Development and Monte Carlo Study of a Procedure for Correcting the Standardized Mean Difference for Measurement Error in the Independent Variable

Educational and Psychological
Measurement

2015, Vol. 75(2) 284–310

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414534067

epm.sagepub.com

William Robert Nugent¹, Matthew Moore¹, and Erin Story¹

Abstract

The standardized mean difference (SMD) is perhaps the most important meta-analytic effect size. It is typically used to represent the difference between treatment and control population means in treatment efficacy research. It is also used to represent differences between populations with different characteristics, such as persons who are depressed and those who are not. Measurement error in the independent variable (IV) attenuates SMDs. In this article, we derive a formula for the SMD that explicitly represents accuracy of classification of persons into populations on the basis of scores on an IV. We suggest an alternate version of the SMD less vulnerable to measurement error in the IV. We derive a novel approach to correcting the SMD for measurement error in the IV and show how this method can also be used to reliability correct the unstandardized mean difference. We compare this reliability correction approach with one suggested by Hunter and Schmidt in a series of Monte Carlo simulations. Finally, we consider how the proposed reliability correction method can be used in meta-analysis and suggest future directions for both research and further theoretical development of the proposed reliability correction method.

¹The University of Tennessee, Knoxville, TN, USA

Corresponding Author:

William Robert Nugent, The University of Tennessee, 308 Henson Hall, Knoxville, TN 37996-3333, USA.
Email: wnugent@utk.edu

Keywords

standardized mean difference, meta-analysis, correcting for measurement error, reliability correction

The standardized mean difference (SMD) is one of the most important effect sizes in meta-analysis. The SMD is a two variable effect size (EFS) used to compare the means of different populations on a dependent variable (DV), with population membership indicated by a dichotomous independent variable (IV), neither of which may be measured the same across studies. The SMD is commonly used to represent outcomes in research examining the efficacy of psychological, educational, and medical interventions and treatments (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001). The SMD is also used to represent differences on a DV between populations composed of persons with different characteristics (Grissom & Kim, 2011), such as males and females in gender differences research (Hedges & Olkin, 1985), and differences between persons who are depressed and those who are not (e.g., Snyder, 2013).

A number of artifact adjustments have been suggested for EFSs in meta-analysis (Hunter & Schmidt, 2004; Schmidt, Le, & Oh, 2009). Lipsey and Wilson (2001) suggest the most useful are corrections for the effects of measurement error. While the effects of measurement error in the DV on the SMD are commonly addressed (e.g., Hedges & Olkin, 1985; Lipsey & Wilson, 2001), the effects of measurement error in the IV are less frequently considered. Hunter and Schmidt (2004), in perhaps the most extensive consideration of this topic, observed that measurement error in the IV: (a) decreases the difference between the population means in the numerator of the SMD and (b) increases the within population variances of scores on the DV in the denominator of the SMD, causing attenuation of the SMD.

The attenuation of SMDs because of measurement error in the IV can have deleterious effects on meta-analysis (Orwin & Cordray, 1985). If different measures of the IV, with differing levels of measurement error, are used in a series of studies, the SMDs for these studies will have differential levels of attenuation. This differential attenuation will propagate through meta-analyses. For example, tests of homogeneity will be affected. A test of homogeneity of differentially attenuated SMDs can suggest heterogeneity, even when the set of SMDs free of the effects of measurement error are truly homogeneous (Hedges & Olkin, 1985). Correlations between differentially attenuated SMDs and explanatory variables will also be attenuated, affecting meta-regression analyses. These considerations underscore the importance of development and use of methods for disattenuating the SMD for the effects of measurement error in the IV (Hedges & Olkin, 1985).

Hunter and Schmidt (2004) sketched a method for correcting the SMD for the effects of measurement error in the IV based on the classical theory correction for attenuation and formulas for converting the SMD to the point-biserial correlation, and vice versa. This approach could be implemented as in the following example.

Suppose a researcher is interested in the relationship between major depressive disorder (MDD) and cognitive deficits (Snyder, 2013). The 1-year prevalence of MDD in the United States is about .09 (Kazdin, 2002). Assume the interview procedure used to classify persons as having, or not having, MDD has sensitivity .717 and specificity .90, mean values for interview methods from a recent review (Swedish Council on Health Technology Assessment, 2012). These sensitivity and specificity values, combined with the prevalence of .09, imply at the population level 84.5% of persons will be classified as not having, and 15.5% as having, MDD (Pepe, 2003); and imply the square root of the reliability coefficient for classifications based on scores from this IV will be about .487 (Phi correlation between observed and true classification; Nunnally, 1978).

Assume in the researcher's study 84.5% of participants are classified as not having, and 15.5% as having, MDD; that the difference between the group means on the DV for those with MDD and those without is +1.95; and the variances of scores on the DV are equal to 109.2 in both MDD and non-MDD groups. The researcher computes the sample SMD, obtaining .187, and then converts this to a point-biserial correlation using formula 3.34 in Lipsey and Wilson (2001), obtaining a value of .067. This is divided by .487 (Hunter and Schmidt, 2004), giving a reliability corrected point-biserial of .139. This is transformed back to a SMD using formula 3.36 from Lipsey and Wilson (2001), giving a reliability corrected SMD of .387. A little algebra shows the reliability corrected SMD obtained using this three step procedure can be condensed into the formula,

$$\text{Reliability corrected SMD} = \frac{\widehat{SMD}}{\sqrt{\text{rel } X + \left((p_0 \times p_1) \times \widehat{SMD}^2 \times (\text{rel } X - 1) \right)}}$$

where $\widehat{SMD}$ is the sample estimated SMD; $\text{rel } X$ is the reliability coefficient for the classification of persons into groups; p_0 the proportion of persons in the non-MDD group; and p_1 the proportion in the MDD group. This example will be considered again later.

In this article, we focus on correcting the SMD in meta-analysis for measurement error in the IV. We expand on Hunter and Schmidt's (2004) examination of the effects of measurement error in the IV on the SMD, and their development of methodology for correcting the SMD for this error. We first conceptualize and formalize measurement of an IV used to classify persons into different populations with respect to their possession of a characteristic of interest, such as depression. We derive a formula for the population SMD that includes representation, in both numerator and denominator, of the accuracy of classification of persons, and which explicitly represents attenuation of the numerator because of measurement error in the IV. We then use a model-based simulation (Axelrod, 2007; Banks, 2009) to examine the effects, on the SMD, of misclassification of persons into populations. The results elaborate Hunter and Schmidt's (2004) observations, showing measurement error in the IV can

attenuate the SMD to a greater degree than measurement error in the DV. We next propose a novel method for disattenuating the SMD for the effects of measurement error in the IV. We compare this proposed method with the method described by Hunter and Schmidt in a series of Monte Carlo simulations. We conclude by considering the following:

- How the proposed method can be used in meta-analysis
- Further theoretical development of the proposed reliability correction method
- Implications of the results of the Monte Carlo simulations for future research on the proposed reliability correction method

Measurement of an IV for Classification

True Population Membership

Individuals are classified into populations based on measurement of an IV. The IV is measured and persons classified, based on IV scores, into different populations. Figure 1 helps conceptualize this measurement. At the top of this figure are two populations, P'_0 and P'_1 . Population P'_1 , conceptualized as “true population P_1 ,” is composed of persons truly possessing some characteristic as indicated by the true scores τ_ψ from a measure ψ of the IV; these persons are, for example, truly depressed. Population P'_0 , “true population P_0 ,” is composed of persons who truly *do not* possess the characteristic; for example, these persons are really *not* depressed. The dashed oval at the top of the figure shows the combined population, $P'_0 \cup P'_1$. Population $P'_0 \cup P'_1$ might be all persons in the United States, and in this population, P'_0 and P'_1 are subpopulations. The proportion of persons in $P'_0 \cup P'_1$ who are members of P'_1 based on the scores τ_ψ is symbolized by $p(\psi)_{P'_1}$, and is the *prevalence* in $P'_0 \cup P'_1$ of the characteristic of interest (e.g., depression). The proportion of persons in $P'_0 \cup P'_1$ who are members of P'_0 is, $p(\psi)_{P'_0} = 1 - p(\psi)_{P'_1}$.

Observed Population Membership

In practice, persons from $P'_0 \cup P'_1$ must be classified, using the measurement procedure, into an observed population of persons who by empirical assessment possess the characteristic of interest, call this population, “observed P_1 ,” or simply P_1 ; and an observed population of persons not possessing the characteristic, call this population “observed P_0 ,” or P_0 . The union of observed P_1 and P_0 is the observed combined population, $P_0 \cup P_1$, shown by the oval encompassing P_0 and P_1 . Observed P_0 and P_1 are subpopulations of $P_0 \cup P_1$. Figure 1 shows the classification of persons from $P'_0 \cup P'_1$ into P_0 and P_1 based on the observed scores X_ψ from measure ψ . In observed P_1 in Figure 1, the symbol, $P'_1 : P_1$, which reads “ P'_1 nested within P_1 ,” represents a subpopulation of P_1 made up of persons who are P'_1 members who have been correctly classified. The subpopulation $P'_0 : P_1$ represents a subpopulation composed of

Figure 1. Illustration of “true populations” P_1^t and P_0^t composed, respectively, of all persons who truly possess a characteristic of interest and those who truly do not possess this characteristic; and the combined population, $P_0^t \cup P_1^t$. Also shown are observed subpopulations P_1 and P_0 created by using the observed scores X_ψ from measure ψ of the independent variable to classify persons into P_1 and P_0 ; the observed combined population $P_0 \cup P_1$; and the subpopulations, $P_0^t : P_0$, $P_1^t : P_0$, $P_0^t : P_1$, and $P_1^t : P_1$, of P_0 and P_1 , respectively, created by classification of persons into P_0 and P_1 .

P_0^t members misclassified into P_1 . Subpopulation $P_0^t : P_0$ represents a subpopulation of P_0 composed of P_0^t members correctly classified. Finally, subpopulation $P_1^t : P_0$ is a subpopulation of P_0 composed of P_1^t members misclassified into P_0 .

Population and Subpopulation Means and Variances

The symbol $\mu_{Y_\alpha}^{P_1^t:P_1}$ represents the mean observed score on the DV in subpopulation $P_1^t : P_1$, and $\mu_{Y_\alpha}^{P_0^t:P_1}$ that in subpopulation $P_0^t : P_1$, where Y_α represents observed scores from measure α of the DV. Similarly, $\mu_{Y_\alpha}^{P_0^t:P_0}$ is the mean DV score in $P_0^t : P_0$ and $\mu_{Y_\alpha}^{P_1^t:P_0}$ is that in subpopulation $P_1^t : P_0$. The mean DV score in P_1^t is $\mu_{Y_\alpha}^{P_1^t}$, while that in P_1 is $\mu_{Y_\alpha}^{P_1}$; and the mean in P_0^t is $\mu_{Y_\alpha}^{P_0^t}$, while that in P_0 is $\mu_{Y_\alpha}^{P_0}$. The variances of DV scores are represented similarly. The variance of scores Y_α in P_1^t is $\sigma^2(Y_\alpha)_{P_1^t}$; in P_0^t it is $\sigma^2(Y_\alpha)_{P_0^t}$; in P_1 it is $\sigma^2(Y_\alpha)_{P_1}$; and in P_0 it is $\sigma^2(Y_\alpha)_{P_0}$. The variance in

subpopulation $P_1^t : P_1$ is $\sigma^2(Y_{\alpha})_{P_1^t : P_1}$, and so forth for subpopulations $P_0^t : P_1, P_1^t : P_0$, and $P_0^t : P_0$.

Reliability of Classification

Hunter and Schmidt (2004) noted that appropriate reliability coefficients need to be used when correcting EFSSs for measurement error. The classical reliability coefficient can be misleading for representing measurement error when scores are used for classification. In this case “reliability” is better represented by quantities indicating classification accuracy (Berk, 1980; Brennan, 2001; Divgi, 1980; Haertel, 2006; Kane & Brennan, 1980). Two population specific indices for representing classification accuracy are the *sensitivity*, or *true positive fraction* (TPF), and the *specificity*, or *true negative fraction* (TNF) (Pepe, 2003). The sensitivity, $sens(X_{\psi})$, or $TPF(X_{\psi})$, specific to population P_1^t is

$$sens(X_{\psi}) = TPF(X_{\psi}) = p(\text{classification} = P_1 | \text{true membership} = P_1^t),$$

the conditional probability a person is correctly identified as possessing the characteristic of interest using the scores X_{ψ} , given he or she is a true member of P_1^t . It is the fraction of persons in P_1^t correctly identified as having the characteristic of interest. The *false negative fraction*, $FNF(X_{\psi}) = 1 - sens(X_{\psi})$, is the fraction of members of P_1^t erroneously inferred to not have the characteristic of interest. The specificity, $spec(X_{\psi})$, or $TNF(X_{\psi})$, specific to population P_0^t is

$$spec(X_{\psi}) = TNF(X_{\psi}) = p(\text{classification} = P_0 | \text{true membership} = P_0^t),$$

the conditional probability a person is correctly identified as *not* possessing the characteristic of interest using the scores X_{ψ} , given he or she is a true member of P_0^t . It is the proportion of persons in P_0^t correctly identified as *not* having the characteristic of interest. The *false positive fraction*, $FPF(X_{\psi}) = 1 - spec(X_{\psi})$, is the proportion of members of P_0^t erroneously inferred to have the characteristic of interest.

Two other indices indicate the accuracy of classification into P_0 and P_1 . The ratio

$$ppv(X_{\psi}) = \frac{p(\psi)_{P_1^t} \times sens(X_{\psi})}{\left(p(\psi)_{P_1^t} \times sens(X_{\psi})\right) + \left[\left(1 - p(\psi)_{P_1^t}\right) \times (1 - spec(X_{\psi}))\right]} \quad (1)$$

is the *positive predictive value* (PPV) of classification of persons into P_1 , based on the scores X_{ψ} . The PPV can be interpreted as the probability a person is in fact a member of P_1^t given he or she has been classified into P_1 ; it is the proportion of persons classified into P_1 who are correctly classified. Similarly, the ratio

$$npv(X_{\psi}) = \frac{\left(1 - p(\psi)_{P_1^t}\right) \times spec(X_{\psi})}{\left(\left(1 - p(\psi)_{P_1^t}\right) \times spec(X_{\psi})\right) + \left[p(\psi)_{P_1^t} \times (1 - sens(X_{\psi}))\right]} \quad (2)$$

is the *negative predictive value*, NPV, of classification of persons into P_0 based on the scores X_ψ . It can be interpreted as the probability a person is truly a member of P_0^t given he or she has been classified into P_0 ; equivalently, it is the proportion of persons classified into P_0 who are correctly classified (Pepe, 2003).

Misclassification Due Only to Random Measurement Error

Assume misclassification of persons into P_1 and P_0 is due only to random measurement error. There are no systematic classification errors. Also assume that a probability of misclassification can be assigned to each person in P_0^t and to each person in P_1^t . Further assume each person in P_0^t has the same probability of being misclassified, $FPF(X_\psi)$; and each person in P_1^t has the same probability of being misclassified, $FNF(X_\psi)$. These assumptions are maintained throughout following argument. Unequal probabilities of, and systematic errors of, misclassification are discussed later.

Thus, persons who are members of P_1^t , who if correctly classified would be placed into P_1 , are effectively randomly selected, as a consequence of random measurement error, to be erroneously classified into P_0 in subpopulation $P_1^t : P_0$. Likewise, persons who are members of P_0^t , and correctly belong in P_0 , are randomly selected as a result of random measurement error to be erroneously classified into P_1 in subpopulation $P_0^t : P_1$. Thus, expected values of the means and variances of scores on the DV in subpopulations $P_1^t : P_1$ and $P_1^t : P_0$ will be the same as in P_1^t , so $\mu_{Y_\alpha}^{P_1^t:P_1} = \mu_{Y_\alpha}^{P_1^t:P_0} = \mu_{Y_\alpha}^{P_1^t}$ and $\sigma^2(Y_\alpha)_{P_1^t:P_1} = \sigma^2(Y_\alpha)_{P_1^t:P_0} = \sigma^2(Y_\alpha)_{P_1^t}$. Similarly, the means and variances in subpopulations $P_0^t : P_0$ and $P_0^t : P_1$ will be the same as in P_0^t , so $\mu_{Y_\alpha}^{P_0^t:P_0} = \mu_{Y_\alpha}^{P_0^t:P_1} = \mu_{Y_\alpha}^{P_0^t}$ and $\sigma^2(Y_\alpha)_{P_0^t:P_0} = \sigma^2(Y_\alpha)_{P_0^t:P_1} = \sigma^2(Y_\alpha)_{P_0^t}$ (Thompson, 2002).

Equality of Collective Populations $P_0^t \cup P_1^t$ and $P_0 \cup P_1$

Consider again Figure 1. The persons in $P_0^t \cup P_1^t$ are the same as those in $P_0 \cup P_1$. For example, assume $P_0^t \cup P_1^t$ is the population of students at a particular university. No matter how the population $P_0^t \cup P_1^t$ of students is arranged into subpopulation P_1 (students who are “depressed”) and subpopulation P_0 (students who are “not depressed”), thereby creating observed $P_0 \cup P_1$ —regardless of how much error there is in classifying students into subpopulations P_0 and P_1 —populations $P_0 \cup P_1$ and $P_0^t \cup P_1^t$ are the same. Since $P_0 \cup P_1$ is the same population as $P_0^t \cup P_1^t$, then $\sigma^2(Y_\alpha)_{P_0 \cup P_1} = \sigma^2(Y_\alpha)_{P_0^t \cup P_1^t}$. As long as no persons leave, or new persons are added to $P_0^t \cup P_1^t$, or $P_0 \cup P_1$, and assuming measuring the IV and classifying persons into observed P_1 and P_0 does not affect the DV, the equality $\sigma^2(Y_\alpha)_{P_0 \cup P_1} = \sigma^2(Y_\alpha)_{P_0^t \cup P_1^t}$ will hold.

Two Versions of the Population SMD

The “Common” Population SMD

The population SMD is traditionally defined as the difference between the means of populations P_1 and P_0 divided by the square root of the mean within population P_1

and P_0 variance (Borenstein, 2009; Rosenthal, 1994). This version of the population SMD, referred to subsequently as the “common” SMD and symbolized by $\delta(Y_\alpha, X_\psi)_{common}$, can be expressed as

$$\delta(Y_\alpha, X_\psi)_{common} = \frac{\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0}}{\sqrt{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1}}}, \quad (3)$$

where $p(\psi)_{P_1} = (p(\psi)_{P_1} \times \text{sens}(X_\psi)) + p(\psi)_{P_0} (1 - \text{spec}(X_\psi))$ is the proportion of persons, at the population level, from $P_0' \cup P_1'$ classified into P_1 , and

$$p(\psi)_{P_0} = (p(\psi)_{P_0} \times \text{spec}(X_\psi)_{SP'}) + p(\psi)_{P_1} (1 - \text{sens}(X_\psi))$$

is the proportion from $P_0' \cup P_1'$ classified into P_0 , based on the observed scores X_ψ (Pepe, 2003); the variance $\sigma^2(Y_\alpha)_{P_1}$ is given by (see the appendix for proof)

$$\begin{aligned} \sigma^2(Y_\alpha)_{P_1} = & \left[ppv(X_\psi) \sigma^2(Y_\alpha)_{P_1'} + (1 - ppv(X_\psi)) \sigma^2(Y_\alpha)_{P_0'} \right] \\ & + \left[ppv(X_\psi) \left(\mu_{Y_\alpha}^{P_1'} - \mu_{Y_\alpha}^{P_1} \right)^2 + (1 - ppv(X_\psi)) \left(\mu_{Y_\alpha}^{P_0'} - \mu_{Y_\alpha}^{P_1} \right)^2 \right], \end{aligned} \quad (4)$$

and the variance $\sigma^2(Y_\alpha)_{P_0}$ by

$$\begin{aligned} \sigma^2(Y_\alpha)_{P_0} = & \left[(1 - npv(X_\psi)) \sigma^2(Y_\alpha)_{P_1'} + npv(X_\psi) \sigma^2(Y_\alpha)_{P_0'} \right] \\ & + \left[(1 - npv(X_\psi)) \left(\mu_{Y_\alpha}^{P_1'} - \mu_{Y_\alpha}^{P_0} \right)^2 + npv(X_\psi) \left(\mu_{Y_\alpha}^{P_0'} - \mu_{Y_\alpha}^{P_0} \right)^2 \right]. \end{aligned} \quad (5)$$

The perhaps complex symbolism for the common population SMD is used to indicate it is based on observed scores X_ψ from measure ψ of the IV and the scores Y_α from measure α of the DV.

As proven in the appendix, the numerator in Equation (3) can be expressed as $(ppv(X_\psi) + npv(X_\psi) - 1) \left(\mu_{Y_\alpha}^{P_1'} - \mu_{Y_\alpha}^{P_0'} \right)$, so the common SMD can be written as

$$\delta(Y_\alpha, X_\psi)_{common} = \frac{(ppv(X_\psi) + npv(X_\psi) - 1) \left(\mu_{Y_\alpha}^{P_1'} - \mu_{Y_\alpha}^{P_0'} \right)}{\sqrt{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1}}}. \quad (6)$$

The common SMD expressed by Equations (3) through (6) formalizes observations of Hunter and Schmidt (2004). The effects of measurement error in the IV on both numerator and denominator are explicitly represented by $ppv(X_\psi)$ and $npv(X_\psi)$, which are functions of the sensitivity, specificity, and prevalence. As will be discussed below, the term $(ppv(X_\psi) + npv(X_\psi) - 1)$ explicitly represents attenuation in the numerator due to the effects of measurement error in the IV.

An Alternate Version of the Population SMD

The use of *SDs* other than that in the denominator of the common SMD has been suggested. For example, the *SD* of scores on the DV in the control group in studies of treatment efficacy has been suggested (Hunter & Schmidt, 2004). An alternative with important advantages is the *SD* of scores on the DV in the combined population $P_0 \cup P_1$, $\sigma(Y_\alpha)_{P_0 \cup P_1}$. As noted above, $\sigma^2(Y_\alpha)_{P_0 \cup P_1} = \sigma^2(Y_\alpha)_{P'_0 \cup P'_1}$, so it follows that $\sigma(Y_\alpha)_{P_0 \cup P_1} = \sigma(Y_\alpha)_{P'_0 \cup P'_1}$. Thus, a principal advantage of this *SD* in the denominator of the SMD is it will not be affected by measurement error in the IV. Returning to the example from above, no matter how the population of students $P'_0 \cup P'_1$ at a particular university is classified into subpopulations who are “depressed” (P_1) and who are “not depressed (P_0)”—regardless of how much error there is in classifying students as “not depressed” or “depressed—” the populations $P_0 \cup P_1$ and $P'_0 \cup P'_1$ contain the same persons, so $\sigma(Y_\alpha)_{P_0 \cup P_1} = \sigma(Y_\alpha)_{P'_0 \cup P'_1}$. As will be shown below, this simplifies correcting the population SMD for the effects of measurement error in the IV, as only the numerator needs disattenuation.

Let an alternate version of the population SMD, with the *SD* of scores in the population $P_0 \cup P_1$ in the denominator, be symbolized as $\delta(Y_\alpha, X_\psi)_{alternate}$. Figure 1 and foregoing argument imply the following expression for this alternate version:

$$\delta(Y_\alpha, X_\psi)_{alternate} = \frac{\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0}}{\sqrt{\sigma^2(Y_\alpha)_{P_0 \cup P_1}}}, \quad (7a)$$

$$= \frac{(ppv(X_\psi) + npv(X_\psi) - 1) \left(\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P'_0} \right)}{\sqrt{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1} + p(\psi)_{P_0} (\mu_{Y_\alpha}^{P_0} - \mu_{Y_\alpha}^{P_0 \cup P_1})^2 + p(\psi)_{P_1} (\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0 \cup P_1})^2}}, \quad (7b)$$

where $\mu_{Y_\alpha}^{P_0 \cup P_1}$ is the mean score on Y_α in population $P_0 \cup P_1$.

Relationship Between Common and Alternate Versions of the SMD

It is straightforward to show the relationship between the common and alternate versions of the SMD is given by Equation (8),

$$\begin{aligned} \delta(Y_\alpha, X_\psi)_{common} &= \delta(Y_\alpha, X_\psi)_{alternate} \sqrt{\frac{\sigma^2(Y_\alpha)_{P_0 \cup P_1}}{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1}}} \\ &= \delta(Y_\alpha, X_\psi)_{alternate} \\ &\quad \sqrt{\frac{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1} + p(\psi)_{P_0} (\mu_{Y_\alpha}^{P_0} - \mu_{Y_\alpha}^{P_0 \cup P_1})^2 + p(\psi)_{P_1} (\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0 \cup P_1})^2}{p(\psi)_{P_0} \sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1} \sigma^2(Y_\alpha)_{P_1}}}, \end{aligned} \quad (8)$$

Figure 2. Graph of values of common (solid lines) and alternate (dashed curves) versions of the population SMD as a function of the between population mean difference, $\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0}$, for four different values of the mean within population variance: 50 (uppermost curves), 100 (two curves immediately under two uppermost curves), 200 (next to lowermost two curves), and 300 (two lowermost curves).

where $\sigma^2(Y_\alpha)_{P_1}$ and $\sigma^2(Y_\alpha)_{P_0}$ are given by Equations (4) and (5). The alternate version will be less than or equal to the common version. The two versions will be equal only when $\mu_{Y_\alpha}^{P_0} - \mu_{Y_\alpha}^{P_0 \cup P_1} = \mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0 \cup P_1} = 0$.

Figure 2 shows a plot of common and alternate SMDs as a function of the difference between the means of P_1 and P_0 , $\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0}$. This difference is scaled on the horizontal axis, while SMD values are on the vertical axis, with common SMD values shown by solid curves and alternate version values by dashed curves. The assumption was made in this graph that the mean within population P_1 and P_0 variances were the following: 50 (uppermost curves), 100 (next uppermost curves), 200 (next to lowermost curves), and 300 (lowermost curves).

The values of the two versions of the population SMD are nearly identical in this graph up to common SMD values of about .75 (marked by the dotted horizontal line), with differences between the two versions of .05 or less. The values differentially and increasingly diverge from this point, with the magnitude of the divergence a function of the magnitude of the within population variances; the smaller the mean within population variance, the greater the divergence. These differences between the two SMD versions are considered below.

The Effects of Measurement Error in the IV: A Simulation

A model-based simulation was conducted to investigate the magnitude by which the population common SMD is attenuated by measurement error in the IV. A model-

Figure 3. Graph showing the attenuation in the population common SMD due to differing levels of measurement error in the IV. The solid curves show the attenuation in the SMD given the prevalence of the characteristic differentiating membership in P_1^t , from that in P_0^t , and is .50, and the dash-dash-dash marked curves that when the prevalence is .09. The dash-dot-dot curve shows the effects of measurement error in the DV on the common SMD.

based simulation uses a mathematical model to investigate the behavior of some real-world system or method under specified conditions (Axelrod, 2007; Banks, 2009; Harrison, Lin, Carrol, & Carley, 2007). In this simulation, the mathematical model was that of the common SMD expressed by Equations (3) through (6), and the method investigated was the representation of the difference between the means of populations P_0 and P_1 by the common SMD under (a) differing levels of measurement error in the IV and (b) different prevalence rates.

In the simulation, the absence of measurement error in the DV was assumed, and the “true” common SMD, defined as its value when there was no measurement error in either DV or IV, was +.50, a value equal to the mean common SMD found in the analysis of over 300 meta-analyses by Lipsey and Wilson (1993). The subpopulations P_1^t and P_0^t in population $P_0^t \cup P_1^t$ were assumed to have means, respectively, of $\mu_{Y_\alpha}^{P_1^t} = 50$ and $\mu_{Y_\alpha}^{P_0^t} = 45$, and equal variances, $\sigma^2(Y_\alpha)_{P_0^t} = \sigma^2(Y_\alpha)_{P_1^t} = 100$. This latter assumption was made since it is one typically made in meta-analysis (Borenstein, 2009).

Figure 3 shows the error in the common SMD, with “error” defined as the difference between the common SMD affected by simulated measurement error in the IV and its “true” value of +.50, plotted as a function of measurement error in the IV for two prevalence rates: .50 and .09. The prevalence of .50 simulated experimentally

created subpopulations of persons, one of which received a treatment and one that did not (solid curves). The prevalence of .09 simulated a low prevalence context, such as the 1-year prevalence of MDD in the United States (dashed curves). The error is represented on the vertical axis; measurement error in the IV as indicated by sensitivity is scaled on the horizontal axis, and as indicated by specificity, with values ranging from 1.0 to 0.40, marked for the curves. For example, the top solid curve shows the error as a function of sensitivity given the prevalence was 0.50 and $\text{spec}(X_{\psi}) = 1.0$.

As this graph shows, the error in the common SMD increases as measurement error in the IV increases; as either the sensitivity or specificity, or both, decrease from 1.0, the error increases. One way of assessing the practical significance of the errors is by comparing them with the *SD*, 0.29, of the distribution of mean common SMDs from Lipsey and Wilson (1993). For example, the error in the common SMD, given a prevalence of .50 and holding the sensitivity constant at 1.0, increased from 0 to $-.09$ as the specificity decreased from 1.0 to 0.80, a magnitude about .3 *SD* in the Lipsey and Wilson distribution. In contrast, given a prevalence of .09, the error in the SMD increased from 0 to about $-.34$, about 1.2 *SD* in the Lipsey and Wilson distribution, as the specificity decreased from 1.0 to 0.80 while holding the sensitivity constant at 1.0. Given a prevalence of .50, the error in the common SMD increased from 0 to $-.21$ as *both* sensitivity and specificity decreased from 1.0 to 0.80, an error covering .70 *SD* in the Lipsey and Wilson distribution. Given a prevalence of .09, the error in the SMD increased from 0 to $-.37$, about 1.3 *SD* in the Lipsey and Wilson distribution, as sensitivity and specificity both decreased from 1.0 to 0.80.

The errors in the common SMD in Figure 3 indicate attenuation, results consistent Hunter and Schmidt's (2004) observations. These results also suggest the effects of measurement error in the IV on the common SMD are moderated by prevalence. A graph of errors in the numerator of the common SMD, similar to Figure 3 and omitted here in the interest of brevity, shows substantial attenuation in the numerator of the SMD as sensitivity and specificity decrease from 1.0. For example, given a prevalence of .09, the difference in the numerator decreased from 5.0 to 1.3 as both sensitivity and specificity decreased to .80. These results suggest the errors in the common SMD in Figure 3 are due prominently to the effects of measurement error in the IV on the numerator of the SMD.

Relative Effects of Measurement Error in IV and DV

The dash-dot-dot curve in Figure 3 shows error in the common SMD as a function of measurement error in the DV, assuming no measurement error in the IV. For this curve the horizontal axis is scaled as the reliability coefficient for scores from the DV in $P_0 \cup P_1$. This curve allows comparison of attenuation in the common SMD introduced by measurement error in the IV with that due to measurement error in the DV. This comparison suggests attenuation caused by measurement error in the IV can be substantially larger than that due to measurement error in the DV, particularly when the prevalence is low.

A Proposed Method for Correcting the Common and Alternate Versions of the SMD for Measurement Error in the IV

Theoretical Rationale

Define the term af as $af = (ppv(X_\psi) + npv(X_\psi) - 1)$. If Equation (6) is divided by af , the numerator becomes $\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P'_0}$; the effects of measurement error in the IV on the numerator are removed, thereby partially reliability correcting the common SMD for the effects of measurement error in the IV. If Equation (7b) is divided by af , it becomes

$$\delta(Y_\alpha, X_\psi)_{alternate} = \frac{(\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P'_0})}{\sqrt{\sigma^2(Y_\alpha)_{P'_0 \cup P'_1}}} = \frac{\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P'_0}}{\sigma(Y_\alpha)_{P'_0 \cup P'_1}}; \quad (9)$$

the alternate version of the SMD is completely reliability corrected for the effects of measurement error in the IV.

A Proposed Method for Reliability Correcting the SMD

The foregoing suggests the following method for disattenuating the numerator of the SMD, thereby partially reliability correcting the common SMD, and completely reliability correcting the alternate SMD, for the effects of measurement error in the IV.

Step 1. Obtain sample estimates of the means and variances of the DV in populations P_0 , P_1 , and $P_0 \cup P_1$, and of $p(\psi)_{P_0}$ and $p(\psi)_{P_1}$. Use these to obtain a sample estimate of the common SMD, $\hat{\delta}(Y_\alpha, X_\psi)_{common}$, or of the alternate SMD, $\hat{\delta}(Y_\alpha, X_\psi)_{alternate}$, depending on which is to be used. Also obtain values of the prevalence of the characteristic of interest in population $P'_0 \cup P'_1$, and of $sens(X_\psi)$ and $spec(X_\psi)$. Use these to compute $ppv(X_\psi)$, $npv(X_\psi)$, and af .

Step 2. An estimate of the numerator disattenuated, partially reliability corrected common SMD, symbolized as $\hat{\delta}^{PR}(Y_\alpha, X_\psi)_{common}$, can then be obtained from

$$\hat{\delta}^{PR}(Y_\alpha, X_\psi)_{common} = \frac{\hat{\delta}(Y_\alpha, X_\psi)_{common}}{af}. \quad (10)$$

An estimate of the reliability corrected alternate SMD, symbolized as $\hat{\delta}^R(Y_\alpha, X_\psi)_{alternate}$, can be obtained from

$$\hat{\delta}^R(Y_\alpha, X_\psi)_{alternate} = \frac{\hat{\delta}(Y_\alpha, X_\psi)_{alternate}}{af}. \quad (11)$$

Reliability Correcting the Unstandardized Mean Difference

Lipsey and Wilson (2001) defined the unstandardized mean difference (UMD) as

$$\mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0};$$

the UMD is the numerator of the population SMD. It follows as a corollary of the foregoing that the UMD can be disattenuated for the effects of measurement error in the IV from

$$\frac{\hat{\mu}_{Y_{\alpha}}^{P_1} - \hat{\mu}_{Y_{\alpha}}^{P_0}}{af}. \quad (12)$$

Conceptual Interpretation of af

Equations (6) and (7) imply the relationship between the difference $\mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0}$ and the difference $\mu_{Y_{\alpha}}^{P'_1} - \mu_{Y_{\alpha}}^{P'_0}$ is given by

$$\mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0} = (ppv(X_{\psi})_{SPi} + npv(X_{\psi})_{SPj} - 1) (\mu_{Y_{\alpha}}^{P'_1} - \mu_{Y_{\alpha}}^{P'_0}) = af (\mu_{Y_{\alpha}}^{P'_1} - \mu_{Y_{\alpha}}^{P'_0}). \quad (13)$$

Thus, af is an attenuation factor quantifying the extent to which the numerator of the population SMD, either common or alternate, is attenuated due to the effects of measurement error in the IV. It also quantifies the attenuation in the UMD due to measurement error in the IV.

The values of af range from -1 to $+1$. If $af = -1$, which occurs when $ppv(X_{\psi}) = npv(X_{\psi}) = 0$, it indicates all persons in P'_1 are misclassified into P_0 , and all persons in P'_0 are erroneously classified into P_1 . Thus, the numerator of the SMD becomes $\mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0} = -1 (\mu_{Y_{\alpha}}^{P'_1} - \mu_{Y_{\alpha}}^{P'_0}) = \mu_{Y_{\alpha}}^{P'_0} - \mu_{Y_{\alpha}}^{P'_1}$. Dividing this by $af = -1$ corrects it for measurement error in the IV: $\frac{\mu_{Y_{\alpha}}^{P'_0} - \mu_{Y_{\alpha}}^{P'_1}}{-1} = \mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0}$. When $af = 1$, which occurs when $ppv(X_{\psi}) = npv(X_{\psi}) = 1$, it indicates there are no classification errors due to random measurement error. When $af = 0$, it indicates the difference in the numerator of the SMD is 0, so the SMD is 0, and therefore no number exists that can divide the SMD to correct it for measurement error in the IV.

Return to the Illustrative Example

Consider again the illustrative example from the introduction. The estimated common SMD was .187, and the Hunter and Schmidt method produced a reliability-corrected common SMD of .387. The alternate SMD would be, in this case, .185, nearly the same as the common SMD. In this example, $af = (.4149 + .9698 - 1) \cong .3849$, so the partially reliability corrected common SMD using the proposed method would be $\frac{.187}{.3849} \cong .49$, and the reliability corrected alternate version of the population SMD

Table 1. Parameters (Rounded to Two Decimal Places) of Simulated Populations in the Eight Monte Carlo Studies.

Simulation	$\mu_{Y_{\alpha}}^{P_0^t} (\sigma_{P_0^t})$	$\mu_{Y_{\alpha}}^{P_1^t} (\sigma_{P_1^t})$	$\mu_{Y_{\alpha}}^{P_0^t \cup P_1^t} (\sigma_{P_0^t \cup P_1^t})$	Prevalence	$r = \frac{\sigma^2(Y_{\alpha})_{P_1^t}}{\sigma^2(Y_{\alpha})_{P_0^t}}$
1	53.98 (10.45)	59.18 (10.45)	56.58 (10.77)	.50	1.0
2	53.98 (10.45)	64.45 (10.45)	59.18 (11.68)	.50	1.0
3	53.98 (10.45)	59.21 (10.45)	54.43 (10.56)	.09	1.0
4	53.98 (10.45)	64.43 (10.45)	54.89 (10.87)	.09	1.0
5	55.04 (5.23)	59.18 (10.45)	57.11 (8.52)	.50	4.0
6	55.04 (5.23)	63.31 (10.45)	59.18 (9.24)	.50	4.0
7	53.18 (5.23)	56.11 (10.45)	54.43 (5.93)	.09	4.0
8	53.18 (5.23)	59.18 (10.45)	53.70 (6.11)	.09	4.0

Note. $\mu_{Y_{\alpha}}^{P_0^t}$ = mean score on DV Y_{α} in population P_0^t ; $\sigma_{P_0^t}$ = SD of scores Y_{α} in population P_0^t ; $\mu_{Y_{\alpha}}^{P_1^t}$ = mean score on DV Y_{α} in population P_1^t ; and so on.

would be $\frac{.185}{.3849} \cong .481$. These values differ from those resulting from the Hunter and Schmidt method, differences considered further below.

A Series of Monte Carlo Simulations

A series of Monte Carlo studies of the two reliability correction methods were conducted (Axelrod, 2007; Banks, 2009; Mooney, 1997). The objectives of these simulations were to (a) obtain Monte Carlo estimates of the sampling distributions of the numerator disattenuated, partially reliability corrected common, and reliability corrected alternate, SMDs obtained using both the proposed method and the Hunter and Schmidt (2004) approach; (b) compare these two methods in terms of bias, efficiency, and the ranges of estimates; and (c) investigate the extent to which disattenuating the numerator of the common SMD effectively reliability corrects it for the effects of measurement error in the IV. Bias was defined as the difference between the mean of the sampling distribution of the estimated reliability corrected SMD and the true value of the measurement error free population SMD. Efficiency was represented in terms of mean squared error (*MSE*; Taboga, 2012).

Methodology

Figure 1 helps in describing the methodology of these simulations. First, DV scores for populations P_0^t , P_1^t , and $P_0^t \cup P_1^t$ were simulated; the means and standard deviations (in parentheses) of these simulated populations are in Table 1, respectively, from left to right. The scores were normally distributed with equal variances in P_0^t and P_1^t , as commonly assumed in meta-analytic models (e.g., Hedges & Olkin, 1985). The scores were assumed to come from a Likert-type scale with a range of scores of about 100 (e.g., the Generalized Contentment Scale [GCS]; Hudson, 1982). The random normal number generator in SPSS version 21 was used to generate the

populations of scores, with a variance of about 109 in simulated P'_0 , and in four simulations in simulated P'_1 , a value found in research with the GCS (e.g., Hudson, 1982; Hudson & Proctor, 1977; Poage, Ketzenberger, & Olson, 2004). Also in Table 1 are prevalence rates in simulated populations $P'_0 \cup P'_1$ of the characteristic of interest, and r , the ratio of the variance of DV scores in population P'_1 to that in population P'_0 , $r = \sigma^2(Y_\alpha)_{P'_1} / \sigma^2(Y_\alpha)_{P'_0}$. These population characteristics are considered further below.

The classification of persons from P'_0 and P'_1 into populations P_0 and P_1 was then simulated, a process simultaneously modeling population $P_0 \cup P_1$. In all simulations it was assumed that $sens(X_\psi) = .55$ and $spec(X_\psi) = .75$. These sensitivity and specificity values, in line with reported sensitivity and specificity values for interview methods used to classify persons as having or not having MDD (e.g., Swedish Council on Health Technology Assessment, 2012), were used to infuse substantial measurement error into the simulated classification of persons from populations P'_0 and P'_1 into P_0 and P_1 . The purpose of infusing this degree of measurement error was to investigate the extent to which the two reliability correction methods removed the effects of this measurement error on the common and alternate SMDs. Consistent with the assumption $sens(X_\psi) = .55$, 45% of persons in population P'_1 were randomly selected to be erroneously classified into observed P_0 . Similarly, consistent with the assumption $spec(X_\psi) = .75$, 25% of persons in population P'_0 were randomly selected to be erroneously classified into observed P_1 .

To investigate their possible effects on the reliability correction methods, three factors were varied in the simulations: prevalence of the characteristic of interest in $P'_0 \cup P'_1$; magnitude of the measurement error free population SMD; and ratio of the variances of scores on the DV in populations P'_1 and in P'_0 , $r = \frac{\sigma^2(Y_\alpha)_{P'_1}}{\sigma^2(Y_\alpha)_{P'_0}}$. The prevalence rates, .09 and .50, from the model-based simulation, the results of which are in Figure 3, were simulated in order to test the hypothesis, suggested by results of the simulation, that prevalence may influence the reliability correction methods. Two measurement error free population common SMD magnitudes were simulated, .50 and 1.0, in order to explore the possibility the reliability correction methods work differently for different magnitude SMDs. The value of .50 was used given it was the mean common SMD found in the analysis of meta-analyses by Lipsey and Wilson (1993), while the SMD of 1.0 was used to simulate a “large” effect size (Cohen, Cohen, West, & Aiken, 2003).

Finally, two values of the variance ratio, r , were simulated: 1.0, consistent with equal variances in populations P'_0 and P'_1 ; and 4.0, simulating a large difference between the variances in P'_0 and P'_1 . The ratio r was varied to investigate the possibility the reliability correction methods performed differently in equal variance and unequal variance contexts. The eight Monte Carlo simulations had the following prevalence, variance ratio, and true common SMD values:

- Simulation (1): prevalence = .50, $r = 1$, true common SMD = .50
- Simulation (2): prevalence = .50, $r = 1$, true common SMD = 1.0
- Simulation (3): prevalence = .09, $r = 1$, true common SMD = .50

- Simulation (4): prevalence = .09, $r = 1$, true common SMD = 1.0
- Simulation (5): prevalence = .50, $r = 4$, true common SMD = .50
- Simulation (6): prevalence = .50, $r = 4$, true common SMD = 1.0
- Simulation (7): prevalence = .09, $r = 4$, true common SMD = .50
- Simulation (8): prevalence = .09, $r = 4$, and true common SMD = 1.0

Once populations P_0 , P_1 , and $P_0 \cup P_1$ were simulated, 6,000 random samples of $n = 300$ cases were obtained from $P_0 \cup P_1$ in each simulation. This modeled a study in which a large sample of persons from $P_0 \cup P_1$ was obtained to investigate the relationship between the characteristic that differentiates membership in P_1^t from that in P_0^t and the DV. For each random sample the sample means $\bar{Y}_{\alpha}^{P_0}$, $\bar{Y}_{\alpha}^{P_1}$, and $\bar{Y}_{\alpha}^{P_0 \cup P_1}$; sample SDs $s(Y_{\alpha})_{P_0}$, $s(Y_{\alpha})_{P_1}$, and $s(Y_{\alpha})_{P_0 \cup P_1}$; and sample sizes were used to estimate the common SMD, using formulas from Lipsey and Wilson (2001), and the alternate SMD from

$$\hat{\delta}(Y_{\alpha}, X_{\psi})_{alternate} = \frac{\bar{Y}_{\alpha}^{P_1} - \bar{Y}_{\alpha}^{P_0}}{s(Y_{\alpha})_{P_0 \cup P_1}}.$$

The reliability corrected common and alternate SMDs were estimated for each random sample using the methods described earlier, giving 6,000 estimates in each simulation. In those simulations in which the prevalence was .50, $af = .3125$ and the square root of the reliability coefficient for classification (Phi correlation between observed classification and true population membership) was .306. In those simulations in which the prevalence was .09, $af = .1227$ and the square root of the reliability coefficient for classification was .189.

Results

The results of the Monte Carlo simulations are shown in Table 2. The first column identifies the simulation number (1 to 8); the SMD being reliability corrected ($\delta_{alternate}$ = alternate version; δ_{common} = common version); and the reliability correction method (PM = proposed method; HS = Hunter and Schmidt, 2004, method). Then, shown from left to right are the following:

1. The means of the sampling distributions, with the differences between means of sampling distributions and true measurement error free SMDs (bias) in parentheses
2. The SDs of the sampling distributions
3. The ranges of estimates of the reliability corrected SMDs
4. 99.9% confidence intervals (CIs) for bias in the estimates. The CIs for reliability corrected SMDs obtained using the proposed method were normal curve based, while those from the Hunter and Schmidt method were bootstrap CIs given these sampling distributions were nonnormally distributed. A 99.9% CI that included 0 was taken to indicate an unbiased estimate, and vice versa. Use of 99.9% CIs gave an overall type I error rate for bias inferences of less than .05 over the 32 CIs

Table 2. Results of Eight Monte Carlo Simulations.

Simulation		Mean of sampling distribution	SD	Range of estimates	Approx. 99.9% CI for bias	MSE	Z
1: $\delta_{\text{alternate}}$	PM	+ .47 (−.01)	.37	−1.0, 1.8	−.023, .008	.14	0.76
	HS	+ .53 (+.05)	.45	−1.2, 3.8	.034, .064	.20	2.59
δ_{common}	PM	+ .48 (−.02)	.38	−1.0, 1.9	−.039, −.009	.14	0.98
	HS	+ .53 (+.03)	.46	−1.2, 4.6	.016, .048	.21	3.01
2: $\delta_{\text{alternate}}$	PM	+ .88 (−.01)	.38	−.46, 2.1	−.029, .002	.14	0.73
	HS	+ 1.1 (+.21)	.83	−.48, 31.7	.152, .248	.73	9.50
δ_{common}	PM	+ .89 (−.11)	.39	−.46, 2.2	−.129, −.097	.16	0.67
	HS	+ 1.1 (+.10)	.70	−.48, 12.5	.072, .129	.50	6.88
3: $\delta_{\text{alternate}}$	PM	+ .50 (0.0)	1.1	−3.4, 4.2	−.046, .042	1.2	0.68
	HS	+ .41 (−.09)	1.3	−5.4, 70.2	−.132, −.048	1.8	11.4
δ_{common}	PM	+ .50 (0.0)	1.1	−3.4, 4.3	−.044, .044	1.2	0.58
	HS	+ .41 (−.09)	1.1	−5.9, 24.2	−.125, −.055	1.2	8.57
4: $\delta_{\text{alternate}}$	PM	+ .99 (+.03)	1.1	−2.7, 5.5	−.015, .075	1.2	0.55
	HS	+ .84 (−.12)	2.4	−2.5, 153.2	−.203, −.037	5.6	18.1
δ_{common}	PM	+ 1.0 (0.0)	1.1	−2.7, 5.8	−.050, .040	1.2	0.84
	HS	+ .82 (−.18)	1.5	−2.5, 70.6	−.234, −.126	2.2	11.6
5: $\delta_{\text{alternate}}$	PM	+ .48 (−.01)	.38	−1.0, 2.1	−.027, .004	.15	0.68
	HS	+ .54 (+.05)	.55	−1.2, 23.4	.024, .076	.31	5.77
δ_{common}	PM	+ .48 (−.02)	.39	−1.1, 2.1	−.033, −.017	.15	0.53
	HS	+ .54 (+.04)	.48	−1.3, 5.5	.008, .072	.23	3.49
6: $\delta_{\text{alternate}}$	PM	+ .89 (0.0)	.38	−.64, 2.2	−.007, .011	.15	0.51
	HS	+ 1.1 (+.21)	.77	−.68, 27.5	.172, .248	.65	8.30
δ_{common}	PM	+ .90 (−.10)	.40	−.64, 2.4	−.106, −.086	.17	0.68
	HS	+ 1.1 (+.10)	.95	−.69, 46.3	.065, .135	.93	10.6
7: $\delta_{\text{alternate}}$	PM	+ .46 (−.04)	1.2	−3.6, 4.8	−.088, .005	1.3	0.87
	HS	+ .37 (−.13)	1.1	−13.7, 23.8	−.168, −.092	1.2	7.90
δ_{common}	PM	+ .46 (−.04)	1.2	−3.7, 5.0	−.086, .008	1.4	1.0
	HS	+ .38 (−.12)	1.2	−5.7, 26.3	−.168, −.072	1.4	8.60
8: $\delta_{\text{alternate}}$	PM	+ .98 (0.0)	1.2	−3.3, 5.4	−.045, .047	1.3	0.68
	HS	+ .82 (−.16)	1.4	−4.7, 43.1	−.205, −.115	2.1	11.0
δ_{common}	PM	+ .99 (−.01)	1.2	−3.3, 5.6	−.060, .033	1.4	0.87
	HS	+ .86 (−.14)	2.2	−5.1, 87.0	−.289, .005	4.8	16.2

Note. PM = proposed reliability correction method; HS = Hunter and Schmidt reliability correction method; $\delta_{\text{alternate}}$ = alternate SMD; δ_{common} = common SMD; SD = standard deviation of sampling distribution; this is also standard error of mean; Z = Kolmogorov–Smirnov Z statistic.

- 5. *MSE* values
- 6. Kolmogorov–Smirnov *Z*-statistics for tests of normality of the sampling distributions

Results Comparing Hunter and Schmidt and Proposed Methods

The proposed reliability correction method had lower bias values, controlling for version of SMD (alternate or common), prevalence, variance ratio (*r*), and magnitude of

the measurement error free population SMD, than the Hunter and Schmidt approach, and the differences in bias were moderated by prevalence, $F(1, 25) = 64.6, p < .001$, with the moderating relationship uniquely accounting for about 48% of the variation in bias.¹ Given a prevalence of .50, the mean bias in estimated reliability corrected common SMDs using the Hunter and Schmidt (2004) method was about, .07 (95% CI: .02, .1); and given a prevalence of .09, about $-.13$ (95% CI: $-.18, -.09$). Given a prevalence of .50, the mean bias using the Hunter and Schmidt method to reliability correct the alternate version of the SMD was about .13 (95% CI: .08, .18); and given a prevalence of .09, about $-.13$ (99% CI: $-.17, -.08$).

In contrast, given a prevalence of .50, the mean bias in estimates of the reliability corrected common SMD using the proposed method to disattenuate the numerator was $-.06$ (95% CI: $-.11, -.02$); and given a prevalence of .09, $-.01$ (95% CI: $-.06, .04$). Given a prevalence of .50, the mean bias in estimated reliability corrected alternate SMDs using the proposed method was $-.008$ (95% CI: $-.06, .04$); and given a prevalence of .09, $-.003$ (95% CI: $-.05, .05$).

The proposed method was overall more efficient than the Hunter and Schmidt (2004) method in terms of *MSE*. Controlling for version of the SMD, prevalence, variance ratio, and magnitude of the measurement error free SMD, the difference between the *MSE* for Hunter and Schmidt (2004) estimated reliability corrected SMDs and that for estimates from the proposed method was .79, $F(1, 26) = 6.93, p < .05$ (95% bootstrap CI for difference: .11 to 1.5).² The *MSE* of estimates of the reliability corrected SMDs was also strongly associated with prevalence, controlling for the other factors in the simulations, $F(1, 26) = 28.2, p < .001$, unique $R^2 = .425$. The difference between the mean *MSE* associated with estimated reliability corrected SMDs for a prevalence of .09 and that for a prevalence of .50, controlling for the factors in the simulation, was about 1.6 (95% bootstrap CI for difference: .9 to 2.3). Estimates in the higher prevalence context were overall more efficient. There was also evidence suggesting the proposed method produced more efficient estimates in the .09 prevalence context, mean *MSE* = 1.28, than the Hunter and Schmidt method, mean *MSE* = 2.54 (95% bootstrap CI for difference: .42 to 2.1).

The proposed method also produced, overall, estimates with narrower ranges of estimates, and lower extreme values, regardless of whether the reliability correction was for the common or alternate SMD. The mean range of estimates for the proposed method was from -2.0 to about 3.4; and for the Hunter and Schmidt method, -3.3 to 40.8.

Summary of Results

The proposed method appeared superior to the Hunter and Schmidt approach in terms of producing unbiased estimates of both common and alternate SMDs disattenuated for the effects of measurement error in the IV in the .09 prevalence context. The proposed method produced estimates of the reliability corrected common SMD with a downward mean bias of about .06 given a prevalence of .50. The proposed approach

produced overall more efficient estimates, in terms of *MSE*, of reliability corrected SMDs than the Hunter and Schmidt method.

The Illustrative Example: Conclusion

The illustrative example, considered earlier at two points in this article, comes from a simulation in which the population parameters were $\mu_{Y_{\alpha}}^{P_1} - \mu_{Y_{\alpha}}^{P_0} = 5.23$ and $\sigma(Y_{\alpha})_{P_1} = \sigma(Y_{\alpha})_{P_0} = 10.45$, so the measurement error free population common SMD was $+.50$. The *SD* in simulated $P_0 \cup P_1$ was 10.56 , so the measurement error free population alternate SMD was $+.495$. As seen earlier, the estimated reliability corrected common SMD using the Hunter and Schmidt method was $+.387$, and using the proposed method $+.49$. The estimated reliability corrected alternate SMD was $+.481$. The results of the Monte Carlo simulations explain the differences between estimates from the two reliability correction methods in this example. The simulations in the Hunter and Schmidt approach, given a prevalence of $.09$, produced overall downwardly biased estimates, whereas the proposed method produced unbiased estimates. Thus, in the example the Hunter and Schmidt reliability corrected estimate of $+.387$ is downwardly biased.

Conclusion

The results of the simulation in Figure 2 suggest attenuation in the SMD due to measurement error in the IV depends on prevalence, can be particularly pronounced when prevalence is low, and can exceed that due to measurement error in the DV. The results also suggest significant attenuation can occur at levels of measurement error, as indicated by sensitivity and specificity values, found in scores from measures currently used. In the illustrative example, the sensitivity and specificity values were $.717$ and $.90$, respectively, mean values from a recent review of interview methods used to identify persons with MDD. The measurement error implied by these values, in context of the MDD prevalence in the United States of about $.09$, led to an attenuation factor of $.385$; the magnitude of the SMD numerator, common or alternate, would be slightly more than one-third its value were there no measurement error in the IV. These findings imply significant attenuation due to measurement error in the IV may exist in SMDs reported in research, especially in low prevalence population comparison studies.

Given that levels of measurement error in the IV may vary across studies, the degree of attenuation in SMDs will vary across studies. As noted earlier, this differential attenuation will propagate through meta-analyses, a problem likely compounded by differential measurement error in DVs. The propagation of differential attenuation of SMDs through meta-analyses can potentially lead to erroneous results and conclusions. The proposed reliability correction method appears a promising approach for disattenuating the SMD, common or alternate, due to measurement error in the IV, thereby increasing the validity of results from meta-analyses.

The results of the Monte Carlo simulations support use of the proposed method of correcting the alternate version of the population SMD for the effects of measurement error in the IV, regardless of prevalence. The results support its use for reliability correcting the common SMD, especially in lower prevalence contexts, by disattenuating the numerator of the SMD for the effects of measurement error in the IV. The proposed method appears most promising for disattenuating SMDs from population comparison studies. The proposed reliability correction method could be implemented in a meta-analysis in a manner analogous to that suggested by Hedges and Olkin (1985) for meta-analyzing SMDs corrected for measurement error in the DV. The weighted reliability corrected estimate of the SMD would be given by formula (40); confidence intervals by formulas (41) and (42); and a test of homogeneity of SMDs corrected for measurement error in the IV using formula (43) in Lipsey and Wilson (2001), but with the term af , substituted for the term $\sqrt{\rho(Y, Y')}$ (the square root of the reliability coefficient for scores on the DV) in each of these formulas. This approach to meta-analyzing alternate or common SMDs that have been corrected for measurement error in the IV using the proposed method needs study in subsequent research.

As currently formulated the proposed method is based on the assumption persons in P'_0 have the same probability of being misclassified into P_1 , and persons in P'_1 have the same probability of being misclassified into P_0 . This assumption is unlikely to hold for some measurement procedures used for classification. One example would be a Likert-type scale, such as the GCS mentioned earlier, that produces a range of scores and classification decisions are made based on a cut score. The GCS has a cutting score of 30, and persons with scores of less than 30 can be classified as "not depressed," whereas persons with scores of 30 or higher are classified as "depressed" (Hudson, 1982). As a truly nondepressed person's score is increasingly lower than 30, the probability he or she will be misclassified as "depressed" goes down, and the probability he or she will be accurately classified as "not depressed" increases. Similarly, as a truly depressed person's score is increasingly greater than 30, the probability he or she will be misclassified as "not depressed" decreases, and the probability he or she will be accurately classified as "depressed" will increase (Pepe, 2003). Thus, all truly nondepressed persons will not have the same probability of being misclassified as "depressed;" and all truly depressed persons will not have the same probability of being misclassified as "nondepressed." The probability of misclassification will increase as persons' scores get closer to the cutting score.

The proposed reliability correction method needs further theoretical development to be applicable in measurement scenarios such as that immediately above. One approach to generalizing the reliability correction method developed above might be to derive expressions for the sensitivity and specificity conditional on values of the observed scores from the measure of the IV. These might be based, for example, on a receiver operating characteristic curve for the relationship between the scores on the IV measure and classification (Pepe, 2003; Swets, 1988). The expressions for these conditional sensitivity and specificity values could then be used to derive expressions

for the conditional PPV and conditional NPV. From these a disattenuation factor similar to af might be derived.

The proposed reliability correction method, like the method sketched by Hunter and Schmidt (2004), can only be used to correct for the effects of random measurement error in the IV. It will not correct for systematic error. Hunter and Schmidt (2004) discussed systematic measurement error under the conceptual umbrella "imperfect construct validity." In this exposition, Hunter and Schmidt considered three approaches to dealing with systematic error in the IV. The reader is referred to this source for in depth consideration of this issue in meta-analysis.

The graph of common and alternate versions of the SMD in Figure 2 suggests the difference between the two versions will be .05 or less for common SMD values of about .75 or lower. This implies the alternate version of the SMD might be used in circumstances in which the common SMD would be .75 or less, and Lipsey and Wilson's (1993) findings suggest this may occur rather frequently, with relatively minimal differences between the two versions of the SMD. Earlier it was argued a principal advantage of use of the alternate SMD is the insensitivity of the denominator to the effects of measurement error in the IV, and the ability to correct this version of the SMD for the effects of measurement error in the IV by dividing it by af . These considerations suggest the alternate SMD could be used fairly frequently in lieu of the common version, and the advantages gained by its use would come at relatively small cost in terms of difference in magnitude between these versions of the SMD, a speculation that needs testing in subsequent research.

Recent work has been done on the development of regression coefficient based EFSs for use in meta-analysis (e.g., Kim, 2011). An interesting line of future research and theoretical development is investigation of the extension of the proposed reliability correction method to correcting regression-based EFSs for the effects of measurement error in the IV. Keef and Roberts' (2004) recently proposed "partial standardized mean difference" appears to be an interesting regression based EFS on which to focus. The partial SMD is, essentially, a SMD that has been adjusted for a covariate. Generalization of the proposed reliability correction method to this particular regression based EFS might open the door to application of the method to other regression based EFSs, such as those developed by Kim (2011). It might also provide a link between reliability correcting meta-analytic EFSs and correcting regression coefficients for the effects of measurement error in the IV.

Finally, the results of the Monte Carlo simulations have implications for future research. Monte Carlo simulations investigating the use of the proposed reliability correction method with both the common and alternate SMDs need to be done with prevalence values different from those in the current simulations. Prevalence rates lower than .09; between .09 and .50; and above .50 need to be done. The bias and efficiency of the two reliability correction methods needs to be investigated as a function of these prevalence rates. The results of the current simulations suggest the hypothesis that the proposed method used to reliability correct the alternate version of the SMD will produce unbiased estimates regardless of prevalence. A specific

research question concerns the prevalence at which disattenuation of the numerator of the common SMD using the proposed reliability correction method ceases to give unbiased estimates of the common SMD corrected for the effects of measurement error in the IV. The results of the current simulations imply this point will be in the neighborhood of .50. The results of the Monte Carlo simulations also suggest the Hunter and Schmidt method might produce unbiased estimates at some prevalence rates between .09 and .50. Factors other than those studied in the current Monte Carlo studies need to be varied in future simulation studies of the proposed method, such as sample size.

Appendix

Numerator of the Common and Alternate SMD

Consider Figure 1. The subpopulation P_0 mean observed score on the DV will be

$$\mu_{Y_\alpha}^{P_0} = \frac{N_{P'_0:P_0} \mu_{Y_\alpha}^{P'_0:P_0} + N_{P'_1:P_0} \mu_{Y_\alpha}^{P'_1:P_0}}{N_{P'_0:P_0} + N_{P'_1:P_0}} = \frac{N_{P'_0:P_0}}{N_{P'_0:P_0} + N_{P'_1:P_0}} \mu_{Y_\alpha}^{P'_0:P_0} + \frac{N_{P'_1:P_0}}{N_{P'_0:P_0} + N_{P'_1:P_0}} \mu_{Y_\alpha}^{P'_1:P_0} \quad (A1)$$

and that for subpopulation P_1 will be

$$\mu_{Y_\alpha}^{P_1} = \frac{N_{P'_1:P_1} \mu_{Y_\alpha}^{P'_1:P_1} + N_{P'_0:P_1} \mu_{Y_\alpha}^{P'_0:P_1}}{N_{P'_1:P_1} + N_{P'_0:P_1}} = \frac{N_{P'_1:P_1}}{N_{P'_1:P_1} + N_{P'_0:P_1}} \mu_{Y_\alpha}^{P'_1:P_1} + \frac{N_{P'_0:P_1}}{N_{P'_1:P_1} + N_{P'_0:P_1}} \mu_{Y_\alpha}^{P'_0:P_1}, \quad (A2)$$

where $N_{P'_0:P_0}$ is the number of persons in subpopulation $P'_0 : P_0$, $N_{P'_1:P_0}$ the number in subpopulation $P'_1 : P_0$, $N_{P'_1:P_1}$ the number in $P'_1 : P_1$, and $N_{P'_0:P_1}$ the number in $P'_0 : P_1$ (Pepe, 2003). But $npv(X_\psi) = \frac{N_{P'_0:P_0}}{N_{P'_0:P_0} + N_{P'_1:P_0}}$ and $1 - npv(X_\psi) = \frac{N_{P'_1:P_0}}{N_{P'_0:P_0} + N_{P'_1:P_0}}$; and $ppv(X_\psi) = \frac{N_{P'_1:P_1}}{N_{P'_1:P_1} + N_{P'_0:P_1}}$ and $1 - ppv(X_\psi) = \frac{N_{P'_0:P_1}}{N_{P'_1:P_1} + N_{P'_0:P_1}}$ (Pepe, 2003); and the assumption that misclassification is due only to random measurement error implies, $\mu_{Y_\alpha}^{P'_1:P_1} = \mu_{Y_\alpha}^{P'_1:P_0} = \mu_{Y_\alpha}^{P'_1}$ and $\mu_{Y_\alpha}^{P'_0:P_0} = \mu_{Y_\alpha}^{P'_0:P_1} = \mu_{Y_\alpha}^{P'_0}$. Thus, $\mu_{Y_\alpha}^{P_0} = (1 - npv(X_\psi)) \mu_{Y_\alpha}^{P'_1:P_0} + npv(X_\psi) \mu_{Y_\alpha}^{P'_0:P_0}$ and $\mu_{Y_\alpha}^{P_1} = ppv(X_\psi) \mu_{Y_\alpha}^{P'_1:P_1} + (1 - ppv(X_\psi)) \mu_{Y_\alpha}^{P'_0:P_1}$, so

$$\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0} = \left[ppv(X_\psi) \mu_{Y_\alpha}^{P'_1} + (1 - ppv(X_\psi)) \mu_{Y_\alpha}^{P'_0} \right] - \left[(1 - npv(X_\psi)) \mu_{Y_\alpha}^{P'_1} + npv(X_\psi) \mu_{Y_\alpha}^{P'_0} \right].$$

After expanding the right hand side, collecting and rearranging terms, and recognizing,

$$(ppv(X_\psi) + npv(X_\psi) - 1) = - (1 - ppv(X_\psi) - npv(X_\psi)),$$

the numerator of the SMD can be written as

$$\mu_{Y_\alpha}^{P_1} - \mu_{Y_\alpha}^{P_0} = (ppv(X_\psi) + npv(X_\psi) - 1) \left(\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P'_0} \right). \quad (A3)$$

Denominator of Common SMD

Consider again Figure 1. The total sums of squares for observed scores Y_α on the DV in observed P_0 will be given by Equation (A4) (Kirk, 1994),

$$\begin{aligned} SST_{P_0} = & \sum_{P'_0:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_0:P_0} \right)^2 + \sum_{P'_1:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_1:P_0} \right)^2 + N_{P'_0:P_0} \left(\mu_{Y_\alpha}^{P'_0:P_0} - \mu_{Y_\alpha}^{P_0} \right)^2 \\ & + N_{P'_1:P_0} \left(\mu_{Y_\alpha}^{P'_1:P_0} - \mu_{Y_\alpha}^{P_0} \right)^2 \end{aligned} \quad (A4)$$

so the variance of scores $\sigma^2(Y_\alpha)_{P_0}$ will be given by Equation (A5),

$$\begin{aligned} \sigma^2(Y_\alpha)_{P_0} = & \frac{\sum_{P'_0:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_0:P_0} \right)^2}{N_{P'_0:P_0} + N_{P'_1:P_0}} + \frac{\sum_{P'_1:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_1:P_0} \right)^2}{N_{P'_0:P_0} + N_{P'_1:P_0}} \\ & + \frac{N_{P'_0:P_0} \left(\mu_{Y_\alpha}^{P'_0:P_0} - \mu_{Y_\alpha}^{P_0} \right)^2}{N_{P'_0:P_0} + N_{P'_1:P_0}} + \frac{N_{P'_1:P_0} \left(\mu_{Y_\alpha}^{P'_1:P_0} - \mu_{Y_\alpha}^{P_0} \right)^2}{N_{P'_0:P_0} + N_{P'_1:P_0}}. \end{aligned} \quad (A5)$$

Now the variances $\sigma^2(Y_\alpha)_{P'_0:P_0}$ and $\sigma^2(Y_\alpha)_{P'_1:P_0}$ are

$$\sigma^2(Y_\alpha)_{P'_0:P_0} = \frac{\sum_{P'_0:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_0:P_0} \right)^2}{N_{P'_0:P_0}} \text{ and } \sigma^2(Y_\alpha)_{P'_1:P_0} = \frac{\sum_{P'_1:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_1:P_0} \right)^2}{N_{P'_1:P_0}}, \quad (A6)$$

so

$$\sigma^2(Y_\alpha)_{P'_0:P_0} = \frac{\sum_{P'_0:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_0:P_0} \right)^2}{N_{P'_0:P_0}} = \frac{\sum_{P'_0:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_0:P_0} \right)^2}{npv(X_\psi) \left(N_{P'_0:P_0} + N_{P'_1:P_0} \right)}, \quad (A7)$$

and

$$\sigma^2(Y_\alpha)_{P'_1:P_0} = \frac{\sum_{P'_1:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_1:P_0} \right)^2}{N_{P'_1:P_0}} = \frac{\sum_{P'_1:P_0} \left(Y_\alpha - \mu_{Y_\alpha}^{P'_1:P_0} \right)^2}{(1 - npv(X_\psi)) \left(N_{P'_0:P_0} + N_{P'_1:P_0} \right)}. \quad (A8)$$

Thus, the variance of scores $\sigma^2(Y_\alpha)_{P_0}$ will be given by

$$\begin{aligned}\sigma^2(Y_\alpha)_{P_0} = & npv(X_\psi)\sigma^2(Y_\alpha)_{P'_0:P_0} + (1 - npv(X_\psi))\sigma^2(Y_\alpha)_{P'_1:P_0} + npv(X_\psi)\left(\mu_{Y_\alpha}^{P'_0:P_0} - \mu_{Y_\alpha}^{P_0}\right)^2 \\ & + (1 - npv(X_\psi))\left(\mu_{Y_\alpha}^{P'_1:P_0} - \mu_{Y_\alpha}^{P_0}\right)^2.\end{aligned}\quad (A9)$$

By a parallel argument, the variance of scores $\sigma^2(Y_\alpha)_{P_1}$ will be

$$\begin{aligned}\sigma^2(Y_\alpha)_{P_1} = & ppv(X_\psi)\sigma^2(Y_\alpha)_{P'_1:P_1} + (1 - ppv(X_\psi))\sigma^2(Y_\alpha)_{P'_0:P_1} + ppv(X_\psi)\left(\mu_{Y_\alpha}^{P'_1:P_1} - \mu_{Y_\alpha}^{P_1}\right)^2 \\ & + (1 - ppv(X_\psi))\left(\mu_{Y_\alpha}^{P'_0:P_1} - \mu_{Y_\alpha}^{P_1}\right)^2.\end{aligned}\quad (A10)$$

Now as noted in the text, the assumption that misclassification occurs only due to random measurement error implies $\mu_{Y_\alpha}^{P'_1:P_1} = \mu_{Y_\alpha}^{P'_1:P_0} = \mu_{Y_\alpha}^{P'_1}$ and $\mu_{Y_\alpha}^{P'_0:P_0} = \mu_{Y_\alpha}^{P'_0:P_1} = \mu_{Y_\alpha}^{P'_0}$, and $\sigma^2(Y_\alpha)_{P'_1:P_1} = \sigma^2(Y_\alpha)_{P'_1:P_0} = \sigma^2(Y_\alpha)_{P'_1}$ and $\sigma^2(Y_\alpha)_{P'_0:P_0} = \sigma^2(Y_\alpha)_{P'_0:P_1} = \sigma^2(Y_\alpha)_{P'_0}$. Thus,

$$\begin{aligned}\sigma^2(Y_\alpha)_{P_0} = & npv(X_\psi)\sigma^2(Y_\alpha)_{P'_0} + (1 - npv(X_\psi))\sigma^2(Y_\alpha)_{P'_1} + npv(X_\psi)\left(\mu_{Y_\alpha}^{P'_0} - \mu_{Y_\alpha}^{P_0}\right)^2 \\ & + (1 - npv(X_\psi))\left(\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P_0}\right)^2\end{aligned}\quad (A11)$$

and

$$\begin{aligned}\sigma^2(Y_\alpha)_{P_1} = & ppv(X_\psi)\sigma^2(Y_\alpha)_{P'_1} + (1 - ppv(X_\psi))\sigma^2(Y_\alpha)_{P'_0} + ppv(X_\psi)\left(\mu_{Y_\alpha}^{P'_1} - \mu_{Y_\alpha}^{P_1}\right)^2 \\ & + (1 - ppv(X_\psi))\left(\mu_{Y_\alpha}^{P'_0} - \mu_{Y_\alpha}^{P_1}\right)^2.\end{aligned}\quad (A12)$$

Therefore, the denominator of the common SMD will be

$$\sqrt{p(\psi)_{P_0}\sigma^2(Y_\alpha)_{P_0} + p(\psi)_{P_1}\sigma^2(Y_\alpha)_{P_1}},$$

with $\sigma^2(Y_\alpha)_{P_0}$ given by Equation (A11) and $\sigma^2(Y_\alpha)_{P_1}$ by Equation (A12).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Results from OLS regression. Results of analyses of residuals consistent with assumptions of normality and homogeneity of variance of residuals, hence confidence intervals are normal curve based.
2. Bootstrap confidence intervals used since residuals analyses of OLS results were consistent with violation of the homogeneity of variance, and possibly violation of normality, of residuals assumptions (Cohen et al., 2003).

References

- Axelrod, R. (2007). Simulation in the social sciences. In J. Rennard (Ed.), *Nature inspired computing for economics and management* (pp. 90-100). Hershey, PA: Idea Group Reference.
- Banks, C. (2009). What is modeling and simulation? In J. Sokolowski & C. Banks (Eds.), *Principles of modeling and simulation: A multidisciplinary approach* (pp. 3-24). Hoboken, NJ: Wiley.
- Berk, R. (1980). *Criterion-referenced measurement: State of the art*. Baltimore, MD: John's Hopkins University Press.
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221-236). New York, NY: Russell Sage Foundation.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. New York, NY: Wiley.
- Brennan, R. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Divgi, D. (1980). Group dependence of some reliability indices for mastery tests. *Applied Psychological Measurement*, 4, 213-218. doi:10.1177/014662168000400208
- Grissom, R., & Kim, J. (2011). *Effect sizes for research* (2nd ed.). New York, NY: Routledge.
- Haertel, E. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education and Praeger.
- Harrison, J., Lin, Z., Carrol, G., & Carley, K. (2007). Simulation modeling in organizational and management research. *Academy of Management Review*, 32, 1229-1245.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.
- Hudson, W. (1982). *The clinical measurement package*. Homewood, IL: Dorsey.
- Hudson, W., & Proctor, E. (1977). Assessment of depressive affect in clinical practice. *Journal of Consulting and Clinical Psychology*, 45, 1206-1207.
- Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.
- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4, 105-126.
- Kazdin, A. (2002). *Anxiety and its disorders: The nature and treatment of anxiety and panic* (2nd ed.). New York, NY: Guilford.
- Keef, S. P., & Roberts, L. A. (2004). The meta-analysis of partial effect sizes. *British Journal of Mathematical and Statistical Psychology*, 57, 97-129.

- Kim, R. S. (2011, June 30). Standardized regression coefficients as indices of effect sizes in meta-analysis (Paper 3109). *Electronic Theses, Treatises and Dissertations*. Retrieved from <http://diginole.lib.fsu.edu/cgi/viewcontent.cgi?article=2989&context=etd>
- Kirk, R. (1994). *Experimental design: Procedures for behavioral sciences* (3rd ed.). Independence, KY: Wadsworth.
- Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209. doi: 10.1037/0003-066X.48.12.1181
- Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Newbury Park, CA: Sage.
- Mooney, C. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage.
- Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Orwin, R., & Cordray, D. (1985). Effects of deficient reporting on meta-analysis. *Journal of Applied Psychology*, 97, 134-147.
- Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford.
- Poage, E., Ketzenberger, K., & Olsen, J. (2004). Spirituality, contentment, and stress in recovering alcoholics. *Addictive Behaviors*, 29, 1857-1862.
- Rosenthal, H. (1994). Parametric measures of effect size. In H. Cooper & L. Hedges, (Eds.), *Handbook of research synthesis* (pp. 231-244). New York, NY: Russell Sage Foundation.
- Schmidt, H., Le, H., & Oh, I.-S. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis methods* (2nd ed., pp. 317-336). New York, NY: Russell Sage Foundation.
- Snyder, H. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive functions: A meta-analysis and review. *Psychological Bulletin*, 139(1), 81-132.
- Swedish Council on Health Technology Assessment. (2012). *Diagnostik och uppföljning av förstämningssyndrom: En systematisk litteraturöversikt* [Diagnosis and monitoring of mood disorders: A systematic literature review]. Retrieved from http://www.sbu.se/upload/Publikationer/Content0/1/Forstamningssyndrom_fulltext.pdf
- Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.
- Taboga, M. (2012). *Lectures on probability theory and mathematical statistics* (2nd ed.). Lyndhurst, NJ: Barnes & Noble.
- Thompson, S. (2002). *Sampling* (2nd ed.). New York, NY: Wiley.