Behav Res (2016) 48:1021–1031
DOI 10.3758/s13428-015-0627-7

The influence of base rates on correlations: An evaluation of proposed alternative effect sizes with real-world data

Kelly M. Babchishin[1] · Leslie-Maaike Helmus[2]

Published online: 17 July 2015

Abstract Correlations are the simplest and most commonly understood effect size statistic in psychology. The purpose of the current paper was to use a large sample of real-world data (109 correlations with 60,415 participants) to illustrate the base rate dependence of correlations when applied to dichotomous or ordinal data. Specifically, we examined the influence of the base rate on different effect size metrics. Correlations decreased when the dichotomous variable did not have a 50 % base rate. The higher the deviation from a 50 % base rate, the smaller the observed Pearson's point-biserial and Kendall's tau correlation coefficients. In contrast, the relationship between base rate deviations and the more commonly proposed alternatives (i.e., polychoric correlation coefficients, AUCs, Pearson/Thorndike adjusted correlations, and Cohen's *d*) were less remarkable, with AUCs being most robust to attenuation due to base rates. In other words, the base rate makes a marked difference in the magnitude of the correlation. As such, when using dichotomous data, the correlation may be more sensitive to base rates than is optimal for the researcher's goals. Given the magnitude of the association between the base rate and point-biserial correlations ($r = -.81$) and Kendall's tau ($r = -.80$), we recommend that AUCs, Pearson/Thorndike adjusted correlations, Cohen's *d*, or polychoric correlations should be considered as alternate effect size statistics in many contexts.

Keywords Correlation · AUC · Cohen's *d* · Base rate · Statistics in prediction studies

Introduction

The relationship between two variables can be indexed by correlation coefficients. They are the simplest and most commonly understood effect size statistics in psychology and are one of the first statistical concepts taught in psychology courses (e.g., Aron, Aron, & Coups, 2006; Coolican, 2013; Eysenck, 2013). In the criminal justice field, for example, knowledge translation (particularly to those who create laws and policies) is an integral activity for a large majority of researchers. For this reason, statistics that are easy to understand, such as correlations, are often preferred to more complicated statistics (e.g., Gendreau & Smith, 2007).

The most commonly known correlation coefficient is Pearson's and it is intended for two continuous, normally distributed variables. For example, the relationship between temperature and level of snowfall in North America can be indexed by a correlation (e.g., $r = -.73$, with warmer temperature associated with less snowfall; Karl, Groisman, Knight, & Heim, 1993). Many common research questions, however, involve either dichotomous (e.g., yes/no) or ordinal (e.g., an item on a measure that has four ascending response options) variables, yet we still tend to use correlation coefficients (e.g., phi, point biserial, tau) when describing the relationship between such variables.

Correlations with dichotomous and ordinal data are commonly used in a variety of ways. For example, in the criminal justice field, they are often used to index the relationship between risk factors or risk scales with dichotomously defined recidivism (e.g., Andrews & Bonta, 2010; Olver, Wong, Nicholaichuk, & Gordon, 2007; Rice, Harris, & Lang,

✉ Kelly M. Babchishin
kelly.babchishin@theroyal.ca

[1] Royal's Institute of Mental Health Research, University of Ottawa, 1145 Carling Avenue, Ottawa, Ontario K1Z 7K4, Canada

[2] Correctional Service of Canada, Ottawa, Ontario, Canada

2013). Similarly, meta-analyses on the prediction of recidivism or other dichotomous outcomes frequently use correlations as their effect size (Andrews et al., 1990; Bonta, Rugge, Scott, Bourgon, & Yessine, 2008; Campbell, French, & Gendreau, 2009; French & Gendreau, 2006; Gendreau, Little, & Goggin, 1996; Gonçalves, Gonçalves, Martins, & Dirkzwager, 2014). Correlations with dichotomous or ordinal data also appear as intermediate steps in other analyses, such as factor analyses of measures with dichotomous or ordinal scale items.

The purpose of the current paper was to illustrate the base rate dependence of correlations when using dichotomous and ordinal data, and also to present and evaluate alternative statistics. Specifically, this study examined the effect of base rates on correlations and other alternative statistics, using data from a large Canadian sample which examined the relationship between dichotomous scale items and an overall summary risk judgement (on an ordinal 3-point scale).

Correlation coefficients: A primer and a summary of the issue

A *phi* correlation coefficient is used to describe the relationship between two dichotomous variables (e.g., presence or absence of a risk factor and recidivism scored as *yes* or *no*), whereas a point-biserial correlation is used to describe the relationship between one dichotomous (e.g., recidivism status) and one continuous (e.g., age) variable. Pearson's correlation is intended to describe the relationship between two continuous variables. Although these correlations are called by different names, they are mathematically equivalent. Calculating Pearson's *r* on data with one continuous and one dichotomous variable will produce the point-biserial correlation. Although there are correlations for multiple types of variables, the current study will focus on the point-biserial.

A further complication arises, however, when one or both variables are on an ordinal scale. In the criminal justice field, a common example of an ordinal variable is a rating of low, moderate, or high risk on an item or risk scale. Nonparametric alternatives of the correlation (Pearson, point-biserial, or phi) are available, including Spearman's rho and Kendall's tau, with Kendall's tau more commonly recommended (e.g., Field & Miles, 2010) because it is more suitable for smaller sample sizes and allows for ties in the rank ordering of scores. For example, using Kendall's tau to examine the relationship between risk groups on the VRAG risk scale (an ordinal measure; Quinsey et al., 2006) and recidivism (a dichotomous outcome) would provide a non-parametric point-biserial correlation. As you can see, there is a version of the correlation coefficient available for most types of data, and they are all interpreted in the same way. This has facilitated the widespread use of correlations as a common metric in research and knowledge translation.

Despite their widespread use, there are complications with correlations computed for dichotomous data. The issues have been discussed at length in the statistical literature (Cohen, 1983; Karabinus, 1975; McGrath & Meyer, 2006), but are presumably not well understood in many areas of psychology judging by the persistent and continued use of correlations on such data, without mention of this issue. The issue is that the size of the correlation is influenced by the distribution of the dichotomous variable (i.e., the base rate). In other words, correlations are base rate sensitive. In forensic psychology, base rates are often discussed in the context of recidivism as the dichotomous outcome variable (e.g., 12 % sexual reoffending rate), but they can also refer to other types of dichotomous variables, such as the endorsement rate on a risk factor or the proportion of a subgroup in a sample (e.g., Aboriginal vs. non-Aboriginal offenders). The further the base rate of the dichotomous item deviates from 50 %, the smaller the correlation will get. Given that the variance of a dichotomous variable is proportional to its base rate, the effect of base rates on correlations is a special case of the restriction of range problem, which has been discussed since correlation coefficients were first developed (Pearson, 1903).

On the one hand, the correlation's sensitivity to base rates has some interpretive meaning, reflecting the reduced statistical power and reliability of the correlation when base rates deviate from 50 % (Cohen, Cohen, West, & Aiken, 2003; McGrath & Meyer, 2006). Specifically, a correlation indexes the extent to which variability in one variable can be predicted from another. The lower the variability in one of the variables, the harder it becomes to predict. If dichotomous predictions are being made (e.g., this offender is "dangerous" or "not dangerous"), then the likelihood of false positives increases with low base rate events. This will be reflected in reduced correlations.

On the other hand, there are many situations where it is desirable to use an effect size statistic that is robust to base rates, particularly when making comparisons across studies, outcomes, or predictors with varying base rates. Additionally, in forensic risk assessment, dichotomous predictions are typically avoided in favor of probabilistic statements about risk on a continuous dimension (Association for the Treatment of Sexual Abusers, 2014; Helmus & Babchishin, 2015), which reduces the relevance of false positive rates (Helmus & Babchishin, 2015). In these situations, it may make more sense to incorporate base rate information in the variance of the effect size (with unequal distributions indicating greater imprecision in effect size estimation), rather than in the magnitude of the effect. In short, although there is inherent meaning in the relationship between base rates and correlations, its impact on real-world data is not well understood or discussed, and in many situations, base-rate-independent statistics may be desired.

In many fields that commonly employ experimental designs, it is not difficult to achieve relatively equal group sizes (i.e., 50 % base rates). Non-experimental psychology research, however, rarely encounters a 50 % base rate. This is especially true when examining low base rate events such as sexual recidivism (often below 10 %) or differences between a majority offender group compared to a unique subsample (e.g., female offenders, Aboriginal offenders). This makes the relationship between base rates and correlations meaningful in most non-experimental research.

Alternatives to correlations

To address the base rate dependence of correlations, different recommendations have been proposed either involving alternative statistics or statistical corrections to account for base rates. For example, Rice and Harris (2005) propose using AUC statistics when one variable is dichotomous. An AUC indexes the discrimination between a variable (whether dichotomous, continuous, or ordinal) and a dichotomous grouping variable. In the context of recidivism, an AUC can be interpreted as the probability that a randomly selected recidivist has a higher score on the risk scale than a randomly selected non-recidivist. In the medical field, AUCs are the most commonly utilized discrimination index (e.g., Pintea & Moldovan, 2009). Although AUCs have become commonly accepted for evaluating the predictive accuracy of risk assessment scales in forensic psychology, they are rarely used for dichotomous variables other than recidivism. Some recent exceptions have applied AUCs to comparisons between offenders scoring high and low on emotional congruence with children (McPhail, Hermann, & Fernandez, 2014) and comparisons between Aboriginal and non-Aboriginal offenders (Babchishin, Blais, & Helmus, 2012; Helmus, Babchishin, & Blais, 2012). In the latter context, an AUC is interpreted as the probability that a randomly selected Aboriginal offender will have a higher risk score than a randomly selected non-Aboriginal offender. Alternately, another common practice in forensic psychology is to report AUCs alongside correlations, rather than instead of them (e.g., Andrews et al., 2012; Olver et al., 2007; Viljoen, Mordell, & Beneteau, 2012).

Another option is to use Cohen's *d*, which is a standardized mean difference, and is typically used for group contrast research (e.g., do recidivists differ from non-recidivists on antisocial tendencies?). Although typically used to index differences between two groups on continuous or ordinal variables, it can also be used to compute group differences on dichotomous variables (Sánchez-Meca, Marín-Martínez, & Chácon-Moscoso, 2003).

Statistical adjustments to correlation coefficients have also been proposed to help account for the attenuation of correlations caused by a base rate's deviation from a perfect 50/50 split. As opposed to abandoning the correlation in favor of alternative statistics, these adjustments have the benefit of retaining the same intuitive metric of correlations. Given that the true correlation is a function of the observed correlation as well as the observed and true standard deviations of the two variables, it is possible to adjust the correlation by standardizing the variance of the dichotomous variable to a common value defined by the researcher. Several adjustments have been proposed (for a review, see Chapters 3 and 4 of Hunter & Schmidt, 2004), but the most widely used is that of Pearson (1903), further popularized by Thorndike (Case 2, 1949) and a number of contemporary measurement texts (Ghiselli et al., 1981; Ley, 1972). Such adjustments can be applied to one or both of the variables and have been used in meta-analyses of correlations (e.g., Bonta, Law, & Hanson, 1998; Campbell, French, & Gendreau, 2009). Its use, however, is likely hindered by the complexity of the calculations (defined further in the Methods section).

In addition, polychoric correlations are alternative correlation coefficients, and have been recommended for factor analyses of measures containing ordinal and dichotomous items (Brown, 2006; Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010). Polychoric correlations are used with two ordinal variables or one dichotomous and one ordinal variable. Whereas Pearson's and Kendall's tau correlation coefficients are computed on observed data, polychoric correlations estimate the relationship between two theorized normally distributed continuous latent variables from two observed variables (Flora & Curran, 2004; Holgado-Tello et al., 2010). Regardless of whether the assumption of continuous latent variables is realistic, the categorical restrictions imposed by ordinal scales reduce data variability, leading to a restriction of range. In contrast, these alternative correlation coefficients are less affected by base rates than Pearson's *r* (Brown, 2006; Flora & Curran, 2004; Holgado-Tello et al., 2010).

Despite the variety of alternatives and adjustments, unadjusted correlations (parametric or non-parametric) continue to be frequently used, with seemingly little concern for how they are influenced by base rates. We suspect that this is likely because the influence of the base rate on the magnitude of correlation coefficients is not well understood, nor are the alternative effect sizes to correlations.

Purpose of study

The current study used a large sample of real-world data to explore the relationship between base rates and correlations, and to evaluate commonly proposed alternative effect size metrics. This study answered the following basic questions about using correlations for dichotomous and ordinal data: Is it really going to make a meaningful difference in my analyses? Are the proposed alternatives that much better at dealing with diverse base rates?

These data came from a research study (Helmus & Forrester, 2014) with a large number of participants (n = 60, 415) and over 100 diverse dichotomous items displaying different base rates (ranging from 0.2 % to 99.7 %). This dataset provided a unique opportunity to explore the extent to which correlations are affected by base rates, and compare point-biserial correlations to alternative effect size options that have been proposed. A secondary purpose was to make recommendations about appropriate effect size statistics when one variable is dichotomous and the other variable is ordinal, which is a common occurrence in psychology research.

Method

Sample

The current study started with all Static Factors Assessments (SFAs) that were scored for Canadian federal offenders between 1997 and 2012. Some offenders had more than one assessment in a given sentence. In these cases, the first SFA assessment per sentence was used (provided that at least some SFA item information was included; N = 64,605). Given the lengthy study period (spanning 15 years), it was not uncommon for offenders to have multiple sentences in that timeframe. To avoid counting any individual twice, the assessment from their most recent sentence was used, reducing the available sample to 60,415 offenders. Ninety-four percent of the offenders were male (n = 56,914).

Measures

Static Factors Assessment (SFA; CSC, 2012) The SFA was developed in 1989 by a national working group in the Correctional Service of Canada (CSC; administers prison sentences of two or more years), and provides an assessment of criminal risk of offenders at admission (Motiuk, 1993). The SFA is scored by the parole officer or primary worker involved with the offender. The current study examined the Criminal History Record (CHR) and Offence Severity Record (OSR) subscales of the SFA. The CHR includes 38 items examining the offender's current and previous criminal offences (e.g., youth and adult convictions and sentences). The OSR includes 71 items examining the extent of harm from the offender's criminal activity (e.g., type of prior and current offences, victim information, harm to victims). After scoring the items of each subscale, the officer forms a single overall summary judgment of the offender's static risk (low, moderate, or high risk). A rating of high risk is intended to reflect cases in which the CHR shows considerable involvement in the criminal justice system and the OSR reflects considerable harm to society and to victims. A rating of low risk is intended to reflect cases in which the CHR reflects little criminal involvement and the OSR reflects little harm to society and victims. Lastly, a rating of moderate risk is intended for offenders who are not low risk but also not high risk (CSC, 2012).

Procedure and plan of analysis

Each observation in the analysis is an effect size estimate examining the relationship between a dichotomous item of the SFA and the officer's overall rating of risk (i.e., low, moderate, and high risk). Given the intent of the scale, each item should be related to the officer's overall risk evaluation. Each effect size estimate represents at least 53,792 assessments. Sample sizes fluctuated due to missing data, with the rate of missing item data ranging from 0.02 % to 11.0 %, with a median of 0.3 %.

Six different effect size estimates were used, including (1) Pearson's point-biserial correlations, (2) Kendall's Tau-b point-biserials, (3) polychoric correlations, (4) AUCs, (5) Cohen's d, and (6) Pearson/Thorndike adjusted correlations. For correlation coefficients, we ignored the direction of the relationship (negative or positive) and only used the absolute value. For example, a correlation of .20 and −.20 were both entered as .20. In addition, we restricted AUCs to be .50 and above, reversing those with negative predictive accuracy. Just as .20 and −.20 represents the same magnitude of effect size for correlations, AUCs of .20 and .80 represent the same magnitude for AUCs. SAS version 9.2 was used to calculate Pearson and Kendall's tau point-biserial correlations, AUCs, and polychoric correlations. The remaining analyses were computed in Excel using descriptive data from SAS.

The following formula was used to compute the Pearson/Thorndike correction for each point-biserial correlation coefficient (Ley, 1972):

$$r' = \left[(r_{xy})\left(\delta'_{\mathrm{x}}/\delta_{\mathrm{x}}\right)\right] / \left[1 - {r_{xy}}^2 + \left({r_{xy}}^2\right)\left(\delta'^2_{\mathrm{x}}/\delta^2_{\mathrm{x}}\right)\right]^{1/2}$$

where r_{xy} is the observed correlation, , δ_{x}' is the average standard deviation of the dichotomous variables (defined further below), and δ_{x} is the observed standard deviation of the dichotomous variable. The observed standard deviation of the dichotomous variable (δ_{x}) was defined as the square root of $p*q$, where p represents the proportion of the sample with the risk factor (the base rate) and q the proportion without the risk factor. In these analyses, the average base rate for the 109 items was 25.4 % and the average standard deviation (δ_{x}') was .332 (i.e., the standard deviation was calculated for all items and the average value of the standard deviation was .332). Conceptually, this approach to averaging standard deviations is most frequently used when the dichotomous variables are intended to measure the same latent construct (which is not the case in the current study). Because the goal of the study was to be illustrative, the condition of convergence on the same latent construct was suspended for the purpose of developing a relatively reliable estimate of the overall base rate.

To examine the effect of the base rate on the effect size measures, we correlated the absolute value of the effect size (reflecting the relationship between the dichotomous SFA item with a final risk judgement on a three-point ordinal scale) with the base rate deviation from a perfect 50/50 split.[1] This correlation essentially summarizes how strongly the effect sizes (point biserial correlations or their alternatives) are related to base rates. Conceptually, a correlation of 0 would mean that the effect size is unaffected by the base rate. Following Cohen (1992), correlations of .10, .30, and .50 were considered small, moderate, and large in magnitude, respectively (and this interpretation was applied to all variations of the correlation). For AUCs, values of .56, .64, and .71 were considered small, moderate, and large, respectively, as they most closely approximate Cohen's conventions (Rice & Harris, 2005). Cohen's *d*'s of .20, .50, and .80 were considered small, moderate, and large as well (Cohen, 1992).

The base rate deviation could range from 0 to 50, with 0 representing no deviation (a perfect 50 % base rate) and 50 reflecting that the deviation was 50 percentage points away from 50 (i.e., a 0 % or 100 % base rate). The further the deviation was from 0, the smaller the correlation coefficients should be, irrespective of the true relationship between the examined variables. A deviation score of 22, for example, was given if an item was endorsed by 72 % of the sample (22 points away from a 50 % base rate, or 72−50 = 22) or by 28 % (also 22 points away from a 50 % base rate, or 50−28 = 22).

Results

For the SFA items, deviations from a 50 % base rate ranged from 0 % to 50 % (*Mdn* = 32 %). As expected based on the previous statistical literature, base rates impacted the correlation coefficients. Figure 1 presents the relationship between the two most commonly used correlation coefficients (Pearson's point-biserial correlation and Kendall's tau) and the deviation from a perfect 50/50 base rate. Each data point represents the correlation coefficient between a dichotomous item of the SFA and the officer's overall rating of risk (i.e., low, moderate, and high risk). A clear pattern emerges: the higher the deviation from an equal base rate, the smaller the correlation. Specifically, both Pearson's point-biserial and Kendall's tau generally truncated to about .20 when base rates were less than 10 % or greater than 90 % (i.e., deviations of 40 % or more). In other words, with extremely low or high base rates, point biserial correlations and Kendall's tau did not exceed approximately .20. In contrast, the relationship between base rate deviations and the more commonly proposed alternatives (polychoric correlation coefficients, AUCs, Pearson/Thorndike adjusted correlations, and Cohen's *d*) were less remarkable, with the AUC being most robust to attenuation due to base rates (see Figs. 2 and 3).

Table 1 presents the Pearson's correlations between the effect size statistics and the base rate deviations. A large negative relationship was found for the traditional correlations (*r* = −.81 for Pearson's and *r* = −.80 for Kendall's tau); in other words, as the base rate deviates further from 50 %, the correlation decreased. The Pearson/Thorndike adjustment and Cohen's *d* reduced the base rate dependence, though it was still moderate (*r* = −.28 for Pearson/Thorndike adjustment and *r* = −.30 for Cohen's *d*). Polychoric correlations reduced the relationship between base rates and effect sizes, though it was still large (*r* = −.49), whereas AUCs reduced the relationship to more of a small effect (*r* = -.19). Such large magnitudes of the relationship between correlation coefficients and base rate deviations were unexpected; specifically, approximately two-thirds of the variance in the point-biserial and tau correlations was explained by the base rate deviation. In other words, the majority of the information provided by the correlation was about the base rate.

A few examples are provided to demonstrate the practical impact of these statistical effects. One item with a particularly low base rate was whether the offender had a previous offence against victims who were handicapped or infirm (this item had a 0.6 % base rate, or 367 offenders out of 59,095). As per the stated purpose of the scale, higher summary risk ratings should be given to offenders who have inflicted considerable harm to society and to their victims (CSC, 2012), and most people would consider offences against handicapped/infirm individuals as particularly egregious. Consequently, individuals with this type of prior offence should be rated as higher risk on the overall scale. Using the traditional Pearson's correlation, researchers would conclude that there is not even a small relationship between prior offences against handicapped/infirm victims and higher summary risk ratings (Pearson and Kendall's *r* = .06). The Pearson/Thorndike correction increased the correlation to .26, whereas the polychoric correlation revealed a moderate to large relationship (Polychoric *r* = .39). The AUC and Cohen's *d*, in contrast, revealed a large effect size (AUC = .72; Cohen's *d* = 0.82).

Similarly, offenders with a current homicide offence should presumably have higher risk ratings on the basis of offence severity. The low base rate for this occurrence (7.3 %), however, would reduce the traditional correlation statistics. Pearson's and Kendall's tau point-biserial correlations displayed a small relationship between current homicide offence and summary risk rating (*r* = .15), as did the Pearson/

[1] The use of Pearson's correlation in this analysis is applied to two continuous variables, which means that base rate dependence is not applicable.

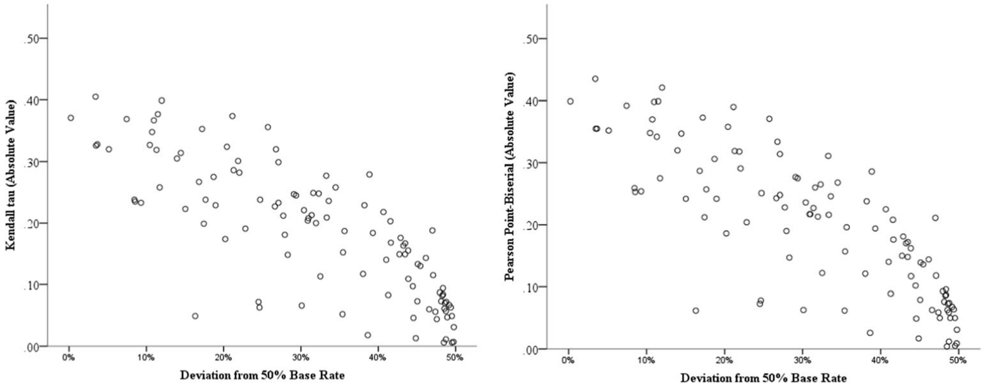

Fig. 1 Influence of base rates on commonly used correlation coefficients

Thorndike adjusted correlation (r = .19), whereas the other statistics demonstrated effects closer to a moderate relationship (polychoric r = .34, AUC = .66, and Cohen's d = 0.59). This effect functions in the same manner for high base rate situations as well. For example, 83 % of the sample had a prior offence. Using Pearson's, Kendall's, and Pearson/Thorndike adjusted correlations, the relationship to the overall risk rating was moderate (Pearson's r = .31; Kendall's tau = .28; Pearson/Thorndike adjusted r = .28). The other effect sizes, however, revealed a large relationship (polychoric r = .49, AUC = .72, Cohen's d = .88).

These examples demonstrate that the attenuating effect of high or low base rates makes an important difference in the conclusions drawn from a research question. Conversely, however, when base rates are closer to the ideal 50 %, then the conclusions were found to be fairly similar, regardless of what statistic is used. For example, 50 % of the sample had a prior violent offence, which presumably should be related to higher risk ratings. For this item, all statistics approached or exceeded the criteria for a moderate effect (Pearson's point-biserial r = .40, Kendall's tau = .37, Pearson/Thorndike adjusted r = .28, polychoric r = .55, AUC = .71, and Cohen's d = 0.87), reflecting less concern about which statistic is used in the case of evenly distributed dichotomous variables.

Table 2 summarizes how the different effect size indicators could lead to different interpretations of the results from the same data. For each of the effect size metrics discussed, the table presents how many effects would be considered trivial (i.e., below the criteria for a small effect), between small and moderate (i.e., at least a small effect size, but not quite moderate), between moderate and large, and those that meet or exceed the criteria for a large effect. The table also presents the highest effect size and the median effect size. For Pearson's and Kendall's point-biserial correlations, none of the effects were considered large, and 31 to 32 of the 109 effects (28–29 % of them) were considered trivial (i.e., less than small). For the other effect sizes, the number of trivial effects varied between 7 for Polychoric correlations (6 % of the effects) and 12 for Pearson/Thorndike adjusted correlations and AUCs (11 %). Using the Pearson/Thorndike adjustment, no effects were considered large, and the median effect (r = .21) was roughly the same as the median Kendall's tau (r = .19) or Pearson point-biserial (r = .21). With polychoric correlations, AUCs, and Cohen's d, at least 12 of the effects (11 %) were considered large, and most effects were at least moderate. Interestingly, although the Pearson/Thorndike adjustment had an exceptionally high correlation with the Cohen's d (r > .99), indicating similarity in the rank ordering of effect sizes, there were substantial differences between these metrics in the magnitude of effect sizes. The median adjusted correlation retained roughly the same value as the median unadjusted correlations, and only four values were considered

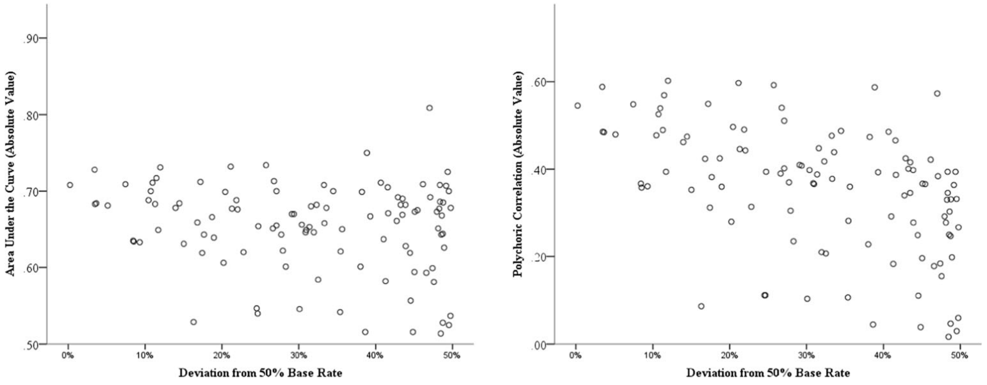

Fig. 2 Influence of base rates on common alternative effect sizes

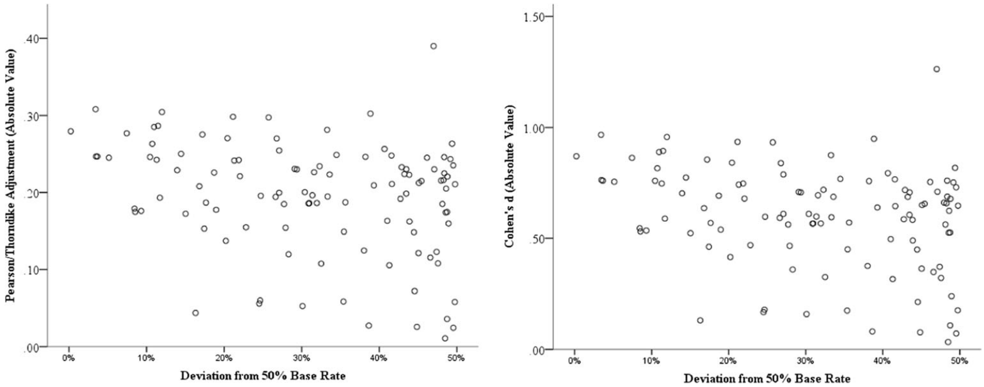

Fig. 3 Influence of base rates on Pearson/Thorndike adjustment and Cohen's *d*

moderate; for Cohen's *d*, however, 80 of the effects were at least moderate in magnitude.

Discussion

The current study utilized real-world data to demonstrate that when the dichotomous variable does not have a 50 % base rate, the correlation is reduced. The further the base rate gets from 50 %, the greater the reduction in the correlation. Our primary research question was whether this base rate sensitivity would make a meaningful difference in the magnitude of the correlation. It did. We found a surprisingly large relationship between the base rate and the correlation. For point-biserial correlations (Pearson's or Kendall's Tau), there was about a −.80 correlation between the effect size and the base rate deviation, meaning that 64 % of the variance in correlations was explained by the base rate. For dichotomous data then, the correlation may be saying a lot more about the base rate than anything else. In particular, once base rates reached less than 10 % or greater than 90 %, correlations generally were unable to exceed .20 (a small effect). Polychoric correlations were less sensitive to base rates, but the relationship still remained fairly large. AUCs appeared to be the most robust statistic (i.e., least influenced by the base rate), although it was not completely independent of the base rate. The Pearson/Thorndike adjusted correlation, Cohen's *d* , and AUC were less sensitive to base rates.

The base rate sensitivity of the correlations could substantively change the interpretation of the results. For example, using a point biserial correlation in our data (either Pearson's or Kendall's tau), we would have concluded that nearly a third of the items on the Static Factors Assessment are trivially related to the final risk rating, and only about one-fifth have a moderate relationship to the final rating (and none have a large relationship). This conclusion would be fairly troubling, given the intent and purpose of the scale. Switching to AUCs for these analyses, however, we would have concluded that only 11 % of the items had a trivial relationship to the final rating, and that two-thirds have moderate or large relationships – these are much more intuitive results given the intended use of the scale.

Even with AUCs, however, there was still a significant and small negative relationship between the base rate deviation and the effect size. Hypothetically, that relationship should be 0. Our findings appear to contradict Rice et al. (2013), who found that AUCs remained fairly stable across diverse follow-up periods, which substantially impacted the base rate of recidivism. A possible reason for the discrepancy in findings may be because Rice and colleagues used data where the

Table 1 Relationship between the deviation from a 50 % base rate and the effect size

Effect size	Deviation from 50 % base rate	Correlation coefficient				
		Pearson point-biserial	Pearson/Thorndike adjustment	Kendall's tau	Polychoric	AUC
Pearson point-biserial	−.810	–	–	–	–	–
Pearson/Thorndike Adjustment	−.284	.715	–	–	–	–
Kendall tau	−.798	.999	.722	–	–	–
Polychoric	−.486	.863	.952	.872	–	–
AUC	−.187	.638	.991	.649	.926	–
Cohen's *d*	−.296	.721	.993	.729	.951	.984

Each effect size is based on 109 observations and all reached statistical significance at $p < .05$

Table 2 Size of effects, based on the statistic used

	Number of effects in each size category					
	Less than small	Small to moderate	Moderate to large	Large or greater	Largest effect size	Median effect size
Pearson point-biserial	31	54	24	0	.44	.21
Kendall Tau	32	57	20	0	.40	.19
Pearson/Thorndike Adjustment	12	93	4	0	.39	.21
Polychoric	7	25	62	15	.60	.38
AUC	12	22	63	12	.81	.67
Cohen's *d*	11	18	64	16	1.26	0.64

Total number of effects = 109. A small effect was a correlation of .10 (Cohen, 1992), an AUC of .56 (Rice & Harris, 2005), or a Cohen's *d* of .20 (Cohen, 1992). A moderate effect was a correlation of .30, an AUC of .64 or a Cohen's *d* of .50. A large effect was a correlation of .50, an AUC of .71, or a Cohen's *d* of .80

two variables remained the same, but the base rate changed purely as a result of methodological adjustments (i.e., follow-up). In our study, the base rate of the dichotomous variable was not artificially changed through the methodology – it changed because the dichotomous variable itself changed (i.e., different risk factors). In these real-world data, it is possible that the residual correlation between base rates and effect sizes reflects a true relationship between infrequent events and the final risk ratings. For example, it is possible that when staff encounter some items that are highly unusual and infrequently endorsed, they may be unsure whether it increases or decreases risk, and consequently these rare items may have less relationship to their final risk ratings. Given this possibility in our data, it is difficult to use the results of this study to attempt to precisely quantify the magnitude of the relationship between base rates and effect sizes. What is more informative are the comparisons of how much the magnitude of the relationship changes depending on which effect size metric is utilized.

Another important limitation of this study was that it examined only the attenuating effect from the dichotomous variable. The second variable (the final risk rating) was an ordinal variable on a three-point scale (low, moderate, and high). Using a trichotomized variable for a hypothetically continuous latent construct would have introduced additional attenuation in our results, which we did not examine. However, given that the attenuating effect of the trichotomous variable was constant for all analyses, it should not affect our findings about the relative contributions of base rates to the effect size metrics we examined.

Recommendations

Given the magnitude of the correlation's base rate sensitivity and its meaningful impact on the interpretation of the results, our study suggests that correlations may not be an ideal effect size measure in research with a dichotomous variable, such as predictive accuracy studies, item-total correlations (when the items are dichotomous), or examining differences between two groups (e.g., Aboriginal and non-Aboriginal offenders, male and female offenders). Importantly, others have argued that base rate sensitivity can be desirable because it is indicative of statistical power and reliability issues (Cohen et al., 2003; McGrath & Meyer, 2006), and they provide benchmarks or adjustments for magnitude of correlations that take into account base rate sensitivity (McGrath & Meyer, 2006; Rice & Harris, 2005). In many situations, however, we believe it is valuable to provide estimates of effect sizes that are insensitive to base rates, particularly when changes in base rates may be arbitrary (e.g., differences in follow-up period) or when comparing findings for groups or variables expected to have differences in base rates (e.g., comparing the predictive accuracy of risk scales for male and female offenders, who may have substantial differences in recidivism rates). For a more nuanced discussion of situations where base rate-independent effect sizes may be more or less desirable, see McGrath and Meyer (2006).

AUCs are more robust to base rates than correlations, are more widely recommended (Rice & Harris, 2005), and are easily computed in many statistical packages, such as SPSS (see "Appendix"). An additional advantage of AUCs is that they are non-parametric, so they can be used for either continuous or ordinal variables as the predictor, without any distributional assumptions (Ruscio, 2008).

Though correlations may be intuitive to interpret, which is appealing for knowledge translation (Gendreau & Smith, 2007), the current study highlights that they are complicated to interpret in light of base rate deviations. To retain the interpretation benefits of using correlations, we recommend considering polychoric, tetrachoric, or Pearson/Thorndike adjusted correlation coefficients, or adjusting the interpretation heuristics of correlations depending on the base rate (e.g., McGrath & Meyer, 2006). Of these, Pearson/Thorndike adjusted correlations were the least influenced by base rates, although they are computationally complex, unavailable in common statistical software thereby requiring hand calculations, and require decisions about what average base rate standard deviation to apply when standardizing the values. Identifying an appropriate average base rate standard deviation is easier when using correlations in a meta-analysis, because numerous studies are available to help identify a plausible average standard deviation (e.g., Bonta et al., 1998; Campbell et al., 2009). In the current study, we had 109 SFA items with varying base rates, which provided a natural sampling from which to produce an average standard deviation. For single studies, however, this decision can appear more arbitrary and may require scanning the literature to form a relevant sampling.

Although using Pearson/Thorndike adjusted correlations (e.g., Ley, 1972) is more practical in meta-analyses, there are other disadvantages to using correlations in this context. Importantly, correlations have a bias in meta-analysis, whereby the variance is inversely affected by the size of the correlation (Borenstein, Hedges, Higgins, & Rothstein, 2009). As a result, studies with larger correlations are given more weight in the meta-analysis (because meta-analyses weight by the inverse of the variance). To help address this, Fisher's *Z* transformations are recommended so that the variance is no longer influenced by the size of the correlation and instead is solely influenced by sample size (Borenstein et al., 2009). In other words, to conduct a meta-analysis using correlations when at least some effects are based on a dichotomous variable and base rate dependence is not desirable, best practice would necessitate first adjusting all correlations, and then transforming to Fisher's *Z* values (although some meta-analysts disagree with the use of Fisher's Z; Hunter & Schmidt, 2004).

Our perspective is that in many contexts, the computational and conceptual complexity inherent in using correlations outweighs the interpretative benefits, and alternative effect sizes should be selected (e.g., if one variable is always dichotomous, use AUCs or Cohen's *d*; if both variables are dichotomous, then use odds ratios). For meta-analysts, standard errors of AUCs are more computationally complicated than Cohen's *d*, are often unreported in primary studies, and there are also different methods of estimating standard errors for AUCs (e.g., depending on the statistical software program being used, or even which version). Given these complexities, we usually favor Cohen's *d* when meta-analysis topics allow for the choice between AUCs or Cohen's *d*s.

For researchers conducting factor analyses with items that are dichotomous or ordinal, recommended best practice is to use polychoric or tetrachoric correlations to develop the correlation matrices underlying the factor analysis (Brown & Benedetti, 1977; Flora & Curran, 2004; Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010; Kubinger, 2003). Using Pearson's correlations would lead to underestimated factor loadings (Holgado-Tello et al., 2010). Kubinger (2003) argues against the use of Pearson's correlation for factor analysis on dichotomous items; he found that, compared to tetrachoric correlations, Pearson's correlations are more likely to lead to artificial factors based on similar distributions (i.e., base rates). Consistent with these recommendations, our study found that polychoric correlations were substantially less sensitive to base rates than Pearson's correlations with dichotomous data. Interestingly, however, the base rate sensitivity remained fairly large. This suggests that using polychoric correlations in factor analysis may be a substantial improvement, but may not fully overcome the base rate dependence of using dichotomous data. Item Response Theory and Structural Equation Modelling for dichotomous data have also been proposed as alternative approaches for factor analysis of dichotomous items (Glockner-Rist & Hoijtink, 2003; Yang, Lu, & Qiao, 2014).

It is important to note that the alternative statistics also have disadvantages. The current study, for example, found that all alternative statistics were still sensitive to base rate deviations, with correlations between base rate deviations and effect sizes ranging from −.19 (AUC) to −.49 (polychoric correlations). Other limitations of these statistics warrant discussion. Cohen's *d*, for example, can be computed on two dichotomous variables but some have argued that this artificially restricts the magnitude of *d* (Fleiss, 1994; Hunter & Schmidt, 2004; but see Sánchez-Meca, Marín-Martínez & Chácon-Moscoso, 2003 for an alternative view). Polychoric correlations may also be limited by the extent to which the latent variables follow the assumed normal distribution (see Maydeu-Olivares, García-Forero, Gallardo-Pujol, & Renom, 2009 for summary of methods of assessing normality of polychoric correlations). Latent trait analyses allow for this normality assumption to be relaxed and, as such, some argue that polychoric correlations do not necessitate underlying continuous latent variables (Finney & DiStefano, 2006; Wall, Guo, & Amemiya, 2012).

Odds ratios are common statistics when analyzing the association between two dichotomous items, but some have argued that they have limited clinical application because physicians often misinterpret them as rate ratios (Sackett, Deeks, & Altmans, 1996; see Helmus & Hanson, 2011 for a primer on odds ratios) and odds ratios may not be the best statistics to index a test's classification accuracy, at least compared to AUCs (Pepe, Janes, Longton, Leisenring, & Newcomb, 2004). Given their instability in the presence of small cell sizes, they also require decisions about sample size adjustments, such as adding .5 to all cells (e.g., Fleiss, 1994). AUCs are also influenced by restriction of range in the predictor variable (Hanson, 2008; see Lobo, Jiménez-Valverde, & Real (2008) for additional criticism of AUCs).

Conclusions

Base rates directly influence the magnitude of the correlation when using dichotomous data. Given the large magnitude of this base rate dependence for point-biserial correlations ($r = -.81$) and Kendall's tau ($r = -.80$) in our dataset, we recommend that point-biserial and Kendall's tau correlations should be avoided for analyses with dichotomous data, unless considering base rate variations is desirable for the analyses (in which case, interpretations should consider the base rate). In most situations, AUCs, Pearson/Thorndike adjusted correlations, Cohen's *d*, or polychoric correlations are likely preferable. AUCs are most (but not necessarily fully) robust to base rate variations. Additionally, for the purpose of comparing effect sizes across studies (informally or through meta-analysis), numerous resources exist for converting and/or comparing these different metrics (Borenstein et al., 2009; Hunter & Schmidt, 2004; Rice & Harris, 2005). All statistics have unique advantages and disadvantages, and researchers should consider these when deciding on appropriate effect size metrics.

Acknowledgments Funding for this project was provided in part by the Canadian Institute for Health Research (CIHR Banting Fellowship). The views expressed are those of the authors and not necessarily those of the Correctional Service of Canada.

Appendix

Table 3 Available software programs to compute alternative statistics

Statistics	Software
AUC	GraphPad Prism 6
	MedCal
	R environment (programs AUC or pROC)
	SAS
	SPSS
	STATA
	UNISTAT
	XLStats
Cohen's *d*	R environment (program effsize)
	STATA (program esize)
Polychoric correlations	FACTOR
	Mplus
	R environment (polycor)
	SAS (macro %POLYCHOR)
	SPSS (macros r_tetra or TetCorr)
	STATA (macro polychoric)
	TetMat

Some software programs required supplementary downloads to be able to compute the effect size. These programs and macros are indicated in parentheses beside the software program. Cohen's *d* can also be easily calculated using means and standard deviations

References

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Newark, NJ: LexisNexus/Anderson.

Andrews, D. A., Guzzo, L., Raynor, P., Rowe, R. C., Rettinger, L. J., Brews, A., & Wormith, J. S. (2012). Are the major risk/need factors predictive of both female and male reoffending? A test with the eight domains of the Level of Service/Case Management Inventory. *International Journal of Offender Therapy and Comparative Criminology, 56,* 113–133. doi:10.1177/0306624X10395716

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28*(3), 369–404.

Aron, A., Aron, E., & Coups, E. J. (2006). *Statistics for psychology.* Upper Saddle River, NJ: Pearson Prentice Hall.

Association for the Treatment of Sexual Abusers. (2014). *ATSA practice guidelines for assessment, treatment interventions, and management strategies for male adult sexual abusers*. Beaverton, OR: Professional Issues Committee, ATSA.

Babchishin, K. M., Blais, J., & Helmus, L. (2012). Do static risk factors predict differently for Aboriginal sex offenders? A multi-site comparison using the original and revised Static-99 and Static-2002 scales. *Canadian Journal of Criminology and Criminal Justice, 54,* 1–43. doi:10.3138/cjccj.2010.E.40

Bonta, J., Law, M., & Hanson, R. K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123,* 123–142. doi:10.1037/0033-2909.123.2.123

Bonta, J., Rugge, T., Scott, T.-L., Bourgon, G., & Yessine, A. K. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation, 47,* 248–270. doi:10.1080/10509670802134085

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, United Kingdom: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford.

Brown, M. B., & Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika, 42,* 347–355. doi:10.1007/BF02293655

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36,* 567–590. doi:10.1177/0093854809333610

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253. doi:10.1177/014662168300700301

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Coolican, H. (2013). *Research methods and statistics in psychology.* New York, NY: Routledge.

Correctional Service of Canada. (2012-06-13a). *Correctional planning and criminal profile. Commissioner's Directive 705-6*. Ottawa, ON: Author.

Eysenck, M. W. (2013). *Simply psychology* (3rd ed.). London, UK: Taylor & Francis Group, Psychology Press.

Field, A., & Miles, J. (2010). *Discovering statistics using SAS.* London: Sage.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, 269–314.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491. doi:10.1037/1082-989X.9.4.466

French, S. A., & Gendreau, P. (2006). Reducing prison misconducts: What works! *Criminal Justice and Behavior, 33,* 185–218.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34,* 575–608.

Gendreau, P., & Smith, P. (2007). Influencing the "people who count": Some perspectives on the reporting of meta-analytic results for prediction and treatment outcomes for offenders. *Criminal Justice and Behavior, 34,* 1536–1559.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioural sciences*. San Francisco, CA: W. H Freeman and Company.

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10,* 544–565. doi:10.1207/S15328007SEM1004_4

Gonçalves, L. C., Gonçalves, R. A., Martins, C., & Dirkzwager, A. J. E. (2014). Predicting infractions and health care utilization in prison: A meta-analysis. *Criminal Justice and Behavior, 41,* 921–942. doi:10.1177/0093854814524402

Hanson, R. K. (2008). What statistics should we use to report predictive accuracy. *Crime Scene, 15*(1), 15–17.

Helmus, L. M., & Babchishin, K. M. (2015). *Primer on risk assessment and the statistics used to evaluate its accuracy.* Unpublished manuscript.

Helmus, L., Babchishin, K. M., & Blais, J. (2012). Predictive accuracy of dynamic risk factors for Aboriginal and non-Aboriginal sex offenders: A comparison using STABLE-2007. *International Journal of Offender Therapy and Comparative Criminology, 56,* 856–876. doi:10.1177/0306624X11414693

Helmus, L., & Forrester, T. (2014). *Construct validity of the Static Factors Assessment*. Unpublished manuscript. Ottawa, ON: Correctional Service of Canada.

Helmus, L., & Hanson, R. K. (2011). More fun with statistics! How to use logistic regression to predict criminal recidivism risk. *Crime Scene, 18*(2), 8–12. Retrieved from http://www.cpa.ca/docs/file/Sections/Criminal%20Justice%20Psychology/Crime%20Scene%202011-10.pdf

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearsons correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44,* 153–166. doi:10.1007/s11135-008-9190-y

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Karabinus, R. A. (1975). The *r*-point biserial limitation. *Educational and Psychological Measurement, 35,* 277–282.

Karl, T. R., Groisman, P. Y., Knight, R. W., & Heim, R. R. (1993). Recent variations of snow cover and snowfall in North America and their relation to precipitation and temperature variations. *Journal of Climate, 6,* 1327–1344. doi:10.1175/1520-0442

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science, 45,* 106–110.

Ley, P. (1972). *Quantitative aspects of psychological assessments: An introduction*. London, UK: Duckworth.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17,* 145–151. doi:10.1111/j.1466-8238.2007.00358.x

Maydeu-Olivares, A., García-Forero, C., Gallardo-Pujol, D., & Renom, J. (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5,* 131.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of the *r* and *d*. *Psychological Methods, 11,* 386–401.

McPhail, I. V., Hermann, C. A., & Fernandez, Y. M. (2014). Correlates of emotional congruence with children in sexual offenders against children: A test of theoretical models in an incarcerated sample. *Child Abuse and Neglect, 38,* 336–346. doi:10.1016/j.chiabu.2013.10.002

Motiuk, L. L. (1993). Where are we in our ability to assess risk? *Forum on Corrections Research, 5*(2), 14–18.

Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale – Sexual Offender Version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment, 19,* 318–329. doi:10.1037/1040-3590.19.3.318

Pearson, K. (1903). Mathematical contributions to the theory of evolution: II. On the influence of natural selection on the variability and correlation of organs. *Royal Society Philosophical Transactions, 200*(series A), 1–66. Available from https://archive.org/details/philtrans02398796

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology, 159,* 882–890. doi:10.1093/aje/kwh101

Pintea, S., & Moldovan, R. (2009). The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9,* 49–66.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r. Law and Human Behavior, 29,* 615–620. doi:10.1007/s10979-005-6832-7

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the *VRAG* and *SORAG*: *The Violence Risk Appraisal Guide – Revised (VRAG-R*). *Psychological Assessment, 25,* 951–965. doi:10.1037/a0032878

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30. doi:10.1037/1082-989X.13.1.19

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence-Based Medicine, 1,* 164–166.

Sánchez-Meca, J., Chacón-Moscoso, S., & Marín-Martínez, F. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8,* 448–467. doi:10.1037/1082-989X.8.4.448

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

Viljoen, J. L., Mordell, S., & Beneteau, J. L. (2012). Prediction of adolescent sexual reoffending: A meta-analysis of the J-SOAP-II, ERASOR, J-SORRAT-II, and Static-99. *Law and Human Behavior, 36,* 423–438. doi:10.1037/h0093938

Wall, M. M., Guo, J., & Amemiya, Y. (2012). Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivariate Behavioral Research, 47,* 276–313. doi:10.1080/00273171.2012.658339

Yang, Y., Lu, X., & Qiao, H. (2014). A robust factor analysis model for dichotomous data. *Journal of Systems Science and Information, 2,* 437–450.