z*-transformation can be applied to calculate the confidence interval of correlation coefficients. However, Mendoza (1993) showed that Fisher's *z*-transformation cannot directly applied to correlations corrected for direct and indirect range restriction scenarios (assumption MAR), and proposed additional correction terms to the Fisher's *z*-transformation. Admittedly computers allow to easily calculate these formulas. However, this examples show that deriving the sampling distribution under the framework of maximum likelihood estimation, especially in the case of missing data, often leads to complex problems relatively quickly. In summary, it can be ascertained that deriving the sampling distribution of the sample correlation coefficient under the framework of maximum likelihood estimation is very complex or maybe sometimes impossible in the case of multivariate distributions with missing data.

In contrast, calculating the standard error using multiple imputation is relatively straightforward. A major advantage is that conventional statistical procedures to calculate the standard error can be applied to the *m* complete datasets. Moreover, the correlation is calculated based on the total sample size *N*, and therefore making Fisher's *z*-transformation much more accurate. Multiple imputation standard errors combine two sources of uncertainty regarding the parameter estimate (Little & Rubin, 2002): The uncertainty within an imputation (the within-imputation variance), and the uncertainty between the *m* imputations (the between-imputation variance). The Fisher's z standard error of one of the *m* complete datasets represents the uncertainty of the data. The increase in Fisher's standard error in the case of missing data results from the between-imputation variance. The parameter estimates and the standard errors can be combined by Rubin's rules (Rubin, 2004). The Appendix shows the equations for computing the estimate of the correlation coefficient, its associated standard error, and the confidence interval for multiple imputed datasets. However, software packages typically implement these procedures, so there is usually no need to compute the parameter estimates by hand.

As mentioned above, multiple imputation and full information maximum likelihood make the same assumptions and have similar statistical properties. The statistical theory underlying these techniques is based partly on large-sample approximations. However, this statement must be restricted because the two missing data techniques differ in their performance in the case of small sample sizes. Simulation studies show that maximum likelihood estimation is inadequate

for small or moderate sample sizes and is likely to result in biased estimates (Graham & Schafer, 1999; Little & Rubin, 1989). The findings suggest that multiple imputation performs more efficiently with small samples. Graham and Schafer (1999, p. 26) pointed out that "limitations of analysis with small sample size lie in the small sample size itself, not with the multiple-imputation procedure". This finding is fundamental for empirical evaluation studies of the predictive validity of selection methods. On the one hand, it supports the use of multiple imputation in small or moderate samples. On the other hand, it makes clear that multiple imputation cannot compensate for having a small number of applicants or small selection ratios. However, multiple imputation allows for the most effective usage of all the data that have been collected.

Therefore, we suggest using a Bayesian multiple imputation technique such as multiple imputation by chained equations to overcome the range restriction problem in small or moderate samples. So far, simulation studies investigating the accuracy of this missing data technique when the sample size is small or moderate are lacking. Additionally, little is known about the correctness of the multiple imputation standard error in the case of range restriction. Our intention is to close these research gaps. The accuracy of the corrected correlation coefficient and of the multiple imputation standard error are important considerations for researchers and evaluators. Thus, our empirical findings will help to increase understanding of the circumstances (e.g. sample size, selection ratio, true population correlation) under which range restriction corrections are appropriate.

Purposes of this study

The first purpose is to examine the accuracy of the population correlation estimates by using multiple imputation by chained equations in terms of small and moderate sample sizes for direct and indirect range restriction scenarios, and for continuous and naturally dichotomous criterion variables. The second purpose is to examine the accuracy of the associated multiple imputation standard error.

Method

We conducted several Monte Carlo simulations to examine the accuracy of the proposed missing data approach and the sampling distribution of the multiple imputation standard error

under different model conditions. Multivariate data were simulated in order to investigate four scenarios: DRR and IRR scenarios with a continuous criterion variable, and DRR and IRR scenarios with a dichotomous criterion variable. Additionally, three factors (continuous criterion) and four factors (dichotomous criterion) that affect the accuracy of the correction as well as the sampling distribution of the standard error were systematically manipulated.

Factor 1: Total sample size, N. Multiple imputation was developed under the framework of large sample theory. In contrast to maximum likelihood estimation, multiple imputation seems to promise a more accurate correction when the sample size is small or moderate. As shown in Equations 3 and 4, sample size also affects the standard error of the correlation coefficient. Therefore, sample size is a very important factor in studying asymptotic estimates. Two different sample sizes were investigated: a small sample with size $N = 50$, and a moderate sample with size of $N = 100$.

Factor 2: Population correlation, ρ_{XY} and ρ_{pb}. The effect of the population correlation on the accuracy of the correction has been documented in a number of empirical studies (Duan & Dunlap, 1997; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, et al., 2016). The correction to the correlation coefficient becomes more precise as the population correlation increases. This effect is valid for DRR and IRR and for continuous and dichotomous criterion variables. As shown in Equation 3, the standard error of the correlation coefficient depends on the magnitude of the correlation coefficient itself and decreases as the correlation coefficient increases. Hence, in the present study, we investigated three levels of ρ_{XY} and $\rho_{\mathrm{pb}XY}$: .20, .40, and .60. According to Cohen's (1988) classification of correlation coefficients in the social sciences, these values represent a small, medium, and large association between predictor and criterion, i.e. a small, medium, and large predictive validity.

Factor 3: Selection ratio, SR. The selection ratio is the ratio of the number of selected applicants to the total sample size *N*. The selection ratio directly affects the proportion of missing values in the criterion variable, and therefore the accuracy of the correction. Correlation estimates become more biased and exponentially less precise when the selection ratio decreases (Pfaffel, Schober, et al., 2016). It is to be expected that this adverse effect increases, when sample sizes become small or moderate. Hence, in the present study, we investigated four levels of the selection ratio: 20%, 30%, 40%, and 50%. The smallest selection ratio of 20% corresponds to subsample sizes of $n = 10$ ($N = 50$) and $n = 20$ ($N = 100$). These two sample sizes have to be considered extremely small because on the one hand, 80% of the criterion values have been

systematically excluded, and on the other hand, 10 and 20 observations are small even for complete data analysis.[1]

Factor 4: Base rate of success, BR. This factor was used in the case of a dichotomous criterion variable. As described above, the effect of range restriction and the effect of variance restriction can sometimes work in opposite directions when the *p*-*q* split is closer to .50 in the selected sample than in the unrestricted sample. Therefore, the effect of the BR is an especially relevant factor to investigate when the criterion variable is dichotomous. In the present study, we varied the BR at three levels: 20%, 50%, and 80%. These three levels represent a small, medium, and large proportion of applicants who would be successful if there were no selection.

Monte Carlo simulation procedure

The Monte Carlo simulations were conducted using the program R (R Core Team, 2016) with 5,000 iterations for each factor combination of sample size, population correlation, selection ratio, and base rate of success. In the case of a continuous criterion variable, there were $2 \times 3 \times 4 = 24$ factor combinations, while in the case of a dichotomous criterion variable, $2 \times 3 \times 4 \times 3 = 72$ factor combinations were investigated. For each of the four scenarios (DRR & IRR × continuous & dichotomous), a random sample with size *N* was generated from a multivariate distribution (see Data Simulation) with a population correlation ρ or ρ_{pb} between predictor *X* and criterion *Y*, and a base rate of success in the case of a dichotomous criterion variable. Then, we simulated the selection by isolating those $n = N \cdot \mathrm{SR}$ cases with the highest values in *X* in the case of a DRR scenario, and in descending order by the third variable *Z* in the case of an IRR scenario. Values of *Y* for non-selected cases were converted into missing values. The selected samples created in this way with $N - n$ missing values in *Y* were used in applying the correction. Next, we used the R package `mice` (multivariate imputation by chained equations, Version 2.25, van Buuren & Groothuis-Oudshoorn, 2011) to generate $m = 20$ imputed datasets. Pfaffel, Kollmayer and colleagues (2016) showed that 20 imputations are sufficient for DRR and IRR corrections using multiple imputation by chained equations. We used the elementary imputation method 'norm' for the imputation of the continuous criterion variable, and the method 'logreg' for the imputation of the dichotomous criterion variable. Finally, the pooled correlation

[1] In a preliminary experiment, we also tested a selection ratio of 10%, but frequent convergence problems led to invalid estimates. In the case of a dichotomous criterion, *Y* was almost always constant in the subsample.

coefficients and the multiple imputation standard errors were calculated using Fisher's *z*-transformation and Rubin's (2004) rules for combining multiple imputation parameter estimates (for details, see the Appendix).

Pfaffel, Kollmayer, and colleagues (2016) reported problems (e.g. constancy of *Y* in the selected sample) in conducting a logistic regression analysis for some factor combinations, especially when the base rate of success and the population correlation were high and the selection ratio was small. They excluded selected samples (with minimum sample size of $n = 50$) with less than five observations in each of the two criterion groups. In the present study, the smallest *n* was 10 when the total sample size *N* was 50 and the selection ratio .20. Requiring at least five observations in each criterion group means that only one *p*-*q* split of 50% in the selected sample is valid for $n = 10$. Thus, there would be no variability in the *p*-*q* split for this factor combination. Consequently, we weakened the prerequisite to at least three observations in each of the two criterion groups.

Data simulation

We simulated multivariate data for a) a normally distributed (continuous) criterion variable and b) for a naturally dichotomously distributed criterion variable. In simulating the multivariate data, we used the procedures presented in the studies by Pfaffel, Schober, et al. (2016) and Pfaffel, Kollmayer, et al. (2016).

a) Continuous criterion: We generated a bivariate (DRR) and a trivariate (IRR) standard normal distribution with Pearson population correlations between *X* and *Y* of .20, .40, and .60 using the `mvrnorm` function of the MASS package (Venables & Ripley, 2002). In the case of IRR, the Pearson correlations between *Z* and *X*, and *Z* and *Y* were varied continuously between .10 and .90. This continuous variation facilitates the aggregation of a parameter estimate over factors and factor levels (more specifically, it facilitates integration over a continuous interval of a parameter). Aggregating parameter estimates over other factors with only a few levels would lead to an underestimation of the variance of the parameter estimate, and therefore to a biased empirical sampling deviation.

b) Naturally dichotomous criterion: The distribution of a naturally dichotomous criterion variable is defined via the two proportions *p* and *q*, no underlying distribution exists. We generated bivariate (DRR) and trivariate (IRR) data where *Y* was naturally dichotomous, and *X* and *Z* were a mixture distribution of two uniform normal distributions, one normal distribution

for each of the two criterion groups. Abrahams and colleagues (1971) also used this mixture distribution to develop the Taylor-Russell tables for dichotomous criterion variables. As shown in Equation 2, the magnitude of the point-biserial correlation coefficient depends on the mean difference in *X* (or in *Z*) between the two criterion groups. Therefore, we generated data with point-biserial population correlations between *X* and *Y* of .20, .40, and .60 based on the difference in mean $M_1 - M_2$ for a given *p*-*q*-split. In the case of IRR, we varied the differences in means and therefore Pearson correlations between *Z* and *X*, and *Z* and *Y* continuously between .10 and .90.

Analysis of the parameters

In order to investigate the accuracy of the missing data approach, we analyzed the residual distribution of the correlation estimates for each factor combination. The concept of accuracy provides quantitative information about the goodness of a parameter estimate and encompasses trueness and precision (Ayyub & McCuen, 2011). Trueness, which is also known as bias or systematic error, describes the distance of an estimated value to the true parameter value. Precision, which is also known as random error, describes the reproducibility of an estimated value. The mean error (ME) of the residuals is a measure of trueness, and the root-mean-square error (RMSE) of the residuals is a measure of precision. Let $\hat{\theta}$ be the value of the parameter estimate and θ the true value of the parameter. The ME and the RMSE can be calculated by

$$\mathrm{ME} = \frac{1}{E}\sum_{i=1}^{E}\left(\hat{\theta}_i - \theta\right) \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{E}\sum_{i=1}^{E}\left(\hat{\theta}_i - \theta\right)^2} \tag{6}$$

where *E* is the number of the Monte Carlo experiments (*E* = 5,000 in the present study), and $\hat{\theta}_i$ is the pooled correlation coefficient from the multiple imputation analysis. When the ME is close to zero, the parameter estimate is to be said unbiased. The smaller the RMSE, the more precise the estimation, i.e. the higher the reproducibility of the estimated parameter value.

The multiple imputation standard error of the correlation coefficient (see Appendix Equation A8) is a theoretical (asymptotic) estimate of the sampling deviation of the sample correlation coefficient. The RMSE is a measure of the empirical sampling deviation of the sample correlation coefficient. In order to investigate the accuracy of the multiple imputation standard error, we compared its average value (over 5,000 Monte Carlo experiments) with the RMSE for each factor combination. When the difference between the theoretical and the empirical value of

the multiple imputation standard error is close to zero, the theoretical value is an accurate measure of the true sampling distribution of the corrected sample correlation coefficient. When the theoretical value of the multiple imputation standard error is smaller than the empirical sampling deviation, the confidence intervals for the population correlation based on the multiple imputation standard error are smaller than they need to be.

Results

Continuous criterion variable

Table 1 summarizes the trueness and the precision of the correction for direct and indirect range restriction scenarios in the case of a continuous criterion variable across 5,000 Monte-Carlo experiments for each factor combination. For both the DRR and IRR scenarios, the correction of the Pearson correlation coefficient is negatively biased, whereby the bias tends to be smaller in the case of an IRR scenario. The bias is higher for a small sample size of $N = 50$ than for a moderate sample size of $N = 100$, and increases as the selection ratio decreases and the true correlation ρ between X and Y increases. The correction is more precise for moderate samples than for small ones and increases as the selection ratio increases. The precision of the correction increases as the true Pearson correlation coefficient between X and Y increases.

Table 1. Mean error (ME) and root-mean-square error (RMSE, in parentheses) for a continuous criterion variable in the case of direct and indirect range restriction scenarios.

		DRR			IRR		
N	SR	ρ = .2	ρ = .4	ρ = .6	ρ = .2	ρ = .4	ρ = .6
50	0.2	-.075 (.499)	-.157 (.506)	-.178 (.466)	-0.088 (.315)	-0.117 (.312)	-0.140 (.293)
	0.3	-.056 (.403)	-.100 (.389)	-.100 (.332)	-0.051 (.248)	-0.074 (.239)	-0.080 (.204)
	0.4	-.035 (.333)	-.067 (.308)	-.067 (.249)	-0.035 (.201)	-0.047 (.187)	-0.053 (.154)
	0.5	-.022 (.273)	-.041 (.249)	-.043 (.190)	-0.022 (.165)	-0.032 (.154)	-0.034 (.117)
100	0.2	-.053 (.391)	-.083 (.368)	-.098 (.316)	-0.053 (.232)	-0.068 (.224)	-0.070 (.185)
	0.3	-.031 (.303)	-.047 (.273)	-.056 (.218)	-0.030 (.176)	-0.037 (.161)	-0.042 (.134)
	0.4	-.021 (.242)	-.031 (.219)	-.035 (.163)	-0.021 (.143)	-0.023 (.125)	-0.027 (.102)
	0.5	-.012 (.199)	-.021 (.175)	-.022 (.126)	-0.016 (.116)	-0.013 (.099)	-0.017 (.078)

Note. N … sample size of the applicant dataset, SR … selection ratio, DRR … direct range restriction scenario, IRR … indirect range restriction scenario.

Table 2 summarizes the comparison of the multiple imputation standard error with the empirical sampling distribution for direct and indirect range restriction scenarios in the case of a continuous criterion variable across 5,000 Monte-Carlo experiments for each factor combination. The results show that the multiple imputation standard error tends to underestimate the sampling deviation of the sample correlation coefficient. This underestimation tends to be smaller in the case of an IRR scenario than for a DRR scenario. The difference between the multiple imputation standard error and the sampling deviation decreases as the selection ratio, the sample size, and the population correlation increase.

Table 2. Average multiple imputation standard error and its absolute bias to the empirical sampling deviation (in parentheses) for a continuous criterion variable in the case of direct and indirect range restriction scenarios.

		DRR			IRR		
N	SR	ρ = .2	ρ = .4	ρ = .6	ρ = .2	ρ = .4	ρ = .6
50	0.2	.361 (-.148)	.353 (-.144)	.327 (-.142)	0.319 (-.043)	0.337 (-.069)	0.299 (-.057)
	0.3	.326 (-.091)	.311 (-.084)	.273 (-.081)	0.268 (-.030)	0.267 (-.040)	0.220 (-.025)
	0.4	.294 (-.053)	.274 (-.048)	.230 (-.041)	0.235 (-.023)	0.222 (-.022)	0.177 (-.014)
	0.5	.263 (-.035)	.241 (-.029)	.196 (-.024)	0.208 (-.016)	0.191 (-.013)	0.149 (-.007)
100	0.2	.317 (-.076)	.298 (-.071)	.260 (-.064)	0.245 (-.031)	0.226 (-.026)	0.195 (-.024)
	0.3	.270 (-.048)	.247 (-.037)	.202 (-.030)	0.194 (-.015)	0.179 (-.014)	0.146 (-.013)
	0.4	.233 (-.023)	.208 (-.020)	.164 (-.015)	0.167 (-.010)	0.151 (-.010)	0.119 (-.006)
	0.5	.203 (-.015)	.178 (-.015)	.137 (-.008)	0.145 (-.006)	0.132 (-.006)	0.101 (-.003)

Note. N … sample size of the applicant dataset, SR … selection ratio, ρ … population correlation between predictor and criterion, DRR … direct range restriction scenario, IRR … indirect range restriction scenario.

Naturally dichotomous criterion variable

Table 3 summarizes the trueness and the precision of the correction for a direct range restriction scenario in the case of a naturally dichotomous criterion variable for each factor combination. However, for a number of factor combinations the number of Monte Carlo experiments was less than 5,000. More than 90% of selected samples did not meet the prerequisite of at least three observations in each criterion group, or there were convergence problems with the logistic regression imputation. No selected sample met the prerequisite at a base rate of success (BR) of 80% and a true point-biserial correlation coefficient of .6. The superscripted numbers in Table 3 and 5 show the percentage of excluded samples. Results of the remaining Monte Carlo experiments show that the correction of the point-biserial correlation

coefficient is negatively biased for factor combinations of sample size, true point-biserial correlation coefficient, and selection ratio when the base rate of success is 20% or 50%, but positively biased when the BR is 80%. As expected, the bias decreases as the selection ratio and the sample size increase. The effect of the direction of the true point-biserial correlation coefficient varied across different base rates of success: For a BR of 20%, the bias of the correction became smaller as ρ_{pb} increases, but for a BR of 50%, bias increases as ρ_{pb} increases. For a BR of 80%, too many data points are missing to assess the direction of the effect. The correction become more precise as the sample size, the selection ratio, and the true correlation between predictor and criterion increase. Comparing the results of the same factor combinations across the three base rates of success to the extent allowed by the data reveals that the correction is most accurate when the BR is 50%. This indicates a non-linear relationship between base rate of success and accuracy.

Table 3. Mean error and root-mean-square error (in parentheses) for a naturally dichotomous criterion variable in the case of direct range restriction scenarios.

		BR = .2			BR = .5			BR = .8	
N	SR	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4
50	0.2	-.282	-.263	-.231	-.040	-.100	---	---	---
		(.330)[19%]	(.308)	(.254)[15%]	(.166)[10%]	(.161)[66%]			
	0.3	-.250	-.212	-.165	-.035	-.064	---	---	---
		(.318)[1%]	(.272)	(.198)	(.188)	(.167)[5%]			
	0.4	-.201	-.156	-.102	-.031	-.062	-.089	.241	---
		(.283)	(.220)	(.132)	(.191)	(.176)	(.153)[11%]	(.275)[46%]	
	0.5	-.154	-.110	-.062	-.028	-.055	-.081	.159	---
		(.240)	(.169)	(.084)	(.180)	(.161)	(.153)	(.226)[10%]	
100	0.2	-.231	-.197	-.175	-.042	-.068	---	.208	---
		(.290)	(.248)	(.204)	(.166)	(.153)[5%]		(.235)[73%]	
	0.3	-.176	-.133	-.093	-.037	-.061	-.091	.193	---
		(.252)	(.191)	(.125)	(.173)	(.166)	(.146)[12%]	(.241)[15%]	
	0.4	-.131	-.087	-.051	-.029	-.048	-.075	.135	.140
		(.210)	(.141)	(.076)	(.163)	(.154)	(.143)	(.216)	(.180)[53%]
	0.5	-.092	-.057	-.030	-.020	-.038	-.052	.092	.077
		(.166)	(.104)	(.049)	(.144)	(.135)	(.117)	(.184)	(.166)[8%]

Note. BR … base rate of success, *N* … sample size of the applicant dataset, SR … selection ratio, DRR … direct range restriction scenario, IRR … indirect range restriction scenario, --- … >80% of selected samples did not meet the prerequisite of at least three observations in each criterion group.

Table 4 shows the difference between the multiple imputation standard error of the estimate of the point-biserial correlation coefficient and its empirical sampling deviation decreases as the sample size and selection ratio increase. The effect of the true point-biserial population correlation is not clear: For a BR of 20% only, this difference tends to decrease as the true population correlation increases. For a BR of 20%, the multiple imputation standard error tends to underestimate the sampling deviation; for base rates of success of 50 and 80%, the multiple imputation standard error tends to overestimates the sampling deviation for all combinations of sample size, selection ratio, and true point-biserial population correlation.

Table 4. Average multiple imputation standard error and its absolute bias to the empirical sampling deviation (in parentheses) for a naturally dichotomous criterion variable in the case of direct range restriction scenarios.

		BR = .2			BR = .5			BR = .8	
N	SR	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4
50	0.2	.277	.270	.256	.276	.268	---	---	---
		(-.021)	(-.057)	(-.057)	(.075)	(.044)			
	0.3	.266	.249	.213	.264	.254	---	---	---
		(-.060)	(-.055)	(-.015)	(.048)	(.050)			
	0.4	.249	.243	.171	.249	.237	.221	.252	---
		(-.056)	(-.041)	(-.011)	(.029)	(.029)	(.030)	(.016)	
	0.5	.231	.198	.142	.231	.216	.194	.245	---
		(-.043)	(-.030)	(-.004)	(.018)	(.018)	(.020)	(.013)	
100	0.2	.240	.223	.200	.238	.232	---	.234	---
		(-.064)	(-.043)	(-.016)	(.055)	(.061)		(.028)	
	0.3	.220	.189	.140	.222	.214	.205	.230	---
		(-.053)	(-.030)	(-.004)	(.030)	(.036)	(.036)	(-.003)	
	0.4	.197	.159	.105	.202	.192	.178	.222	.209
		(-.039)	(-.018)	(-.002)	(.017)	(.020)	(.023)	(-.003)	(.041)
	0.5	.176	.135	.088	.183	.167	.143	.209	.203
		(-.025)	(-.010)	(-.001)	(.012)	(.012)	(.010)	(-.008)	(.034)

Note. N … sample size of the applicant dataset, SR … selection ratio, BR … base rate of success, ρ … population correlation between predictor and criterion.

Table 5 summarizes the trueness and precision of the correction for an indirect range restriction scenario in the case of a naturally dichotomous criterion variable for each factor combination. The results show a similar pattern as the correction for a direct range restriction scenario. The correction of the point-biserial correlation coefficient is negatively biased for factor combinations of sample size, true point-biserial correlation coefficient, and selection ratio when

the base rate of success is 20% or 50%, but positively biased when the BR is 80%. The bias decreases as the selection ratio and the sample size increase. The correction becomes more precise as the sample size, the selection ratio, and the true point-biserial correlation coefficient between predictor and criterion increase. Similar to DRR, the correction is least biased when the BR is 50%. In contrast to DRR, the correction is not most precise for a BR of 50%. For a moderate sample size of $N = 100$, the precision of the correction tends to decrease as the base rate of success increases.

Table 5. Mean error (ME) and root-mean-square error (RMSE, in parentheses) for a naturally dichotomous criterion variable in the case of indirect range restriction scenarios.

		BR = 0.2			BR = 0.5			BR = 0.8	
N	SR	$\rho_{pb} = .2$	$\rho_{pb} = .4$	$\rho_{pb} = .6$	$\rho_{pb} = .2$	$\rho_{pb} = .4$	$\rho_{pb} = .6$	$\rho_{pb} = .2$	$\rho_{pb} = .4$
50	0.2	-.151 (.233)$^{34\%}$	-.142 (.211)$^{34\%}$	-.164 (.200)$^{33\%}$	-.018 (.171)$^{68\%}$	-.057 (.164)$^{69\%}$	-.083 (.137)$^{70\%}$	---	---
	0.3	-.136 (.192)	-.122 (.171)	-.121 (.153)	-.015 (.140)$^{45\%}$	-.048 (.137)$^{47\%}$	-.062 (.116)$^{47\%}$	---	---
	0.4	-.099 (.142)	-.086 (.125)	-.085 (.113)	-.026 (.125)$^{29\%}$	-.045 (.122)$^{30\%}$	-.053 (.103)$^{30\%}$	---	---
	0.5	-.068 (.105)	-.059 (.091)	-.059 (.083)	-.027 (.108)$^{14\%}$	-.040 (.108)$^{14\%}$	-.043 (.090)$^{15\%}$	.086 (.156)$^{72\%}$	.051 (.127)$^{73\%}$
100	0.2	-.127 (.119)$^{17\%}$	-.112 (.166)$^{17\%}$	-.119 (.153)$^{16\%}$	-.024 (.141)$^{46\%}$	-.049 (.139)$^{46\%}$	-.061 (.114)$^{47\%}$	---	---
	0.3	-.092 (.137)	-.076 (.115)	-.074 (.102)	-.021 (.116)$^{30\%}$	-.041 (.114)$^{30\%}$	-.049 (.099)$^{30\%}$	.101 (.169)$^{75\%}$	.063 (.132)$^{77\%}$
	0.4	-.056 (.092)	-.047 (.078)	-.047 (.071)	-.022 (.102)$^{19\%}$	-.033 (.099)$^{18\%}$	-.039 (.084)$^{18\%}$	.080 (.150)$^{62\%}$	.048 (.122)$^{64\%}$
	0.5	-.035 (.065)	-.030 (.056)	-.030 (.050)	-.023 (.085)$^{6\%}$	-.028 (.081)$^{6\%}$	-.029 (.069)$^{6\%}$	.052 (.125)$^{51\%}$	.022 (.107)$^{52\%}$

Note. BR … base rate of success, *N* … sample size of the applicant dataset, SR … selection ratio, DRR … direct range restriction scenario, IRR … indirect range restriction scenario, --- … ≥80% of selected samples did not meet the prerequisite of at least three observations in each criterion group.

Table 6 shows the results for the accuracy of the multiple imputation standard error of the estimate of the point-biserial correlation coefficient for a dichotomous criterion variable in the case of an indirect range restriction scenario. The difference between the multiple imputation standard error and its empirical sampling deviation decreases as the sample size and the selection ratio increase. For a BR of 20% and 50%, this difference decreases as the true point-biserial population correlation increases, but the effect for a BR of 80% is not clear.

Table 6. Average multiple imputation standard error and its absolute bias to the empirical sampling deviation (in parentheses) for a naturally dichotomous criterion variable in the case of indirect range restriction scenarios.

		BR = .2			BR = .5			BR = .8	
N	SR	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4	ρ_{pb} = .6	ρ_{pb} = .2	ρ_{pb} = .4
50	0.2	.248	.237	.218	.247	.235	.208	---	---
		(-.020)	(-.006)	(.002)	(.033)	(.041)	(.052)		
	0.3	.221	.208	.185	.225	.210	.182	---	---
		(-.019)	(-.006)	(.007)	(.036)	(.035)	(.039)		
	0.4	.196	.181	.156	.206	.192	.162	---	---
		(-.004)	(-.004)	(.013)	(.032)	(.027)	(.029)		
	0.5	.177	.162	.137	.189	.175	.146	.195	.172
		(.002)	(.008)	(.014)	(.022)	(.017)	(.020)	(.003)	(.018)
100	0.2	.198	.183	.162	.201	.187	.160	---	---
		(-.017)	(.013)	(-.003)	(.037)	(.034)	(.027)		
	0.3	.163	.147	.122	.178	.165	.140	.183	.162
		(-.007)	(.004)	(.003)	(.032)	(.027)	(.022)	(.002)	(.014)
	0.4	.139	.122	.099	.160	.147	.121	.173	.152
		(.002)	(.002)	(.005)	(.022)	(.019)	(.015)	(.013)	(.013)
	0.5	.124	.108	.087	.142	.130	.105	.163	.146
		(.003)	(.002)	(.005)	(.016)	(.015)	(.010)	(.024)	(.019)

Note. N … sample size of the applicant dataset, SR … selection ratio, BR … base rate of success, ρ … population correlation between predictor and criterion.

Discussion

Statistical problems in estimating the predictive validity of a selection method become worse when the number of applicants in the unrestricted dataset is moderate or small because statistical estimates are only based on a subsample of applicants. In this paper, we proposed using the state-of-the-art missing data approach multiple imputation by chained equations to correct correlations for direct and indirect range restriction scenarios when the sample size is small or moderate. Approaches to overcoming the range restriction problem, including multiple imputation techniques, have been developed within the framework of large sample theory. However, some findings on the comparison between maximum likelihood and multiple imputation suggest that multiple imputation is more efficiently with small samples. Additionally, correction methods have been widely studied for continuous criterion variables but not for dichotomous criterion variables. Therefore, the primary purpose of this research was to examine the accuracy of correlation coefficients corrected for range restriction scenarios using multiple

imputation by chained equations in small or moderate samples and for continuous and dichotomous criterion variables. To the best of our knowledge, no empirical studies so far have investigated the accuracy of the multiple imputation standard error of the population correlation estimate in the case of direct (DRR) and indirect (IRR) range restriction scenarios. Therefore, the second purpose of this study was to examine the accuracy of the multiple imputation standard error of the population correlation estimate. We conducted Monte Carlo simulations to accomplish both purposes for four scenarios: a DRR and an IRR scenario with a continuous and a dichotomous criterion variable. Sample size, selection ratio, true population correlation, and base rate of success were systematically varied in an experimental design.

In the case of a continuous criterion variable, the corrected Pearson correlation coefficient systematically underestimated the true correlation between predictor and criterion for both direct and indirect range restriction scenarios, especially when the selection ratio was small with 20 percent selected applicants. The correction was more precise for moderate samples than for small samples and gradually increased as the selection ratio and the true correlation coefficient increased. Our results are consistent with the findings of the simulation studies by Chan and Chan (2004), who investigated Thorndike's correction formula for a selection scenario on *X* (DRR). The extent of this bias is similar for both approaches, e.g. for $N = 100$, SR = .2, and $\rho = .2$: -.053 and -.059 (p. 374). This means that the underestimation of the correlation coefficient due to range restriction cannot be fully corrected in either approach. The multiple imputation standard error of the corrected correlation coefficient tended to be smaller than the empirical sampling deviation, which means that confidence intervals for the population correlations are smaller than they should be. This bias was lower for moderate than for small samples and gradually decreased as the selection ratio increased.

In the case of a naturally dichotomous criterion variable, multiple imputation by chained equations could not be applied for a large number of selected samples because the criterion variable was constant or nearly constant. This was often the case when the sample size and the selection ratio were small, and the base rate of success was high. The estimate of the population correlation is strongly biased for both direct and indirect range restriction scenarios. The results show that the number of individuals in the selected samples are too small for an accurate correction. Consequently, our findings indicate that correcting for range restriction when the criterion is dichotomous is not a trustworthy method for small sample sizes and for combinations of small selection ratios and low or high base rates of success. In contrast, the multiple

imputation standard error of the corrected point-biserial correlation coefficient was accurate over a wide range of factor combinations for direct and indirect range restriction scenarios.

This study's findings provide empirical evidence about the accuracy of correcting for range restriction using multiple imputation by chained equations, and support researchers and evaluators in their assessment of conditions under which corrected correlation coefficients can be trusted. The results show that interpreting the population correlation estimates can sometimes lead to invalid conclusions about the predictive validity of selection methods if the number of applicants is small or moderate and the selection is rigorous, especially in the case of a dichotomous criterion variable. However, this does not mean that selections should be made only on a large number of applicants or that small selection ratios should be avoided. The predictive validity of a selection method can be high even for a highly competitive selection (i.e. a small selection ratio) with a small number of applicants. The problem is simply that a satisfactory statistical evaluation of the predictive validity is not possible under some conditions. The missing data approach cannot compensate for having small samples (Graham & Schafer, 1999) in which the most criterion values are systematically missing. It would be naive to believe that the predictive validity of a selection method can be statistically assessed for a small number of individuals regardless of which approach is used to correct for range restriction. However, multiple imputation allows for the most effective usage of all collected data. Although this correction can lead to biased estimates in small sample sizes, the missing data approach is currently the best-known approach for handling a dichotomous criterion.

Some of the methodological limitations of our study should be mentioned. However, these limitations also point to promising avenues for further research. The Monte Carlo simulations we conducted considered a limited number of combinations of sample size, population correlation, and base rates of success. In the case of a naturally dichotomous variable, the results indicate a non-linear relation between accuracy of the corrected point-biserial correlation coefficient and the base rate of success. Further research should investigate this effect in more detail. For our data simulation with a naturally dichotomous criterion variable, we assumed a mixture distribution of the predictor based on two normal distributions for each criterion group. This distribution was also used to develop the Taylor-Russell tables for a naturally dichotomous criterion variable (Abrahams et al., 1971) and in the simulation study by Pfaffel, Kollmayer, and colleagues (2016). Although many reasons speak in favor of the assumption of normally distributed values for the criterion groups, other distributions are quite

conceivable and should be also investigated. Finally, we generated multivariate datasets with a minimum number of variables, which is not typical for real datasets. The correction using multiple imputation by chained equations should become more accurate for datasets with more variables, e.g. more predictors and covariates, or even more than one criterion. However, generating multivariate data, especially multivariate data with non-identically distributed variables, is often difficult but necessary in simulation studies. Further research should investigate the accuracy of the correction in datasets with more predictors, covariates, and criteria.

In conclusion, our study shows that the proposed missing data approach is accurate for estimating the predictive validity of a selection method for a continuous criterion variable, even for a small number of applicants when the selection ratio is not too small. For a dichotomous criterion variable, a small or moderate number of applicants sometimes leads to biased estimates or an inability to carry out the correction. The multiple imputation standard error of the estimate of the predictive validity is accurate over a wide range of conditions for both kinds of criterion variables and for direct and indirect range restriction scenarios.

References

Abrahams, N. M., Alf, E. F., & Wolfe, J. J. (1971). Taylor-Russell tables for dichotomous criterion variables. *Journal of Applied Psychology*, *55*(5), 449–457. http://doi.org/10.1037/h0031761

Aitken, A. C. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, *4*(2), 106–110. http://doi.org/10.1017/S0013091500008063

Allen, N. L., & Dunbar, S. B. (1990). Standard Errors of Correlations Adjusted for Incidental Selection. *Applied Psychological Measurement*, *14*(1), 83–94. http://doi.org/10.1177/014662169001400109

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

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement*, *4*(3), 385–398. http://doi.org/10.1177/014662168000400309

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the Effect Size of d for Range Restriction and Unreliability. *Organizational Research Methods*, *4*(1), 46–61. http://doi.org/10.1177/109442810141003

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Erlbaum.

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

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

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507. http://doi.org/10.2307/2331838

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

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: SAGE Publications, Inc.

Greener, J. M., & Osburn, H. G. (1979). An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement*, *3*(1), 31–41. http://doi.org/10.1177/014662167900300104

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement*, *7*(2), 227–237. http://doi.org/10.1177/014662168300700210

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

Kemery, E. R., Dunlap, W. P., & Griffeth, R. W. (1988). Correction for variance restriction in point-biserial correlations. *Journal of Applied Psychology*, *73*(4), 688–691. http://doi.org/10.1037/0021-9010.73.4.688

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics. Vol. 1: Distribution theory* (4th ed., Vol. 1). New York: Macmillan.

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://doi.org/10.1017/S0080454100006385

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer Science & Business Media.

Levin, J. (1972). The occurrence of an increase in correlation by restriction of range. *Psychometrika*, *37*(1), 93–97. http://doi.org/10.1007/BF02291414

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://doi.org/http://dx.doi.org/10.1037/0021-9010.66.6.655

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2–3), 292–326. http://doi.org/10.1177/0049124189018002004

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed). Hoboken, N.J: Wiley.

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.

Millsap, R. E. (1989). Sampling variance in the correlation coefficient under range restriction: A Monte Carlo study. *Journal of Applied Psychology*, *74*(3), 456–461. http://doi.org/10.1037/0021-9010.74.3.456

Muthén, L. K., & Muthén, B. O. (2015). *Mplus User's Guide.* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*(1), 201–211. http://doi.org/10.1214/aoms/1177706717

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*, *11*(3), e0152330. http://doi.org/10.1371/journal.pone.0152330

Pfaffel, A., Schober, B., & Spiel, C. (2016). A comparison of three approaches to correct for direct and indirect range restrictions: A simulation study. *Practical Assessment, Research & Evaluation*, *21*(6), 1–15.

R Core Team. (2016). A language and environment for statistical computing (Version 3.3.0) [64-bit]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Raju, N. S., Steinhaus, S. D., Edwards, J. E., & DeLessio, J. (1991). A logistic regression model for personnel selection. *Applied Psychological Measurement*, *15*(2), 139–152. http://doi.org/10.1177/014662169101500204

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

Robitzsch, A., Grund, S., & Henke, T. (2015). Some additional multiple imputation functions, especially for "mice." Retrieved from http://cran.stat.ucla.edu/web/packages/miceadds/

Rosseel, Y. (2012). An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). http://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581. http://doi.org/10.2307/2335739

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (Vol. 1, pp. 20–34). American Statistical Association.

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.

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.

Ulrich, R., & Wirtz, M. (2004). On the correlation of a naturally and an artificially dichotomized variable. *British Journal of Mathematical and Statistical Psychology*, *57*(2), 235–251. http://doi.org/10.1348/0007110042307203

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/

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer Science & Business Media.

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 Appendix shows the equations for computing the estimate of the Pearson correlation coefficient, its associated standard error, and the confidence interval for multiple imputed datasets. Equations A1 to A9 are implemented in the method `micombine.cor()` of the R package `miceadds` (Robitzsch, Grund, & Henke, 2015). The multiple imputation point estimate of the Pearson correlation coefficient $\bar{r}$ (or of the point-biserial correlation coefficient) is the arithmetic average of the m Fisher z-transformed correlation estimates

$$\bar{r} = \tanh\left(\frac{1}{m}\sum_{t=1}^{m}\operatorname{artanh}(\hat{r}_t)\right) \tag{A1}$$

where $\hat{r}_t$ is the correlation estimate (see Equation 1) from the complete dataset t, artanh is the inverse hyperbolic tangent function (the Fisher z-transformation), and tanh is the hyperbolic tangent function, which converts the Fisher z-value back into a correlation coefficient. The corresponding Fisher z-transformed point estimate $\bar{z}_r$ is calculated by

$$\bar{z}_r = \frac{1}{m}\sum_{t=1}^{m}\operatorname{artanh}(\hat{r}_t) \tag{A2}$$

The within-imputation variance W is the arithmetic average of the squared standard error of the m complete datasets

$$W = \frac{1}{m}\sum_{t=1}^{m}\frac{1}{N-3} \tag{A3}$$

and the between-imputation variance B is the sample variance of the Fisher z-transformed correlation estimates across the m datasets

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\operatorname{artanh}(\hat{r}_t) - \bar{z}_r)^2 \tag{A4}$$

These two components of uncertainty can be combined into a single quantity, the total-imputation variance T of the Fisher's z-transformed parameter estimate $\bar{z}_r$:

$$T = W + \frac{m+1}{m}B \tag{A5}$$

Consequently, the Fisher's z-transformed multiple imputation standard error is the square root of the total-imputation variance

$$SE_{\text{Fisher } z} = \sqrt{T} \tag{A6}$$

The lower and the upper bound of the $1 - \alpha$ asymmetric confidence interval can be calculated by

$$\text{CI}_{1-\alpha} = \tanh\left(\operatorname{artanh}(\bar{z}_r) \pm z_{1-\alpha/2} \cdot SE_{\text{Fisher } z}\right) \tag{A7}$$

where $z_{1-\alpha/2}$ is the value of the cumulative normal distribution at half of the significance level α. The z-value for a 95% confidence interval is approximately 1.96. Based on the confidence

interval, the standard error of the point estimate of the Pearson correlation coefficient can by calculated as

$$SE_{\bar{r}} = \frac{(\text{upper bound CI}_{1-\alpha} - \text{lower bound CI}_{1-\alpha})}{2 \cdot z_{1-\alpha/2}} \tag{A8}$$

In order to test the null hypothesis that $\bar{r}$ is equal to 0, a one sample *t*-test has to be applied to the corresponding Fisher *z*-transformed point estimate $\bar{z}_r$, because $\bar{z}_r$ is to be assumed *t*-distributed with $df = N - 2$ if the sample size is not too small and the magnitude of the correlation coefficient is not too extreme.

$$t = \frac{\bar{z}_r}{SE_{\text{Fisher } z}} \tag{A9}$$

Abstract

A common problem in the assessment of the predictive validity of selection methods is the so-called range restriction problem, which causes biased estimates of the population correlation between predictor and criterion. To overcome this problem, researchers proposed correction formulas, which have been widely studied for continuous criterion variables but not for dichotomous ones. The few existing approaches dealing with a dichotomous criterion are often inapplicable because they require information about the typically unknown base rate of success. Thus, there is a lack of scientific research on suitable methods able to correct correlation coefficients for range restriction in the case of a dichotomous criterion variable.

The present doctoral thesis aims to overcome this deficiency by viewing the range restriction problem as a missing data problem. In order to establish the missing data approach in this research field, the first paper compares common correction formulas with two state-of-the-art missing data techniques for a continuous criterion variable. The second paper shows that the missing data technique multiple imputation by chained equations provides effective and accurate estimates of the population correlation in the case of dichotomous criterion variables. In addition, multiple imputation by chained equations is able to estimate the base rate of success. The third paper focuses on the accuracy of the population estimates with small and moderate sample sizes and provides empirical evidence about conditions under which the estimation has limitations.

The empirical studies are introduced by a description of the theoretical background concerning the range restriction problem and traditional approaches to overcoming this problem. Then, the goals of the dissertation are described in detail, followed by a synopsis of the three publications. Finally, the findings of the three papers are summarized and methodological strengths and limitations of the empirical studies are discussed.

Kurzfassung

Ein bekanntes Problem bei der Bewertung der prognostischen Validität von Auswahlverfahren ist das sogenannte range restriction-Problem, das verzerrte Schätzer der Populationskorrelation zwischen Prädiktor und Kriterium verursacht. Um dieses Problem zu lösen, haben Forscher Korrekturformeln vorgeschlagen, die weitgehend für stetige jedoch nicht für dichotome Kriteriumsvariablen erforscht wurden. Die wenigen Ansätze, die ein dichotomes Kriterium betrachten sind meist nicht anwendbar, weil sie Informationen über die typischerweise unbekannte Basiserfolgsrate benötigen. Der wissenschaftlichen Forschung fehlen folglich geeignete Methoden Korrelationskoeffizienten für das range restriction-Problem im Falle von dichotomen Kriteriumsvariablen zu korrigieren.

Die vorliegende Dissertation beabsichtigt dieses Defizit zu beseitigen indem das range restriction-Problem als ein Missing Data Problem betrachtet wird. Um den Missing Data Ansatz in diesem Forschungsbereich zu etablieren, vergleicht der erste Artikel bekannte Korrekturformeln mit zwei state-of-the-art Missing Data Methoden für steige Kriteriumsvariablen. Der zweite Artikel zeigt, dass die Missing Data Methode Multiple Imputation mit Markovketten effektive und genaue Schätzungen der Populationskorrelation im Falle künstlicher und natürlicher dichotomer Kriteriumsvariablen ermöglicht. Zusätzlich kann mittels Multipler Imputation mit Markovketten die Basiserfolgsrate geschätzt werden. Der dritte Artikel konzentriert sich auf die Genauigkeit der Populationsschätzer in kleinen und moderaten Stichprobengrößen und gibt empirische Evidenz über Bedingungen unter denen die Schätzung nur begrenzt möglich ist.

Den empirischen Untersuchungen wird eine Einleitung durch eine Beschreibung des theoretischen Hintergrunds über das range restriction-Problem sowie traditionelle Ansätze zur Überwindung dieses Problems vorangestellt. Dann werden die Ziele dieser Dissertation im Detail beschrieben, gefolgt von einer Zusammenfassung der drei Artikel. Schließlich werden die Ergebnisse der drei Artikel zusammengefasst und methodische Stärken sowie Grenzen der empirischen Untersuchungen diskutiert.