Meta-Analysis of Correlation Coefficients

Running head: META-ANALYSIS OF CORRELATION COEFFICIENTS

Meta-Analysis of Correlation Coefficients: A Cautionary Tale on Treating Measurement Error

Qian Zhang

Florida State University

Correspondence should be addressed to Qian Zhang, 3204F Stone Building, Department of Educational Psychology and Learning Systems, College of Education, Florida State University, 1114 W. Call Street, Tallahassee, FL 32306. E-mail: qzhang4@fsu.edu. I thank Dr. Betsy Becker for her helpful comments on an earlier version of this article. The study has not been previously disseminated.

Abstract

A scale to measure a psychological construct is subject to measurement error. When meta-analyzing correlations obtained from scale scores, many researchers recommend correcting measurement error. We considered three caveats when conducting meta-analysis of correlations: (1) the distribution of true scores can be non-normal, resulting in a violation of the normality assumption for raw correlations and Fisher's z transformed correlations; (2) coefficient alpha is often used as the reliability, but correlations corrected for measurement error using alpha can be inaccurate when some assumptions (e.g., tau-equivalence) of alpha are violated; and (3) item scores are often ordinal, making the disattenuation formula potentially problematic. Via three simulation studies, we examined the performance of two meta-analysis approaches with raw correlations and z scores. In terms of estimation accuracy and coverage probability of the mean correlation, results showed that (1) the true score distribution alone had slight influence; (2) when the tau-equivalence assumption was violated and coefficient alpha was used for correcting measurement error, the mean correlation estimate can be biased and coverage probability can be low; and (3) discretization of continuous items can result in under-coverage of the mean correlation even when tau-equivalence was satisfied. With more categories and/or items on a scale, results can improve when tau-equivalence was met or not. Based on these findings, we then gave recommendations when conducting meta-analysis of correlations. KEY WORDS: meta-analysis of correlations; measurement error; coefficient alpha; ordinal scale

Meta-Analysis of Correlation Coefficients: A Cautionary Tale on Treating Measurement Error

Meta-analysis is a statistical technique to combine information from different studies (Glass, MacGaw, & Smith, 1984). Specifically, it can be used to provide an overall estimate of population effect size, and examine whether and how effect sizes across studies differ. Effect sizes that are commonly used in meta-analysis include Pearson product moment correlation, standardized mean difference, odds ratio, and so forth. In this study, we focus on meta-analyzing Pearson product moment correlation (in the remainder of the article, we use "correlation" for simplicity). It goes without doubt that correlations are very commonly used in social science studies to depict bi-variate linear relationships. Meta-analysis of correlations are also widely used and discussed in various fields. Examples in recent years include meta-analysis of correlations between sleep and work (Litwiller, Snyder, Taylor, & Steele, 2017), emotional intelligence and job performance (Joseph, Jin, Newman, & O'Boyle, 2015), boldness and psychopathic personality (Lilienfeld et al., 2016), openness and academic achievement (Gatzka & Hell, 2018), and so forth.

Meta-analysis of correlations involves two constructs X and Y. In psychological and educational studies, often times X and Y are measured by scales consisting of multiple items. Scales for measuring X and Y are subject to measurement error (Nunnally, 1978). Based on classical test theory, the variance of observed scores is a sum of the true score variance and the measurement error variance. The ratio of the true score variance to the observed score variance is the reliability, which can be used to gauge to what extent scales measure the true scores. We consider the scenario that primary studies provide reliability estimates of the scales and focus on measurement error as an artifact when conducting meta-analysis.[1]

Denote reliabilities of the scales for X and Y as rel_X and rel_Y respectively. It is well established that an observed nonzero correlation coefficient is attenuated such that $r_{XY} = r_{X_T Y_T} \times \sqrt{rel_X rel_Y}$, where r_{XY} is the observed correlation and $r_{X_T Y_T}$ is the underlying/true correlation (Spearman, 1904). To obtain the correlation corrected for unreliability, theoretically we can use

$$r_{X_T Y_T} = \frac{r_{XY}}{\sqrt{rel_X rel_Y}}. \tag{1}$$

Equation 1 is referred to as correction for attenuation due to measurement error. When conducting meta-analysis of correlations, there have been debates on whether to correct measurement error. Researchers who oppose correcting measurement error argue that the disattenuated effect size would be obtained under the conditions of perfect measurement and therefore consider the corrected effect size as "hypothetical" (e.g., LeBreton, Scherer, & Lawrence, 2014; Nunnally, 1978). Muchinsky (1996) argued that the corrected correlation (in his term, validity) coefficients are biased in an "ideal world" while adding that when the research interest is about a particular sample only with no desire of generalization, corrected correlation coefficients are upward biased. This may imply that for the purpose of generalization, which often is the case, correction of measurement error seems reasonable.

Researchers advocating for correcting measurement error point out usefulness of correction when focusing on reliable variance of the measures (e.g., Hunter & Schmidt, 2004; Oswald, Ercan, McAbee, Ock & Shaw, 2015). Moreover, primary studies may use different scales on the same construct, resulting in "study effects" when meta-analyzing correlations. Using observed correlations obtained from the scale scores may make them not directly comparable across studies. Therefore, Ahn (2008) recommended obtaining the pure effect size with accounting for the study effects. Even with the same scale, reliability estimates are likely to be different across primary studies. An important field, reliability generalization, is centered on obtaining a typical reliability, variability of reliabilities, and sources of variability across studies (Vacha-Haase, 1998). Thus, not correcting measurement error may result in misleadingly larger variability of effect sizes when they are actually homogeneous, or larger/smaller variability than the truth with heterogeneous effect sizes across studies. Focusing on meta-analysis of correlations while accounting for potentially varied reliabilities across studies, we mainly study approaches to correcting measurement error.

To correct measurement error of observed correlations based on Equation 1, researchers need to have r_{XY} that is usually obtained using scale scores of X and Y and reliability estimates for the scales of X and Y for each study. As Oswald et al. (2015) correctly argued, however, psychometric corrections may not be entirely appropriate. As far as we are concerned, there are several caveats when using the disattenuation formula for correcting measurement error. These caveats may have individual and/or combined influence

on estimation of the mean effect size, which is the focal parameter in this study.

Caveat 1: The true scores of X and Y may be non-normal.

In many of the previous methodological studies about meta-analysis of correlations, the true scores of X and Y are assumed to be normally distributed (e.g., Field, 2001, 2005; Hafdahl, 2010; Hall & Brannick, 2002). In reality, the true scores of X and Y may be normal (e.g., intelligence) or non-normal (e.g., depression). Lambert (1996) examined the influence of non-normality and dependency between effect sizes on between-studies variance estimate and found that data distribution did influence the variance estimate and therefore validity generalization result. Estimation and statistical inference of the mean effect size may also be influenced by the distribution of the true scores as the commonly used interval estimates and tests of the mean effect size are often based on the normality assumption (Sánchez-Meca & Marín-Martínez, 2008). Furthermore, in practice, researchers of meta-analysis can not obtain correlations based on true scores directly. With the use of measurement scales of X and Y, the underlying distribution of true scores can interact with other factors to have influence on the results of meta-analysis of correlations. In caveats 2 and 3, we will discuss other factors that can also impact estimation of the mean correlation by themselves and/or interacting with the distribution of true scores.

Caveat 2: The assumption of linear association between the observed and true scores may be violated, which can influence estimates of the individual corrected correlations.

In many cases, ordinal items are used for a scale. For example, educational or cognitive assessments are often scored as binary responses (correct/incorrect); personality scales such as PANAS (Watson, Clark, & Tellegen, 1988) consist of two 10-item scales to measure both positive and negative affects, with each item rated on a 5-point scale of 1 (not at all) to 5 (very much). Assuming items on a scale measure a single construct (i.e., unidimensionality),[2] a continuous item is often times regarded as linearly associated with the corresponding latent variable such that $Item_j = \lambda_j T + e_j$ ($j = 1, ...J$), where $Item_j$ is the score for item j on the scale of X or Y, λ_j is the coefficient of the true score T, or the factor loading under the confirmatory factor analysis framework (Bollen, 1989), e_j is the measurement error for item j, and J is the number of items on the scale. Typically, it is assumed that T is independent of e_j. Then, the scale score,

such as the sum of item scores, is also linearly associated with the true scores:

$$\sum_{j=1}^{J} Item_j = T\sum_{j=1}^{J}\lambda_j + \sum_{j=1}^{J} e_j. \quad (2)$$

The sum of measurement errors across items $\sum_{j=1}^{J} e_j$ is also independent of T. When T and $\sum_{j=1}^{J} Item_j$ (the scale score) are linearly associated and $\sum_{j=1}^{J} e_j$ is independent of T, the correction for attenuation formula presented by Spearman (1904) holds.

Nevertheless, when items are ordinal, it is not legitimate to assume that the association between an item and the corresponding latent variable is linear. Instead, as is shown by ample relevant research in the latent variable modeling framework (e.g., Finney & DiStefano 2013; Flora & Curran 2004), a nonlinear structural equation model should be used for depicting the relationship between each ordinal item and the latent variable. Specifically, an underlying unobserved continuous variable $Item_j^*$ together with $C-1$ thresholds produce an observed ordinal item $Item_j$ with C categories. For example, when a score of $Item_j^*$ is between τ_{c-1} and τ_c, $Item_j = c$. In this case, it is legitimate to assume that $Item_j^*$ instead of $Item_j$ is linearly associated with T, and thus the scale score $\sum_{j=1}^{J} Item_j$ should not be linearly associated with T, the true latent score. The correlation obtained using the scale scores with ordinal items ($Item_j$) is usually smaller than that using the underlying continuous items ($Item_j^*$) due to the loss of information by categorizing $Item_j^*$.

When researchers are presented with a complete data set of ordinal item scores, approaches have been provided to estimate the correlation between two latent variables of X and Y (e.g., for an excellent comparison of the existing estimation methods, see Rhemtulla, Brosseau-Liard, & Savalei, 2012). For researchers conducting meta-analysis, however, they probably do not have access to the original data. In this latter case, if one wishes to correct for unreliability before synthesizing correlation coefficients, they often refer to the disattenuation formula (Equation 1) presented earlier, which technically does not hold for ordinal scales given the nonlinear relationship between the observed scale scores and the true scores. The nonlinear relationship between each ordinal item and the latent construct also contributes to the next caveat about the reliability estimate used for correcting measurement error.

Caveat 3: Primary studies often report coefficient alpha as the estimated reliability without checking its assumptions, which can make the corrected correlation inaccurate when some assumptions about alpha are violated.

Reliability of the scale based on Equation 2 is

$$\begin{aligned} rel &= \frac{(\sum_{j=1}^{J} \lambda_j)^2 \text{var}(T)}{(\sum_{j=1}^{J} \lambda_j)^2 \text{var}(T) + \text{var}(\sum_{j=1}^{J} e_j)} \\ &= \frac{(\sum_{j=1}^{J} \lambda_j)^2 \text{var}(T)}{(\sum_{j=1}^{J} \lambda_j)^2 \text{var}(T) + \sum_{j=1}^{J} \text{var}(e_j)}, \end{aligned} \tag{3}$$

where var$(\cdot)$ denotes variance. Equation 3 is also coefficient omega (McDonald, 1999).

It is worth emphasizing the assumptions about the scale for appropriately obtaining coefficient omega: (1) items on the scale measure a single latent construct of interest (uni-dimensionality); (2) the latent construct is the only common cause of the inter-item correlation, and thus error scores e_j are independent across items; and (3) as stated earlier, item scores are continuous and are linearly related with the latent construct. When these assumptions are met for the scales of X and Y, we can replace rel_X and rel_Y with coefficient omega estimated under the confirmatory factor analysis framework (e.g., using R package `MBESS`) in the correction for attenuation formula.

However, coefficient alpha (Cronbach, 1951) instead is still the most often reported measure of reliability in social science studies. coefficient alpha is

$$\alpha = \frac{J^2 \bar{\sigma}_{jj'}}{\sigma^2}, \tag{4}$$

where $\bar{\sigma}_{jj'}$ is the average covariance between item j and $j^{'}$ ($j \neq j^{'}$), and σ^2 is the variance of the scale scores. One main reason why coefficient alpha is so popular is probably because its calculation is built in almost all statistical software (e.g., SAS, SPSS, Stata, R, etc.).

To examine the current practice of meta-analyzing correlations, we did a literature review in five major psychological journals between 2009 and 2019: the *Journal of Personality and Social Psychology*, the *Journal of Applied Psychology*, the *Journal of Consulting and Clinical Psychology*, *Health Psychology,*

and *Developmental Psychology*. In total 70 studies about meta-analysis of correlations were found. When synthesizing individual correlations, 35 (50%) studies did not correct measurement error, and 9 (12.9%) studies claimed that they corrected for measurement error but were not clear about how the correction was done. Among the other studies that clearly stated reliability measures, more than 2/3 used coefficient alpha for correcting measurement error.

Despite the popularity of coefficient alpha, many studies have commented on its limitations as a measure of reliability (e.g., Graham, 2006; Green & Yang, 2009; Sijtsma, 2009). These limitations stem from the potential violation of assumptions when alpha is used as a reliability estimate. The assumptions for coefficient omega also apply for alpha. For example, common to both coefficient omega and alpha, with ordinal scales, items cannot be assumed to be linearly related to the latent construct. Despite this, alpha is still frequently reported as the reliability estimate with ordinal scales. Using it for correcting measurement error is problematic due to the nonlinear relationship between items and the latent variable. In this case, the reported alpha coefficient using $Item_j$ ($j = 1, ...J$) is typically smaller than the reliability estimate using $Item_j^*$.[3]

Furthermore, coefficient alpha also requires that all items have the same linear relationship with the latent construct, or tau-equivalence. This means that all λ_j's are the same in Equation 2. By replacing λ_j with λ (a constant) in Equation 3, coefficient omega reduces to coefficient alpha:

$$rel = \frac{J^2\lambda^2 var(T)}{\sigma^2} = \frac{J^2\bar{\sigma}_{jj'}}{\sigma^2}.$$

As demonstrated in Yang (2018), alpha is the lower bound of coefficient omega, provided that errors are uncorrelated. Though appearing simpler, the tau-equivalence assumption is restrictive and likely to be violated in practice. Thus, often times reliability is underestimated with continuous items. With ordinal scales, reliability is also underestimated with the violation of tau-equivalence assumption. Furthermore, variances and covariances of categorical item scores depend on the thresholds for discretization as well as the underlying distribution of continuous item scores. For example, even when the variances and covariances of the underlying continuous item scores are the same under normal and non-normal distributions (thus alpha is the same), discretizing continuous items may change relative frequencies of

item scores and/or ranks of participants in different ways with underlying normal and non-normal items. Therefore, variances and covariances of categorical items and thus coefficient alpha are likely to differ as well under different underlying item distributions. In turn, individual correlations corrected by alpha and the meta-analyzed correlation would also be influenced.

For scales that are appropriately designed to measure a single construct, ideally the assumptions of unidimensionality and uncorrelated errors are met. Individual correlation coefficient corrected for attenuation may be influenced by non-normality of true scores, the nonlinear relationship between an ordinal item and the latent construct, and violation of tau-equivalence when coefficient alpha is used as the reliability estimate. Consequently the meta-analysis results should also be influenced. Therefore, in the study, we aim to consider each of these caveats and their combined influence on meta-analysis of correlations.

In the next section, we introduce two approaches to meta-analysis of correlations with correcting measurement error. A series of simulation studies are then conducted to evaluate performance of the two approaches. We end the article with discussion of the findings, recommendations, limitations, and future directions.

Meta-Analysis of Correlations with the Correction of Measurement Error

When synthesizing correlations across a total of NS studies, two methods are commonly used. One method uses raw correlations directly (Hunter & Schmidt, 2004), and the other method uses Fisher's z transformed scores (L. V. Hedges & Olkin, 1985; L. V. Hedges & Vevea, 1998). Two types of models can be used for meta-analyzing correlations using either the raw or z transformed correlations: fixed-effects models and random-effects models. Fixed-effects models assume the same true effects across studies, whereas random-effects models assume that the true effect sizes differ across studies or a nonzero between-studies variance. When the between-studies heterogeneity exists, ignoring it and fitting a fixed-effects model can lead to inflated type I error rate of testing the mean effect size (e.g., Field, 2003; Hunter & Schmidt, 2000). On the other hand, studies have shown that fitting a random-effects model when effects are the same can still yield accurate mean effect estimates and inferences (e.g., Baird & Maxwell, 2016; Ke, Zhang, & Tong, 2019; Zhang, Wang, & Bergeman, 2018). In fact, the fixed-effects model is a

special case of the random-effects model when the between-studies variance is zero. We focus on the mean effect estimates and therefore adopt random-effects models throughout the study.

When using random-effects models, the between-studies variance needs to be estimated. Many estimators of the between-studies variance have been proposed (e.g., the DerSimonian and Laird estimator, the Hedges estimator, the Hunter and Schmidt estimator, and the maximum likelihood estimator; DerSimonian & Laird 1986; Hunter & Schmidt 2004; L. V. Hedges 1983; Viechtbauer 2005). We choose restricted maximum likelihood (REML) estimator, for it yields unbiased variance estimates when the number of studies is large enough (Viechtbauer, 2005). Furthermore, REML is the default estimator in some popular statistical software for meta-analysis (e.g., the package `'metafor'` in R; Viechtbauer, 2010). In each iteration when finding the REML estimate of the between-studies variance, a negative estimate is set equal to zero (Veroniki et al., 2015). With the between-studies variance estimate, the within-study sampling error variance estimate, and individual correlations across studies, the mean correlation can be estimated. Details are to be shown using the raw and z transformed correlations respectively.

Meta-analysis using raw correlations

When raw correlations are used in meta-analysis, denote the observed correlation from study i as r_{oi} ($i = 1, ...,$NS). A meta-analysis model at the level of observed correlations is (Hunter & Schmidt, 2000)

$$r_{oi} = \rho_{oi} + e_{oi}, \tag{5}$$

where ρ_{oi} is the attenuated population correlation; e_{oi} is the corresponding within-study sampling error for study i with a mean of zero and a variance of $\sigma^2_{ro,i}$. According to Hunter and Schmidt (2004), $\sigma^2_{ro,i}$ is

$$\sigma^2_{ro,i} = \frac{(1 - \rho^2_{oi})^2}{N_i - 1}, \tag{6}$$

where N_i is the sample size in study i. In estimating $\sigma^2_{ro,i}$, researchers recommended replacing ρ_{oi} with the simple (unweighted) mean correlation averaged across studies: $\bar{r}_o = \sum_{i=1}^{NS} r_{oi}/NS$ (Becker & Fahrbach, 1994; Furlow & Beretvas, 2005; Hafdahl, 2008). Thus, an estimated sampling error variance estimate

without correction for measurement error is

$$\hat{\sigma}^2_{ro,i} = \frac{(1-\bar{r}^2_o)^2}{N_i - 1}. \tag{7}$$

The correlation corrected for measurement error using Equation 1 is $r_{ci} = r_{oi}/\sqrt{rel_{Xi}rel_{Yi}}$,where rel_{Xi} and rel_{Yi} are reliabilities for the scales of X and Y in study i. Therefore, the meta-analysis model at the level of the corrected sample correlation r_{ci} is

$$r_{ci} = \rho_{ci} + e_{ci}, \tag{8}$$

where $\rho_{ci} = \rho_{oi}/\sqrt{rel_{Xi}rel_{Yi}}$ is the corrected population correlation and $e_{ci} = e_{oi}/\sqrt{rel_{Xi}rel_{Yi}}$ is the corrected within-study sampling error for study i. The within-study sampling error variance e_{ci} is

$$\sigma^2_{rc,i} = \frac{\sigma^2_{ro,i}}{rel_{Xi}rel_{Yi}}. \tag{9}$$

The estimated sampling error variance with the correction of measurement error is (Hunter & Schmidt, 2004)

$$\hat{\sigma}^2_{rc,i} = \frac{\hat{\sigma}^2_{ro,i}}{\hat{rel}_{Xi}\hat{rel}_{Yi}}, \tag{10}$$

where $\hat{rel}_{Xi}$ and $\hat{rel}_{Yi}$ are the estimated reliabilities of the scales for X and Y respectively in study i.

The between-studies model of the corrected population correlations is

$$\rho_{ci} = \bar{\rho}_c + u_{ci}, \tag{11}$$

where $\bar{\rho}_c$ is the mean corrected population correlation and is the parameter of interest in this study. u_{ci} is the deviation between ρ_{ci} and $\bar{\rho}_c$. Let τ^2_c be the between-studies variance of ρ_{ci}, or the variance of u_{ci}. When $\tau^2_c = 0$ or > 0, ρ_{ci} across studies are fixed (and equal to $\bar{\rho}_c$) or random across studies, respectively.

When raw correlations are used in meta-analysis, estimation of $\bar{\rho}_c$ takes two steps (Raudenbush, 2009). In the first step, the between-studies variance is estimated using REML. REML aims to obtain an

unbiased estimate of τ_c^2 by defining the likelihood on residuals. Specifically, it is assumed that the residuals e_{ci} and u_{ci} from Equations 8 and 11 follow normal distributions with means of zero and variances of $\sigma_{rc,i}^2$ and τ_c^2 respectively. $\sigma_{rc,i}^2$ is estimated by $\hat{\sigma}_{rc,i}^2$, and therefore τ_c^2 is the only parameter to be estimated under REML (see e.g., Raudenbush, 2009 for details). In practice, the normality assumption of residuals might be violated, for example when $\bar{\rho}_c$ is relatively large. In the study, we will examine the performance of meta-analysis of raw correlations using REML for estimating τ_c^2 when $\bar{\rho}_c$ ranges from null to large such that the normality assumption is potentially violated under certain conditions.

In the second step, the Weighted Least Squares (WLS) estimate [4] for $\bar{\rho}_c$ (denoted by $\hat{\bar{\rho}}_c$) is obtained (e.g., under the R package `metafor`; Viechtbauer, 2010):

$$\hat{\bar{\rho}}_c = \frac{\sum_{i=1}^{NS}(\hat{\sigma}_{rc,i}^2 + \hat{\tau}_c^2)^{-1} r_{ci}}{\sum_{i=1}^{NS}(\hat{\sigma}_{rc,i}^2 + \hat{\tau}_c^2)^{-1}}. \tag{12}$$

One commonly used standard error estimate of $\hat{\bar{\rho}}_c$ is (L. V. Hedges & Vevea, 1998)

$$\text{se}(\hat{\bar{\rho}}_c) = \sqrt{\frac{1}{\sum_{i=1}^{NS}(\hat{\sigma}_{rc,i}^2 + \hat{\tau}_c^2)^{-1}}}. \tag{13}$$

Therefore, a 95% confidence interval of $\bar{\rho}_c$ is: $\hat{\bar{\rho}}_c - 1.96\text{se}(\hat{\bar{\rho}}_c) \leq \bar{\rho}_c \leq \hat{\bar{\rho}}_c + 1.96\text{se}(\hat{\bar{\rho}}_c)$,which can be used for statistically inferring the mean effect size.[5]

We term meta-analysis using the raw correlations as an "r-based" approach.

Meta-analysis using Fisher's z transformed scores

Alternatively, each individually corrected correlation r_{ci} is transformed into a Fisher's z score using

$$z_{rc,i} = \frac{1}{2}\ln(\frac{1 + r_{ci}}{1 - r_{ci}}) = \tanh^{-1}(r_{ci}). \tag{14}$$

A meta-analysis model with Fisher's z transformed correlations is

$$\begin{aligned} z_{rc,i} &= z_{\rho c,i} + e_{zrc,i}, \\ z_{\rho c,i} &= \bar{z}_{\rho c} + u_{z\rho c,i}. \end{aligned} \tag{15}$$

With bi-variate normally distributed X and Y scores, $z_{rc,i}$ is approximately normally distributed with a mean of $z_{\rho c,i} = \tanh^{-1}(\rho_{ci})$, where ρ_{ci} is the population correlation coefficient, and a within-study sampling error variance $\sigma^2_{zc,i}$. When there is no measurement error in either the X or Y scales, the asymptotic within-study sampling error variance for the Fisher's z transformed correlation in study i is

$$\sigma^2_{z,i} = (\frac{1}{\sqrt{N_i - 3}})^2. \tag{16}$$

In comparison to the r-based approach, $\sigma^2_{z,i}$ enjoys the property that it is invariant over population effect sizes. When using the disattenuated raw correlations for obtaining the Fisher's z transformed correlations, however, asymptotically Equation 16 is no longer the sampling error variance. To the best of my knowledge, however, correcting sampling error variance with z transformed disattenuated correlations has not been thoroughly examined.[6] Therefore, in order to provide cautions when conducting meta-analysis using the Fisher's z transformed corrected raw correlations, we will first examine whether using $\sigma^2_{z,i}$ could be problematic; then we will discuss a proposed approach to correcting sampling error variance with the z transformed disattenuated correlations in the Discussion section. Using either Equation 16 or the proposed approach, we denote in general the estimated within-study error variance of $z_{rc,i}$ as $\hat{\sigma}^2_{zc,i}$.

Let τ^2_{zc} be the between-studies variance of $z_{\rho c,i}$, or the variance of $u_{z\rho c,i}$. When $\tau^2_{zc} = 0$ or > 0, $z_{\rho c,i}$ (or ρ_{ci}) are fixed or random across studies respectively. Similar to the r-based approach, the between-studies variance needs to be estimated in order to estimate the mean effect size. REML is used again with assuming residuals $e_{zrc,i}$ and $u_{z\rho c,i}$ are normal. With REML estimate of τ^2_{zc} (denoted by $\hat{\tau}^2_{zc}$), the estimated within-study error variance $\hat{\sigma}^2_{zc,i}$, and $z_{rc,i}$, the WLS estimate of $\bar{z}_{\rho c}$ (denoted by $\hat{\bar{z}}_{\rho c}$) is

$$\hat{\bar{z}}_{\rho c} = \frac{\sum_{i=1}^{NS}(\hat{\sigma}^2_{zc,i} + \hat{\tau}^2_{zc})^{-1} z_{rc,i}}{\sum_{i=1}^{NS}(\hat{\sigma}^2_{zc,i} + \hat{\tau}^2_{zc})^{-1}}. \tag{17}$$

A commonly used standard error of $\hat{\bar{z}}_{\rho c}$ is (L. V. Hedges & Vevea, 1998)

$$\text{se}(\hat{\bar{z}}_{\rho c}) = \sqrt{\frac{1}{\sum_{i=1}^{NS}(\hat{\sigma}^2_{zc,i} + \hat{\tau}^2_{zc})^{-1}}}. \tag{18}$$

Therefore, a 95% confidence interval of $\bar{z}_{\rho c}$ is $\hat{\bar{z}}_{\rho c} - 1.96\text{se}(\hat{\bar{z}}_{\rho c}) \leq \bar{z}_{\rho c} \leq \hat{\bar{z}}_{\rho c} + 1.96\text{se}(\hat{\bar{z}}_{\rho c})$. Denote the lower and upper limits as $\hat{\bar{z}}_{\rho cL}$ and $\hat{\bar{z}}_{\rho cU}$ respectively.

To transform the synthesized z score ($\hat{\bar{z}}_{\rho c}$) back to the raw correlation metric, first we have $r_{Tc} = (e^{2\hat{\bar{z}}_{\rho c}} - 1)/(e^{2\hat{\bar{z}}_{\rho c}} + 1) = \tanh(\hat{\bar{z}}_{\rho c})$, the inverse of Fisher's z transformation of $\hat{\bar{z}}_{\rho c}$. Similarly, we have r_{TcL}=tanh($\hat{\bar{z}}_{\rho cL}$) and r_{TcU}=tanh($\hat{\bar{z}}_{\rho cU}$), the inverse of Fisher's z transformation for the lower and upper limits of the 95% confidence interval of $\bar{z}_{\rho c}$. r_{Tc} and the confidence interval [r_{TcL}, r_{TcU}] have been found to yield a biased estimate and a relatively low coverage rate of the mean effect size, particularly when the population mean effect size and/or the between-studies variance are large (Hafdahl, 2010). Researchers have recommended alternative ways to obtain more accurate mean effect size and its confidence interval estimates when transforming back to the r metric (e.g., integral of z-to-r transformation, Taylor series methods; Hafdahl, 2009, 2010; Law, 1995). In the study, we use the mean effect size estimate developed based on Taylor series approximation of the tanh function (see Equations 12-14 and Appendix of Law, 1995 for technical details):

$$\hat{m}_{\rho c} = r_{Tc}[1 + (r_{Tc}^2 - 1)\hat{\tau}_{zc}^2]. \tag{19}$$

The 95% confidence interval of $\bar{\rho}_c$ is obtained as $[\hat{m}_{\rho cL},\ \hat{m}_{\rho cU}]$, where $\hat{m}_{\rho cL} = r_{TcL}[1 + (r_{TcL}^2 - 1)\hat{\tau}_{zc}^2]$ and $\hat{m}_{\rho cU} = r_{TcU}[1 + (r_{TcU}^2 - 1)\hat{\tau}_{zc}^2]$. We term meta-analysis using Fisher's z transformed scores as a "z-based" approach.

It is worth emphasizing that the two approaches assume that (1) X and Y scores are bi-variate normally distributed for each study; (2) the true scores and observed scores of X and Y are linearly associated so that the disattenuation formula is justifiable (Equation 1); and (3) the reliability estimate used for correcting measurement error is accurate. The estimated mean effect size, however, can be influenced with potential violation of the normality assumption for the true scores of X or Y, use of ordinal items so that the association of the true and observed scores may no longer be linear, and/or the use of coefficient alpha as the reliability estimate with potential violation of its assumptions. Based on our knowledge, no studies have examined performance of meta-analysis approaches using the raw and Fisher's z transformed correlations considering together these potential violations of assumptions.

Furthermore, one can also examine whether individual correlations across studies are homogeneous. A traditional test of homogeneity is to use Cochran's *Q*-test (L. V. Hedges & Olkin, 1985) based on either the raw or z transformed correlations. The Q statistic asymptotically follows a chi-square distribution with NS-1 degrees of freedom (L. V. Hedges & Olkin, 1985; L. V. Hedges & Vevea, 1998; Rosenthal, 1991). Q statistics have been criticized in general of being too conservative when the sample sizes across studies and/or the total number of studies NS are small, and conversely too liberal when the sample sizes across studies and/or the total number of studies NS are large (Higgins, Thompson, Deeks, & Altman, 2003). In a study comparing performance of different approaches to testing homogeneity among various effect sizes, however, Viechtbauer (2007) found that when Fisher's z transformed correlations are used, the Q statistic is not sensitive to the sample sizes across studies or the number of studies and could well control the type I error rates. Note that this finding is based on correlations with no measurement error, and no studies have been done regarding homogeneity test considering the above-mentioned caveats. To fill the research gap, we conduct a series of Monte Carlo simulation studies.

Simulation Studies

Three simulation studies were conducted in sequence to explore potential influence from the aforementioned caveats. For these simulation studies, we evaluated the two meta-analysis approaches in terms of estimation accuracy of the mean correlation estimate using raw bias, empirical coverage probability of the mean effect size (proportion of confidence intervals covering the true average effect size across replications), and homogeneity test result.

Simulation study 1: Investigating influence of the true score distribution only on mean correlation estimate and inference

In simulation study 1, we first examined whether the distribution of true scores for X and Y can influence mean effect size estimate and inference. Therefore, both normal and non-normal true scores underlying items of a scale were considered together with factors including effect size, number of studies, and average sample size. In this simulation study, the influence of measurement error or discretization of continuous items was not considered yet.

Distribution of the true scores. For generating true scores in study i, we used the following approach:

$$x_{ti} = \omega_{1i}, \text{ and}$$

$$y_{ti} = \rho_i \omega_{1i} + \sqrt{1 - \rho_i^2}\omega_{2i},$$

where x_{ti} and y_{ti} are the true scores of X and Y in study i, and ω_{1i} and ω_{2i} are independent standardized random variables. The true correlation between X and Y is therefore ρ_i. For generating normal true scores, ω_{1i} and ω_{2i} followed a standard normal distribution. For generating non-normal true scores, data were first generated from a chi-square distribution with one degree of freedom and then standardized to obtain ω_{1i} and ω_{2i} with a skewness of 2.8 and an excess kurtosis of 12.[7]

Effect size and between-studies variability. Based on Cohen (1988), four levels of average population correlation ($\bar{\rho}$) were included: 0, .1, .3, .5, ranging from zero to large effect sizes. When correlations were fixed in the population, generated sample correlations contain sampling variability only. In contrast, when correlations were random in the population, in addition to sampling variability, sample correlations also contained between-studies variability.

When correlations were heterogeneous, we generated a random super-population as a normal distribution of z scores with an average Fisher's z transformed correlation $\bar{z}_\rho$, and a standard deviation (τ_z). We fixed τ_z at .16, a frequently reported and relatively large between-studies variability of correlations based on a review in Field (2005). In obtaining the population $\bar{z}_\rho$, the Taylor series method discussed in Hafdahl (2010) and Law (1995) was again used. Specifically, at the population level, the Taylor series method of transforming from $\bar{z}_\rho$ to the mean correlation is $\bar{\rho} \doteq \rho_T[1 + (\rho_T^2 - 1)\tau_z^2]$, where $\rho_T = \tanh(\bar{z}_\rho)$. Given this relationship as well as population values of $\bar{\rho}$ and τ_z, we can find solutions for $\bar{z}_\rho$ numerically. In particular, when $\bar{\rho}$ = 0, .1, .3, .5, the corresponding $\bar{z}_\rho$= 0, .103, .317, and .562 respectively. Each generated individual score z_i was transformed to the raw correlation metric using ρ_i= $\tanh(z_i)$ as the population correlation for study i.

Number of studies. The number of studies NS had the following three values: 5, 20, and 50, from small to relatively large.

Average sample size across studies. The sample sizes in individual studies of a meta-analysis are usually different. Positively skewed distributions of sample sizes were generated because in the literature of meta-analyzing correlations there can be a few studies with large sample sizes. Therefore, the same as Hafdahl (2010), we set N_i ($i = 1, ...,$NS) to the nearest integer of $[(\chi^2_{3,i} - 3)/\sqrt{2 \times 3}](\bar{N}/2) + \bar{N}$, where $\chi^2_{3,i}$ is a random number generated from a chi-square distribution with three degrees of freedom. Generated N_i's are positively skewed with an average of $\bar{N}$ and a variance of $(\bar{N}/2)^2$. Four levels of $\bar{N}$ were considered: 30, 50, 100, and 200. These average sample sizes were selected to be close to the sizes in a single study for detecting large, medium, and small correlations with a power of .8 (Cohen, 1988), following the same reason as Field (2001) and Field (2005). Furthermore, in each replicated data set under a certain condition the same set of sample sizes were used. This is consistent with the frequentist inference in meta-analysis of correlations (see also e.g., Bonett, 2008; Hafdahl, 2010). In each condition, whenever the generated sample size was less than 3 (very rarely happened even given the smallest $\bar{N} = 30$), it was discarded until a total of NS sample sizes were generated.

We would like to emphasize that when correlations differ across studies in the population, the proportion of between-studies variance (the ratio of the between-studies variance to the sum of the between-studies variance and the actual within-study sampling error variance, which is similar to the I^2 index of heterogeneity; Higgins et al., 2003) under the r-based approach and the z-based approach can be different. Specifically, for the r-based approach, the actual within-study sampling error variance is dependent on both the population effect size and the sample size in a study (Hunter & Schmidt, 2004, Equation 3.4). Furthermore, when using the true scores of X and Y that are normally distributed, based on Equation 14 in Law (1995), the between-studies variance for the raw correlations is approximately .026, .025, 0.021, and .014 with population mean correlations of 0, .1, .3, and .5 respectively and $\tau_z = .16$; in contrast, for the z-based approach, the within-study sampling error variance is solely dependent on N_i, and the between-studies variance is $.16^2$. Therefore, it seems necessary to examine the typical proportion of between-studies variance for both the r-based and z-based approaches.

Based on normally distributed true scores of X and Y, given $\tau_z = .16$ and various levels of NS, $\bar{N}$ and $\bar{\rho}$, each cell in table 1 shows the mean percentage of the between-studies variance averaged across

studies and over 1000 replications considering variation in generating ρ_i. As can be seen from the table, as $\bar{N}$ increased, the proportion of between-studies variance increased for both approaches; moreover, the proportion tends to be slightly smaller under the z-based approach than for the r-based approach given a certain NS and $\bar{N}$.

—Insert table 1 here—

Number of items per scale and corresponding reliability. Levels of the number of items per scale included 4 and 12. This allows us to examine influence of the length of the scale on meta-analysis of correlations. Observed continuous item scores were generated using a two-factor confirmatory factor analysis model, each factor with J items. For each study i, the factor scores were the true scores x_{ti} and y_{ti}, and the correlation between the two factors was the true correlation ρ_i. We used a fixed factor loading of .6 and measurement error variance of .55 for each item of scales measuring X and Y (standardized factor loading is .63).[8] Therefore, based on Equation 3, the true reliability when $J = 4$ and 12 was .72 and .89 respectively.

For the r-based approach, the model in Equation 5 was used and the sampling error variance was estimated using Equation 7. For the z-based approach, z transformed uncorrected correlations replaced $z_{rc,i}$ in Equation 15 and the sampling error variance was obtained using Equation 16. All other steps were essentially the same as those described under the two approaches in the previous section except that uncorrected instead of corrected raw or z-transformed correlations were used. For examining influence of the true score distributions, the obtained mean correlation estimates and coverage probabilities were compared to $\bar{\rho}_o = \bar{\rho}\sqrt{rel_X rel_Y}$, where rel_X and rel_Y are the true reliabilities for X and Y and $\bar{\rho}$ is the nominal population mean effect size specified earlier.

The above-mentioned factors were cross-tabulated to generate 384 conditions in total. For each condition, 1000 data sets were generated. To evaluate estimation accuracy, we used raw bias which is defined as $(\sum_{k=1}^{1000} \bar{r}_{ok})/1000 - \bar{\rho}_o$, where $\bar{r}_{ok}$ is the estimated $\bar{\rho}_o$ in replication k. In addition, the 95% coverage probability (CP) of $\bar{\rho}_o$ was computed as the proportion of replications with estimated intervals covering the true value. Reasonable empirical CPs should be between .925 and .975 (Bradley, 1978).

Simulation study 1 results. Table 2 contains results about raw bias and CPs when the number of items per scale was 4. We selected the conditions when $\bar{\rho} = 0$ and .5 ($\bar{\rho}_o = 0$ and .445) for discussion. For the generated 1000 data sets in every condition, the convergence rates for REML under both the r-based and z-based approaches were greater than 99% and only the converged results were used for result summary. When correlations were homogeneous across studies, coverage rates of $\bar{\rho}$ were in general close to the acceptable range of [.925, .975] for both zero and nonzero effect sizes. A few exceptions occurred with small NS and large $\bar{N}$ or large NS and small $\bar{N}$ for the r-based approach. When correlations were heterogeneous across studies, coverage rates lower than .925 were mostly found when NS=5 for both the r-based and z-based approaches and under the r-based approach when the true scores were skewed with large NS and small $\bar{N}$.

Comparing the r-based and z-based approaches, the former had less accurate estimates of $\bar{\rho}$ and lower coverage rates when $\bar{\rho}$ was large, NS was large and $\bar{N}$ was small (e.g., $\bar{\rho} = .5$, NS=50, and $\bar{N} = 30$). The difference of bias and coverage rates between the two approaches decreased a bit when the number of items per scale increased to 12 (A table of results can be found in the supplemental materials). Patterns of results when $\bar{\rho} = .1$ and .3 were similar to the condition when $\bar{\rho} = 0$ and therefore were not presented.

—Insert table 2 here—

Simulation study 2: Investigating consequences of using coefficient alpha for correcting measurement error when the tau-equivalence assumption is violated

Following simulation study 1, in simulation study 2, we mainly examined the influence of tau-equivalence violation when coefficient alpha was used for correcting measurement error. All conditions from simulation study 1 were still used in simulation study 2. In addition, in parallel to the condition that factor loadings were the same, or tau-equivalence was satisfied, we also considered the situation that factor loadings were not the same, or the tau-equivalence assumption was violated. Specifically, for a scale measuring X or Y, half of the items had a loading of .3 and a measurement error variance of .91, and the other half had a loading of .9 and a measurement error variance of .19. This results in standardized factor loadings of .3 and .9 respectively, corresponding to relatively small and large sizes. The two sets of factor loadings and measurement error variances described in simulation studies 1 and 2 were selected such that

reliabilities remained the same. It is expected that when the tau-equivalence assumption is violated, a relatively short scale tends to result in relatively large bias of reliability estimated by coefficient alpha (Graham, 2006; Yang, 2018).

The number of conditions in study 2 increased to be 768. For each replicated data set, an observed correlation for X and Y was computed based on the scale scores of X and Y. A corrected correlation coefficient was obtained using Equation 1 with coefficient alpha as the reliability estimate for both scales of X and Y. In a replicated data set, for each given N_i, when the computed coefficient alpha was outside the range of (0, 1), standard deviations of the scale scores were zero, and/or the corrected correlation coefficient was outside the range of (-1, 1), the corresponding data set was discarded and a new one was generated until NS studies were reached.[9] Estimates and coverage probabilities were compared to the true nominal $\bar{\rho}$. For the generated 1000 data sets in every condition, the convergence rates for REML under both the r-based and z-based approaches were greater than 99% and only the converged results were used for result summary.

Estimation accuracy of $\bar{\rho}$ and coverage probability when the tau-equivalence assumption is met. When tau-equivalence was satisfied and the number of items per scale $J = 4$, under the homogeneous case for the r-based approach (table 3), bias was close to zero across all conditions; CPs were generally in the reasonable range. Under the heterogeneous case for the r-based approach, the conditions when $\bar{\rho}$ was large with skewed true scores or when NS=5 tended to yield coverage rates under 92.5%; when increasing the number of items per scale, coverage rates slightly improved under these situations (see supplemental materials).

For the z-based approach, bias was generally comparable to that under the r-based approach and was slightly greater when NS=5 and $\bar{N} = 30$ for large $\bar{\rho}$. In comparison to the r-based approach, CPs tended to be smaller than 92.5% under both the homogeneous and heterogeneous cases when NS=5. When $J = 12$, CPs mostly increased to be within the range of [92.5%, 97.5%] with $\bar{\rho} \leq .3$ under the homogeneous case; however, CPs were still mostly lower than 92.5% for the heterogeneous case (see supplemental materials).

Therefore, with the correction of measurement error using coefficient alpha, when the tau-equivalence assumption was met, the two approaches did not have substantial bias. CPs of the r-based

approach were relatively low when the number of items was small, $\bar{\rho}$ was large, true scores were skewed and NS was small; CPs of the z-based approach tended to be low when NS was small. The z-based approach using Equation 16 as the estimated within-study sampling variance did not have substantial impact on the estimates and inferences of the mean effect size provided large enough NS.

—Insert table 3 here—

—Insert table 4 here—

Estimation accuracy of $\bar{\rho}$ and coverage probability when the tau-equivalence assumption is not met. When tau-equivalence was not satisfied, both the r-based and z-based approaches deteriorated from the conditions with meeting the tau-equivalence assumption. Specifically, both approaches generally yielded positive bias when $\bar{\rho} \neq 0$. In terms of CPs, there are several common findings for the two approaches: (1) CPs were lower than those with meeting the tau-equivalence assumption; (2) as $\bar{\rho}$ increased, CPs tended to decrease; (3) as NS and $\bar{N}$ increased, CPs tended to decrease for medium to large $\bar{\rho}$. For example, CPs were close to zero when tau-equivalence was violated, effect sizes were homogeneous, $J = 4$, $\bar{\rho} = .5$, $\bar{N} = 200$, and NS=50 (table 5). Findings (2) and (3) were probably because $\bar{\rho}$ was biased when it was nonzero. Bias decreased and CPs increased when the number of items increased. CPs when $\bar{\rho}$ was medium to large and/or NS = 5, however, can still be below .925 in many conditions. Furthermore, for the r-based approach, under the homogeneous case bias tended to be larger and CPs tended to be lower when true scores were skewed than when they were normal with medium to large $\bar{\rho}$; this pattern is clearer when $J = 12$ than when $J = 4$. Detailed results about the r-based approach can be found in tables 5 and 6. Details about the z-based approach can be found in the supplemental materials.

—Insert table 5 here—

—Insert table 6 here—

Type I error rate of the homogeneity test. We focus on reporting the type I error rate instead of power for the homogeneity test of effect sizes (i.e., when they were fixed in the population). This is because the statistical power of the homogeneity test depends on the size of heterogeneity, which was not manipulated in this study as we focus primarily on estimating the mean effect size.

Type I error rate in the range of [.025, .075] is deemed reasonable (Bradley, 1978). For the r-based

approach, when true scores were normal, results showed that the type I error rates were in general reasonable when NS=5; type I error rates tended to be smaller when NS increased to 50. When the true scores were skewed, type I error rates were dependent on $\bar{\rho}$, J, NS and $\bar{N}$. In particular, when $\bar{\rho}$, J, NS and/or $\bar{N}$ increased, type I error rates tended to increase. For example, when $\bar{\rho} = .5$, $J = 12$, NS=50 and $\bar{N} = 200$, the type I error rate was .905 (table 7). Meeting or violating tau-equivalence seemed not to have as much influence on type I error rates of the homogeneity test.

In contrast, for the z-based approach, type I error rates were in general inflated across conditions. As NS increased, inflation of type I error rate became more serious. We use the situation when $\bar{\rho} = 0$ and NS=5 and 50 for illustration (figure 2). For instance, when $\bar{\rho} = 0$ and NS=50, the type I error rates of homogeneity test were all 1 when the true scores were skewed across all other conditions. Type I error rates using the z-based approach under other conditions were not better. This indicates that using Equation 16 as the estimated within-study sampling error variance for the z-based approach seems to be problematic for the homogeneity test.

—Insert figure 1 here—

—Insert table 7 here—

—Insert figure 2 here—

Simulation study 3: Investigating the influence of coarse categorization on mean correlation estimates when coefficient alpha is used for correcting measurement error

In simulation study 3, we examined the influence of coarse categorization. In order to highlight its impact, we selected the most ideal situation among those used in the former two studies in terms of J, $\bar{N}$, and NS. In particular, $J = 12$, $\bar{N} = 200$, and NS=50. Two item categories were used: 2 and 5. Furthermore, we considered both symmetric and asymmetric ordinal items scores. For generating symmetric items, with 2 categories, the cutoff was the 50th percentile (skewness = 0, excess kurtosis = -2);[10] with 5 categories, the cutoffs were the 10th, 30th, 70th, and 90th percentiles (skewness = 0, excess kurtosis = -.5). For generating asymmetric items, with 2 categories, the cutoff was the 80th percentile (skewness = 1.50, excess kurtosis = .25); with 5 categories, the cutoffs were the 40th, 65th, 85th, and 95th percentiles (skewness = .76, excess kurtosis = -.41). Item distributions under 2 and 5 categories are shown

in Figure 3.

—Insert figure 3 here—

In practice, items on the same scale can have similar or different distributions. Therefore, four scenarios were included: (1) All the items were symmetric as in Figure 3a; (2) All the items were right skewed as in Figure 3b; (3) Half of the items were symmetric as in Figure 3a and the other half were right skewed as in Figure 3b; (4) Half of the items were right skewed as in Figure 3b, and the other half were mirror symmetric to the distributions in Figure 3b (right skewed becoming left skewed with the same kurtosis yet opposite skewness). The last scenario mimicked the case that the distribution of scale scores is symmetric although the item scores are not. Distributions of the scale scores under the four scenarios are shown in Figure 4 with 2-category items and Figure 5 with 5-category items.

—Insert figure 4 here—

—Insert figure 5 here—

Therefore, there were 2 types of item categories and 4 types of distribution combinations for the categorical items. Other conditions included $\bar{\rho}$= 0, .1, .3, and .5, two levels of between-studies variance and tau-equivalence status, which were specified in the same way as in simulation study 2 to generate the underlying continuous items. All these conditions were cross-tabulated for generating data. In total, there are 64 conditions. Under a certain condition, the cutoff values mentioned above were applied to the generated continuous item scores in each primary study of a replicated data set. Estimates and coverage probabilities were compared to the true nominal $\bar{\rho}$.

Estimation accuracy of $\bar{\rho}$ and coverage probability. With 2-category items, even when tau-equivalence was met, coverage rates can be very low. For example, even under the condition that all the items were symmetric, having a skewed true score distribution can yield coverage probabilities of less than 5% under both the r-based and z-based approaches (e.g., $\bar{\rho} = .3$ and effect sizes were homogeneous; table 8). When tau-equivalence was not met, in general bias increased in magnitude comparing to those under tau-equivalence holding constant other factors. It is worth mentioning that under the situation when the scale scores were symmetric (when all the items were symmetric or a half were left skewed and the other half were right skewed), a situation that many researchers believe to be relatively better than having

asymmetrically distributed scale scores (Micceri, 1989), the mean correlation can be much overestimated and the coverage probability can be close to or equal to zero. For example, with the correction of measurement error using coefficient alpha, when $\bar{\rho} = .3$ and the true scores were skewed, the bias was .078 and .116 and the coverage probabilities were .001 and 0 when all the items were symmetric and when a half were left skewed and the other half were right skewed respectively.

When the true scores were normal, increasing the number of categories to 5 helped improve estimation accuracy and coverage probabilities a lot with or without tau-equivalence. When the true scores were skewed, however, most of the CPs were still not in the acceptable range with 5 categories with or without meeting tau-equivalence (table 9).

—Insert table 8 here—

—Insert table 9 here—

Type I error rate of the homogeneity test. Type I error rates using the r-based approach were presented (table 10). For the z-based approach, with coarsely categorized items, type I error rates were still seriously inflated and not better than those with continuous items (see figure 2). When tau-equivalence was met and the true scores were normal, type I error rates were in general reasonable with 5-category items; with 2-category items, type I error rates were inflated when these items were right skewed and $\bar{\rho} \neq 0$. When tau-equivalence was met, inflation of type I error rates occurred more frequently when the true scores were skewed than when they were normal.

When true scores were normal, comparing the condition when tau-equivalence was not met to that when tau-equivalence was met, type I error rates generally increased. When the true scores were skewed, type I error rates tended to be inflated when $\bar{\rho}$ was medium to large.

Thus, with categorical items, performance of type I error rates depended on the number of categories of an item, the distribution of observed data, the distribution of true scores and whether tau-equivalence was met.

—Insert table 10 here—

Discussion

The paper examined meta-analysis approaches using raw correlations and Fisher's z transformed correlations (r- and z-based approaches) when individual correlations were obtained using scale scores. The focal parameter of interest in the study is the mean correlation. Random-effects models were adopted for both approaches. Restricted maximum likelihood (REML) estimation was used for estimating the between-studies variance, followed by weighted least squares (WLS) estimation of the mean effect size. Three simulation studies were conducted. The first study concerns the impact of true score distribution only, the second study investigates the impact of violating the tau-equivalence assumption when using coefficient alpha for correcting measurement error, and the third one examines the influence of discretization of continuous items and how it interacts with the distribution of true scores and violation of tau-equivalence. In the first simulation study, mean effect size estimate obtained using uncorrected correlations was compared to the true attenuated population mean correlation; in the second and third simulation studies, mean effect size estimate obtained using correlations corrected by coefficient alpha was compared to the population mean effect size under perfect measurement. R code for data generation can be found at https://github.com/qzhang4/metacorr.

Summary of findings

Based on the simulation study results, there are several findings. First, considering the distribution of true scores only, skewed scores made the mean effect size estimate and inference less accurate especially when $\bar{\rho}$ was large. Second, although researchers often use coefficient alpha as the reliability estimate to correct measurement error, violation of tau-equivalence can result in low coverage and upward bias. Coverage probabilities tended to be lower and bias tended to be greater in magnitude as the population mean effect size increased. Third, when observed correlations were corrected using coefficient alpha, the r-based approach and the z-based approach had less accurate estimates and lower coverage probabilities when tau-equivalence was not satisfied than when it was satisfied. The homogeneity test result depends on true score distribution. In addition, the r-based approach had better controlled type I error rate than the z-based approach under each of the considered conditions. Fourth, discretization of continuous items

interacted with the distribution of the true scores to influence mean effect size estimates and coverage probabilities. Specifically, when the true scores were normal, results can improve a lot with increasing the number of categories with tau-equivalence met or not; in contrast, when the true scores were skewed, increasing the number of categories did not improve results much with violation of tau-equivalence. Coverage probability of the true mean effect size can be very close to or equal to zero with skewed true scores, even when ordinal scale scores were symmetric.

We would like to emphasize that the conclusions above were made under specific estimation approaches. For example, Equation 16 was used for estimating the sampling error variance of the Fisher's z transformed corrected and uncorrected raw correlations; REML is used for estimating the between-studies variance. When one or more of the estimation approaches are changed, we expect to see some differences in the results. However, the negative impact of using coefficient alpha with violation of tau-equivalency and of discretizing continuous items should still hold.

Recommendations based on the findings

Based on the findings, several recommendations are made with respect to factors considered in the simulation studies. These recommendations are made for researchers of primary studies and meta-analysts.

Advice to researchers of primary studies to facilitate meta-analysis of correlations. Extending from former research on meta-analyzing correlations with normal true scores (e.g., Field, 2001, 2005; Hafdahl, 2010; Hall & Brannick, 2002), the study found that the distribution of the true scores played a key role in accuracy of the average correlation when ordinal scales are used for measuring X and Y. Therefore, for researchers of primary studies, we suggest careful examination of the theoretical distribution of the underlying constructs when meta-analyzing correlations obtained from ordinal scales. For example, Micceri (1989) showed based on 440 achievement/ability, criterion/mastery, and psychometric measures with sufficiently large samples (most of them were greater than 450) that more than 80% of the measures deviated from normality and were asymmetric. Therefore, it is critical *not to ignore* the influence of underlying distributions of X and Y on individual correlation estimates across primary studies, in particular when ordinal scales are used. Information about the theoretical distribution of psychological

constructs can be based on substantive theory or findings from prior studies in the same field. Moreover, a tutorial of using statistical software (e.g., SAS, SPSS, R) and a web application to compute skewness and kurtosis can be found in Cain, Zhang, and Yuan (2016). With these statistical tools, we encourage researchers of primary studies to report skewness and kurtosis in addition to other descriptive statistics such as means and variances to have a clear description about the distribution of interested variables.

We would like to also emphasize that all the conclusions in this study rests on validity, which is not less important than the issue of reliability in psychological and educational measurement. By employing a modern test theory model (e.g., confirmatory factor analysis, or CFA model), one can examine the structure of constructs and validity provided large enough sample sizes (Borsboom & Mellenbergh, 2002). Specifically, only when it is found that unidimensional models fit the scales of X and Y can one use a single correlation to depict their linear association. Meanwhile, employing the CFA model also allows examination of tau-equivalence. With a large enough sample size (e.g., ≥ 100; Rhemtulla et al., 2012), we strongly encourage researchers of primary studies to evaluate tau-equivalence (e.g., using a CFA model, loadings should be equal or approximately equal) especially when the number of categories per item is small (e.g., 2), number of items per scale is small and/or the theoretical true scores are asymmetric. Selection of approaches of fitting CFA models with symmetric/asymmetric categorical item scores and true scores are discussed in Rhemtulla et al. (2012) in detail. Reporting of model fit and model parameters (e.g., factor loadings and error variances) under CFA models would be helpful for meta-analysis researchers to evaluate the quality of measurement of the scales and to accurately correct measurement error. For example, with provided factor loading and error variance estimates meta-analysts can use coefficient omega that relaxes the tau-equivalence assumption as an alternative to coefficient alpha for estimating reliability (McDonald, 1999).

Recommendations to meta-analysis researchers. For meta-analysts, we give the following cautions/recommendations when meta-analyzing correlations. The tau-equivalence assumption is important when using coefficient alpha as the reliability estimate. When tau-equivalence is violated, reliability estimated using coefficient alpha and thus the corrected individual correlations are influenced. With a relatively small number of continuous items (e.g., 4) per scale, violation of tau-equivalence alone

typically resulted in an overestimated mean correlation and a low coverage probability when the mean population correlation is nonzero; when the number of items is large enough (e.g., $\geq$12), the estimated mean correlation becomes more accurate and coverage rate would increase. For the latter case, however, coverage probability with large effect sizes on average still tended to be lower than 90%. Therefore, no matter tau-equivalence is met or not, when the primary studies have continuous items and large enough numbers of items per scale (e.g., >12), meta-analyzing correlations that are corrected by coefficient alpha typically results in reasonably accurate mean effect size estimates and inferences when the mean effect sizes range from zero to medium and scores are close to normality. With ordinal scales, we found that the situation with a null to medium mean effect size, a relatively large number of categories per item (e.g., 5), large enough number of items (e.g., 12) and normal true scores generally yielded estimated mean effect size with bias close to zero and relatively stable coverage probabilities in the nominal range with all kinds of item distributions considered. This conclusion regarding ordinal scales holds no matter tau-equivalence is met or not. Therefore, meta-analysts using ordinal scales need not worry too much about violation of tau-equivalence in this situation.

A proposed approach to correcting sampling error variance using Fisher's z transformed correlations

In the paper, we used the z-based approach with Equation (16) as the sampling error variance. It has been found in the study, however, that this z-based approach had inflated type I error rate of the homogeneity test. There are some discussions regarding correcting measurement error with Fisher's z transformed correlations (e.g., Hall & Brannick, 2002; L. Hedges, 1988), yet little about correcting the sampling error variance. In light of the study by Bonett (2008), we derived a formula to correct the sampling error variance of $z_{rc,i}$ in each study by using the delta method (Casella & Berger, 2002) based on the relation $z_{rc,i} = \tanh^{-1}(r_{ci})$ and the estimated sampling error variance of r_{ci}:

$$\breve{\sigma}^2_{zc,i} = \frac{1}{(1 - r^2_{ci})^2}\hat{\sigma}^2_{rc,i}, \tag{20}$$

where r_{ci} is the raw correlation in study i corrected for measurement error and $\hat{\sigma}^2_{rc,i}$ is from Equation 10. In practice, however, when r_{ci} is close to 1 or -1, $\breve{\sigma}^2_{zc,i}$ corresponding to this r_{ci} can be huge. This may result

in unstable estimation. Therefore, another formula to correct the sampling error variance is proposed as:

$$\tilde{\sigma}^2_{zc,i} = \frac{1}{(1-\bar{r}_c^2)^2}\hat{\sigma}^2_{rc,i}, \tag{21}$$

where $\bar{r}_c = \sum_{i=1}^{NS} r_{ci}/NS$ is the simple mean of the corrected correlations across studies. A small simulation was conducted to examine performance of the z-based approach using Equations 20 and 21 to obtain sampling error variance estimates with measurement error. We considered the scenario when NS = 50, number of items per scale $J = 12$, and true scores were normal. Continuous items were simulated without discretization. Other factors were the same as those used in simulation study 2.

In each replication, results in which the ratio of the the largest to the smallest corrected sampling error variance across primary studies that was greater than 1000 were discarded. Across conditions, the largest rate of having such data was 11.6% and 0.5% using Equations 20 and 21, respectively. In terms of estimation accuracy and coverage probability, the z-based approach using Equation 21 had more accurate mean effect size estimates and reasonable coverage probabilities than that using Equation 20 in general with or without meeting tau-equivalence; this approach had results close to the r-based approach for most conditions yet slightly worse when $\bar{\rho} \geq .3$ and tau-equivalence was violated (see table 6).

In terms of the homogeneity test, it was found that the z-based approach using Equation 20 tended to be very conservative (type I error rates close to or equal to 0); in contrast, the z-based approach using Equation 21 tended to be liberal when the average sample sizes were small, and had in general well-controlled type I error rates when $\bar{N} \geq 100$. Comparing to the z-based approach with Equation 16 as the sampling error variance estimate, the type I error rates were reduced substantially and became much more reasonable. Both the proposed formulas are expected to yield reasonable homogeneity test results when $\bar{N}$ becomes large enough (e.g., >200). Considering both parameter estimation and the homogeneity test, it seems that the approach using Equation 21 is more promising. This approach needs to be examined more and can be compared to the sampling error variance formula proposed by L. Hedges (1988).

Limitations and future directions

As is true for any simulation study, the current study examined a limited number of factors and levels. In the future, other conditions may be considered regarding other issues of using coefficient alpha as the reliability estimate (e.g., multi-dimensionality; Sijtsma, 2009). Also, it is worthwhile to consider using coefficient omega (Equation 3) for correcting measurement error. Factor loadings may be obtained using CFA with the provided correlations matrices in primary studies (Ahn, 2008). Furthermore, we studied correcting measurement error for meta-analyzing correlations between one pair of variables. More generally, it would be helpful to study how to correct measurement error for each element in a correlation matrix that are obtained with scale scores, and other potential issues with meta-analysis of correlation matrices (e.g., missing elements in a correlation matrix) for examining multivariate relations (e.g., Becker, 1992; M. W.-L. Cheung & Chan, 2005; Ke et al., 2019). In addition, artifacts besides measurement error are often present in meta-analysis; how to accurately correct these artifacts together can be of future interest.

References

Ahn, S. (2008). *Application of model-driven meta-analysis and latent variable framework in synthesizing studies using diverse measures* (Unpublished doctoral dissertation). Michigan State University.

Baird, R., & Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods*, *21*, 175-188. doi: https://doi.org/10.1037/met0000070

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*, 341-362. doi: https://doi.org/10.3102/10769986017004341

Becker, B. J., & Fahrbach, K. (1994). A comparison of approaches to the synthesis of correlation matrices. In *the annual meeting of the American Educational Research Association.* New Orleans, LA..

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley. doi: https://doi.org/10.1002/9781118619179

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13*, 173–181. doi: 10.1037/a0012868

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs. *Intelligence*, *30*(6), 505–514. doi: 10.1016/s0160-2896(02)00082-x

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi: https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716–1735. doi: 10.3758/s13428-016-0814-1

Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd edition)*. Thomson Learning.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*, 40-64. doi: 10.1037/1082-989X.10.1.40.

Cheung, S. F., Chan, D. K.-S., & Sun, R. W. (2018). Meta-analyzing dependent correlations with correction for artifacts that multiplicatively attenuate the true correlation. *Behavior Research Methods*, *51*, 793–810. doi: 10.3758/s13428-018-1111-y

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (second ed.)*. Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi: http://dx.doi.org/10.1007/BF02310555

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A monte carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161-180. doi: https://doi.org/10.1037//1082-989x.6.2.161

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 105–124. doi: 10.1207/s15328031us0202_02

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444-467. doi: https://doi.org/10.1037/1082-989x.10.4.444

Finney, S., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.). In *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. Structural equation modeling: A second course* (p. 439-492). Charlotte, NC, US: IAP Information Age Publishing.

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

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods*, *10*(2), 227–254. doi: 10.1037/1082-989x.10.2.227

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *7*, 1-13.

Gatzka, T., & Hell, B. (2018). Openness and postsecondary academic performance: A meta-analysis of facet-, aspect-, and dimension-level correlations. *Journal of Educational Psychology*, *110*, 355-377. doi: http://dx.doi.org/10.1037/edu0000194

Glass, G. V., MacGaw, B., & Smith, M. L. (1984). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement*, *66*, 930-944. doi: http://dx.doi.org/10.1177/0013164406288165

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121-135. doi: http://dx.doi.org/10.1007/s11336-008-9098-4

Hafdahl, A. R. (2008). Combining heterogeneous correlation matrices: Simulation analysis of fixed-effects methods. *Journal of Educational and Behavioral Statistics*, *33*(4), 507–533. doi: 10.3102/1076998607309472

Hafdahl, A. R. (2009). Improved fisher z estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, *62*, 233–261. doi: 10.1348/000711008x281633

Hafdahl, A. R. (2010). Random-effects meta-analysis of correlations: Monte carlo evaluation of mean estimators. *British Journal of Mathematical and Statistical Psychology*, *63*, 227–254. doi: 10.1348/000711009x431914

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*, 377–389. doi: 10.1037/0021-9010.87.2.377

Hedges, L. (1988). The meta-analysis of test validity studies: Some new approaches. In H. Braun & H. Wainer (Eds.), *Test validity.* Lawrence Erlbaum.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388-395. doi: http://dx.doi.org/10.1037/0033-2909.93.2.388

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504. doi: https://doi.org/10.1037//1082-989x.3.4.486

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.)*, *327*, 557-560. doi: 10.1136/bmj.327.7414.557

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292. doi: 10.1111/1468-2389.00156

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications. doi: https://doi.org/10.4135/9781483398105

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology*, *100*, 298-342. doi: http://dx.doi.org/10.1037/a0037681

Ke, Z., Zhang, Q., & Tong, X. (2019). Bayesian meta-analytic SEM: A one-stage approach to modeling between-study heterogeneity in structural parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 348-370. doi: https://doi.org/10.1080/10705511.2018.1530059

Lambert, R. G. (1996, April). The robustness of the standard error of summarized, corrected validity coefficients to non-independence and non-normality of primary data. *Dissertation Abstracts International Section A: Humanities and Social Sciences*(AAM9602815), 3928.

Law, K. S. (1995). The use of Fisher's Z in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics*, *20*, 287-306. doi: https://doi.org/10.3102/10769986020003287

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology*, *7*(4), 478–500. doi: 10.1111/iops.12184

Lilienfeld, S. O., Smith, S. F., Sauvigne, K. C., Patrick, C. J., Drislane, L. E., Latzman, R. D., & Krueger, R. F. (2016). Is boldness relevant to psychopathic personality? Meta-analytic relations with non-psychopathy checklist-based measures of psychopathy. *Psychological Assessment*, *28*, 1172-1185. doi: http://dx.doi.org/10.1037/pas0000244

Litwiller, B., Snyder, L. A., Taylor, W. D., & Steele, L. M. (2017). The relationship between sleep and work: A meta-analysis. *Journal of Applied Psychology*, *102*, 682-699. doi: http://dx.doi.org/10.1037/apl0000169

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166. doi: https://doi.org/10.1037//0033-2909.105.1.156

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*(1), 63–75. doi: 10.1177/0013164496056001004

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Oswald, F. L., Ercan, S., McAbee, S. T., Ock, J., & Shaw, A. (2015). Imperfect corrections or correct imperfections? Psychometric corrections in meta-analysis. *Industrial and Organizational Psychology*, *8*(2), e1–e4. doi: 10.1017/iop.2015.17

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*, 432–446. doi: 10.1037/0021-9010.76.3.432

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In (pp. 295–315). New York, NY, US: Russell Sage Foundation.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354-373. doi: 10.1037/a0029315

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications. doi: https://doi.org/10.2307/2348600

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*, 31–48. doi: 10.1037/1082-989x.13.1.31

Seide, S. E., Röver, C., & Friede, T. (2019). Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC Medical Research Methodology*, *19*, 16. doi: 10.1186/s12874-018-0618-3

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120. doi: 10.1007/s11336-008-9101-0

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72-101. doi: http://dx.doi.org/10.2307/1412159

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6–20. doi: 10.1177/0013164498058001002

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*, 55–79. doi: 10.1002/jrsm.1164

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261-293. doi: https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metaforpackage. *Journal of Statistical Software*, *36*, 1-48. doi: 10.18637/jss.v036.i03

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070. doi: https://doi.org/10.1037//0022-3514.54.6.1063

Yang, Y. (2018). Omega. in Bruce B. Frey (Ed.),. In *Sage encyclopedia of educational research, measurement, andevaluation.* (p. 1177-1180). SAGE.

Zhang, Q., Wang, L., & Bergeman, C. S. (2018). Multilevel autoregressive mediation models: Specification, estimation, and applications. *Psychological Methods*, *23*, 278-297. doi: 10.1037/met0000161

Footnotes

[1]In practice, reliability estimates may not be available and/or other artifacts (e.g., range restriction) in addition to measurement error exist. We encourage readers to refer to relevant studies that comprehensively consider various artifacts and/or missing information on these artifacts when meta-analyzing correlations (e.g., Cheung, Chan, & Sun, 2018; Hunter & Schmidt, 2004; Hall & Brannick, 2002; Raju, Burke, Normand, & Langlois, 1991).

[2]The conclusion under Caveat 2 can be generalized to multi-dimensional scales. We focus on uni-dimensional scales in the study.

[3]To handle this issue, polychoric covariance/correlation matrices can be used for estimating the covariances/correlations among the underlying continuous items $Item_j^*$ that are assumed to be normally distributed. Gadermann, Guhn, and Zumbo (2012) showed the use of polychoric covariance matrices for obtaining reliability for the underlying items $Item_j^*$ to correct for underestimation of reliability due to ordinal items.

[4]It is also the REML estimate when the normality of residuals e_{ci} and u_{ci} is assumed.

[5]In the study, we selected this simple and popular way of obtaining an interval estimate for $\bar{\rho}_c$. It is worth noting that researchers have proposed alternative interval estimates of $\bar{\rho}_c$ (e.g., the *t* distribution-based confidence interval, the weighted variance confidence interval, the Bayesian credible interval; Sánchez-Meca & Marín-Martínez, 2008; Seide, Röver, & Friede, 2019) to account for the uncertainly in estimation of $\hat{\sigma}^2_{rc,i}$ and $\hat{\tau}^2_c$ and/or violation of data assumptions. Choosing different interval estimates would influence the statistical inference of $\bar{\rho}_c$. However, it is not our focus to compare different approaches of obtaining interval estimates in this study. Interested readers may examine performances of using various interval estimates for $\bar{\rho}_c$. Choosing interval estimates is also relevant for the meta-analysis method using Fisher's z transformed correlations, which is to be discussed in the next section.

[6]When Fisher's z transformed correlations are used, Hall and Brannick (2002) discussed correcting the mean effect size estimate and the between-studies variance estimate. L. Hedges (1988) discussed correcting individual z-transformed correlation estimates and proposed to use a formula similar to Equation 10 for estimating the corrected sampling error variance for measurement error by replacing $\hat{\sigma}^2_{ro,i}$

with $\sigma^2_{z,i}$ despite the different metrics of raw and z-transformed correlations. Thus far, however, we are not aware of any studies that have examined comprehensively the sampling error variance estimator using Fisher's z transformed corrected correlations.

[7]When generating scores of ω_{1i} and ω_{2i}, we used population means and standard deviations instead of their sample means and standard deviations for standardization to avoid the dependence of standardized scores on sample sizes.

[8]Here tau-equivalence is met. For the purpose of simulation study 1, however, meeting or violating the tau-equivalence assumption was not considered as we use the true reliabilities (based on equation 3 given assumptions were met for coefficient omega) for evaluating the estimated mean correlation. This is only one way to generate specific levels of reliabilities. In the next simulation study, the influence of violating the tau-equivalence assumption when using coefficient alpha will be considered.

[9]Across conditions, issues occurred most frequently when $\bar{N}$ and the number of items were small (i.e., $\bar{N} = 30$ or 50, $J = 4$). Under these scenarios, the larger $\bar{\rho}$ was, the more likely these issues would occur. With $\bar{N} \geq 100$ and $J = 12$, these issues barely happened. We would like to also note that discarding data that are not applicable may have influence on meta-analysis results. For example, we compared the results with and without discarding data based on estimates of uncorrected correlations (as in simulation study 1) under the condition that $\bar{\rho} = .5$, $\bar{N} = 30$, NS=5, and $J = 4$. Results showed that discarding data yielded larger bias in magnitude than not discarding data (the differences were .05 and .13 respectively for normal and skewed true scores). Coverage probabilities were comparable with and without discarding data. One alternative approach that may be investigated is to replace problematic estimates of corrected correlations by admissible values that are very close to 1 in magnitude.

[10]All skewness and excess kurtosis for ordinal items were obtained under a sample size of 1,000,000.

Table 1: The mean percentage of between-studies variance averaged across studies over 1000 replications for the r-based and z-based approaches

NS	$\bar{N}$	r-based approach				z-based approach
		$\bar{\rho} = 0$	$\bar{\rho} = .1$	$\bar{\rho} = .3$	$\bar{\rho} = .5$	
5	30	45.27	45.26	45.47	45.38	42.35
	50	52.41	52.46	52.39	52.47	49.92
	100	68.17	68.14	68.01	67.79	66.48
	200	81.14	81.12	80.98	80.86	80.17
20	30	41.09	41.10	41.18	41.30	37.96
	50	54.66	54.65	54.67	54.51	52.28
	100	70.19	70.15	70.06	69.80	68.64
	200	79.65	79.65	79.49	79.18	78.62
50	30	40.33	40.35	40.40	40.48	37.15
	50	50.14	50.14	50.12	50.16	47.49
	100	69.08	69.07	68.94	68.70	67.48
	200	82.54	82.52	82.41	82.11	81.65

Table 2: Examining the influence of true score distribution with respect to mean correlation estimates and coverage probabilities with 4 continuous items on scales of X and Y.

		Meta-analysis using raw correlations								Meta-analysis using z transformed scores							
		$\bar{\rho} = 0$				$\bar{\rho} = .5$				$\bar{\rho} = 0$				$\bar{\rho} = .5$			
		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed	
NS	$\bar{N}$	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
									Homogeneous case								
5	30	-.004	.964	-.002	.951	-.008	.958	-.008	.934	-.004	.968	-.001	.959	-.002	.968	-.004	.939
	50	.000	.960	-.001	.966	-.003	.968	-.016	.937	.000	.962	-.001	.970	-.001	.968	-.012	.938
	100	.000	.955	-.002	.957	.002	.958	-.009	.928	.000	.958	-.002	.964	.003	.960	-.006	.931
	200	-.001	.964	-.001	.964	.000	.955	-.005	**.920**	-.001	.964	-.001	.965	.001	.955	-.004	**.923**
20	30	.002	.962	-.002	.962	-.005	.950	-.018	**.916**	.002	.964	-.001	.965	.005	.951	-.008	.941
	50	.003	.952	.001	.953	-.003	.965	-.013	.926	.003	.953	.002	.957	.002	.964	-.005	.932
	100	-.001	.964	.000	.965	-.001	.961	-.007	.925	-.001	.964	.000	.964	.001	.963	-.004	.931
	200	.000	.963	-.001	.963	-.001	.953	-.002	.932	.000	.964	-.001	.963	.000	.959	-.001	.938
50	30	.000	.961	.002	.960	-.006	.960	-.020	**.898**	.000	.964	.002	.961	.005	.952	-.007	.946
	50	-.001	.957	.000	.951	-.003	.958	-.011	.929	-.001	.958	.001	.951	.003	.960	-.004	.949
	100	.000	.949	.001	.956	-.001	.949	-.007	.926	.000	.951	.001	.956	.002	.947	-.003	.932
	200	.000	.964	.000	.966	-.001	.955	-.003	.934	.000	.965	.000	.967	.001	.953	-.001	.957
									Heterogeneous case								
5	30	-.001	.939	.002	**.922**	-.007	.936	-.030	.932	-.001	.944	.002	.931	.000	.942	-.022	.931
	50	-.001	**.921**	.002	**.919**	-.002	.925	-.017	**.910**	-.001	.926	.002	.926	.001	.933	-.013	**.911**
	100	-.001	**.890**	.001	**.900**	.002	**.903**	-.010	**.906**	-.001	**.895**	.001	**.905**	.003	**.905**	-.009	**.916**
	200	.002	**.898**	-.001	**.894**	-.002	**.879**	-.003	**.881**	.002	**.900**	-.001	**.896**	-.002	**.878**	-.003	**.878**
20	30	.000	.950	.002	.929	-.004	.940	-.020	**.923**	.000	.954	.003	.931	.006	.941	-.009	.935
	50	-.001	.933	-.001	.942	-.003	.939	-.011	**.922**	-.001	.933	.000	.942	.003	.941	-.004	.931
	100	.000	.927	.000	**.915**	-.001	.943	-.007	.933	.000	.927	.000	**.916**	.002	.941	-.004	.938
	200	.000	.925	-.001	.927	-.002	.932	-.003	.937	.000	.925	-.001	.928	.000	.929	-.002	.936
50	30	-.001	.937	.001	.947	-.006	.942	-.021	**.892**	-.001	.936	.001	.950	.006	.934	-.009	.942
	50	.000	.942	.000	.934	-.003	.936	-.013	**.914**	.000	.943	.001	.934	.003	.939	-.006	.948
	100	-.001	.951	.000	.941	-.002	.946	-.007	.932	-.001	.951	.000	.943	.002	.943	-.003	.937
	200	.000	.950	.000	.956	-.001	.944	-.003	.945	.000	.951	.000	.957	.001	.941	-.001	.948

Note. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 3: Meta-analysis with correction of measurement errors using the r-based approach: mean correlation estimates and coverage probabilities when the tau-equivalence assumption is met with 4 continuous items on scales of X and Y.

NS	$\bar{N}$	$\bar{\rho}=0$				$\bar{\rho}=.1$				$\bar{\rho}=.3$				$\bar{\rho}=.5$			
		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed	
		Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
									Homogeneous case								
5	30	.004	.969	.001	.964	-.006	.944	.006	.965	.005	.964	.014	.951	.002	.953	-.003	.932
	50	-.006	.960	-.001	.962	.002	.966	.009	.953	.003	.956	.017	.941	.003	.960	.006	**.926**
	100	.002	.946	.000	.958	-.001	.966	.009	.954	.002	**.977**	.009	.955	.000	.967	.012	.936
	200	.000	.962	.000	.959	.000	.955	.000	.969	.005	.952	.004	.954	.001	.965	.002	**.924**
20	30	.000	.957	-.002	.946	.003	.961	.007	.967	.001	.962	.015	.947	.002	.960	.010	.943
	50	.000	.959	.001	.950	-.001	.953	.006	.961	.004	.961	.016	.939	.005	.973	.008	.943
	100	.000	.963	.000	.958	-.002	.974	.003	.959	.002	.956	.007	.928	.003	.966	.009	.931
	200	-.001	.954	.000	.959	-.001	.962	.002	.972	.000	.962	.003	.943	.002	.962	.004	.933
50	30	-.001	.961	-.002	.965	.002	.963	.008	.957	.004	.962	.017	.932	.003	.970	.008	.929
	50	.000	.963	-.001	.962	.002	.950	.007	.947	.004	.941	.014	.933	.007	.952	.012	.931
	100	.000	.957	.000	.949	.000	.954	.003	.944	.002	.952	.008	.935	.003	.948	.009	**.924**
	200	.000	.958	.000	.949	.001	.968	.002	.945	.000	.952	.004	.931	.001	.964	.003	.943
									Heterogeneous case								
5	30	.000	.928	.000	**.921**	.002	.929	.002	.930	.000	.950	.016	.929	.000	.948	.012	**.919**
	50	.000	**.907**	.002	.925	-.002	.926	.008	.939	.005	**.914**	.007	**.916**	.007	.926	.010	**.898**
	100	-.001	**.908**	.007	**.909**	.002	**.919**	.007	**.900**	.003	**.915**	.014	**.901**	.001	**.910**	.005	**.879**
	200	-.006	**.889**	-.003	**.882**	-.003	**.879**	.003	**.887**	-.001	**.890**	.005	**.882**	-.002	**.883**	.006	**.882**
20	30	.001	.938	-.003	.935	.003	.945	.005	.948	.003	.945	.011	.944	-.004	.954	.006	**.922**
	50	-.001	.932	.000	.926	.003	.934	.004	.939	.001	.942	.012	**.924**	.001	.946	.010	**.912**
	100	.000	**.922**	-.001	.949	-.002	.933	.000	.953	.001	.930	.009	**.919**	.003	**.922**	.007	**.923**
	200	.000	.937	.002	.948	.001	.938	.002	.932	.001	.944	.001	.938	.000	.938	.004	.935
50	30	.000	.947	-.002	.950	.001	.942	.004	.939	.004	.948	.013	.929	.000	.949	.005	.939
	50	-.001	.939	-.001	.961	.002	.941	.005	.940	.003	.928	.012	.935	.004	.947	.008	.926
	100	.001	.947	.000	.936	.001	.942	.002	.945	.002	.948	.007	.939	.003	.944	.007	**.922**
	200	.000	.953	-.001	.928	.001	.945	.002	.939	.000	.945	.004	.937	.001	.947	.003	.939

Note. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 4: Meta-analysis with correction of measurement errors using the z-based approach: mean correlation estimates and coverage probabilities when the tau-equivalence assumption is met with 4 continuous items on scales of X and Y.

		$\bar{\rho} = 0$				$\bar{\rho} = .1$				$\bar{\rho} = .3$				$\bar{\rho} = .5$			
NS	$\bar{N}$	T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed	
		Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
Homogeneous case																	
5	30	.003	**.920**	-.002	.925	-.009	**.908**	.004	**.917**	.001	**.918**	.000	**.917**	-.010	**.907**	-.034	**.911**
	50	-.007	.925	-.001	**.918**	.003	.930	.010	**.908**	.003	**.909**	.010	**.911**	.003	**.902**	-.015	**.898**
	100	.001	**.911**	.000	**.913**	-.002	.928	.009	**.914**	.003	**.924**	.009	**.908**	.001	**.912**	.010	**.902**
	200	.001	**.914**	.000	**.920**	.000	**.895**	.000	.926	.006	**.891**	.004	**.905**	.001	**.901**	.002	**.893**
20	30	.002	.934	-.001	.930	.004	.938	.004	.948	.003	.945	.001	.943	-.003	.943	-.017	.935
	50	.000	.938	.003	.930	.000	.934	.008	.945	.007	.945	.017	.933	.008	.946	-.004	.957
	100	.000	.939	.000	.932	-.001	.951	.006	.949	.003	.932	.010	**.921**	.006	.945	.008	.947
	200	-.001	.927	.000	.949	.000	.930	.002	.943	.001	.937	.004	.935	.003	.927	.004	.940
50	30	.000	.956	-.001	.943	.003	.952	.006	.953	.007	.949	.010	.944	.000	.958	-.019	.938
	50	.000	.955	.000	.946	.003	.942	.011	.939	.009	.930	.015	.937	.011	.954	.005	.961
	100	.000	.932	.001	.939	.002	.932	.005	.938	.004	.931	.011	.932	.007	.933	.009	.943
	200	.000	.940	.000	.939	.001	.952	.002	.944	.001	.940	.006	.930	.003	.944	.004	.959
Heterogeneous case																	
5	30	.001	**.900**	.004	**.897**	-.002	**.899**	-.001	**.899**	-.014	**.905**	-.018	**.904**	-.010	.927	-.010	**.913**
	50	.000	**.891**	.003	**.900**	-.003	**.897**	.004	**.906**	.002	**.890**	-.003	**.910**	.003	**.894**	-.010	**.880**
	100	-.001	**.880**	.006	**.885**	.001	**.893**	.006	**.887**	.002	**.897**	.011	**.900**	-.001	**.887**	-.001	**.881**
	200	-.006	**.891**	-.003	**.882**	-.003	**.875**	.003	**.888**	-.002	**.874**	.003	**.881**	-.003	**.864**	.003	**.877**
20	30	.002	.946	-.001	.941	.003	.940	.002	.952	.000	.941	-.009	.943	-.018	.944	-.030	.934
	50	.000	.936	.001	.941	.005	.939	.005	.948	.002	.949	.009	.944	.002	.951	-.004	.945
	100	.001	**.923**	-.001	.953	.000	.936	.001	.946	.003	.932	.010	.937	.004	.941	.004	.947
	200	.000	.934	.002	.955	.002	.938	.003	.939	.002	.945	.002	.938	.001	.941	.003	.950
50	30	.000	.948	-.001	.948	.000	.945	.002	.948	.002	.946	-.001	.953	-.005	.949	-.021	.942
	50	-.001	.944	-.001	.958	.003	.945	.007	.955	.007	.948	.010	.952	.006	.964	-.003	.966
	100	.001	.952	.001	.941	.002	.938	.003	.940	.004	.954	.009	.945	.005	.955	.004	.954
	200	.000	.949	-.001	.933	.001	.942	.002	.947	.001	.941	.004	.945	.003	.950	.003	.951

Note. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 5: Meta-analysis with correction of measurement error using the r-based approach: mean correlation estimates and coverage probabilities when the tau-equivalence assumption is NOT met with 4 continuous items on scales of X and Y.

NS	$\bar{N}$	$\bar{\rho} = 0$				$\bar{\rho} = .1$				$\bar{\rho} = .3$				$\bar{\rho} = .5$			
		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed	
		Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
		Homogeneous case															
5	30	.002	.961	.003	.955	.008	.956	.011	.962	.025	.956	.038	.948	.044	.944	.032	.941
	50	-.005	.959	-.002	.962	.011	.952	.011	.954	.031	.942	.035	**.923**	.050	**.921**	.053	**.886**
	100	.000	.960	-.003	.964	.014	.955	.009	.956	.030	.943	.038	**.903**	.050	**.898**	.055	**.856**
	200	.002	.955	.000	.959	.010	.950	.015	.953	.027	**.913**	.031	**.912**	.050	**.797**	.052	**.827**
20	30	.000	.960	-.002	.954	.010	.943	.015	.954	.024	.943	.040	**.906**	.044	**.896**	.046	**.857**
	50	.000	.959	-.001	.955	.010	.944	.016	.950	.030	**.914**	.040	**.868**	.050	**.803**	.048	**.798**
	100	.001	.956	.001	.949	.010	.952	.013	.945	.030	**.862**	.036	**.843**	.050	**.649**	.055	**.666**
	200	.000	.947	.000	.961	.009	.947	.010	.938	.029	**.791**	.034	**.740**	.048	**.479**	.052	**.495**
50	30	-.001	.959	-.003	.959	.012	.943	.014	.950	.032	**.886**	.040	**.864**	.039	**.826**	.035	**.835**
	50	.000	.970	-.003	.962	.011	.949	.014	.936	.032	**.806**	.040	**.771**	.049	**.554**	.051	**.596**
	100	-.001	.946	.001	.953	.010	.933	.013	**.921**	.029	**.763**	.036	**.617**	.048	**.277**	.055	**.349**
	200	.000	.965	.001	.964	.009	**.913**	.011	**.903**	.029	**.523**	.033	**.516**	.047	**.038**	.049	**.155**
		Heterogeneous case															
5	30	.001	**.920**	.004	.934	.007	.945	.003	**.917**	.018	.930	.013	.938	.036	**.911**	.018	.928
	50	-.003	**.912**	.010	.929	.012	.928	.022	**.905**	.036	**.920**	.038	**.884**	.051	**.871**	.043	**.872**
	100	.003	**.909**	-.009	**.905**	.008	**.864**	.008	**.909**	.038	**.870**	.040	**.899**	.046	**.849**	.047	**.843**
	200	.001	**.897**	.004	**.881**	.014	**.878**	.008	**.890**	.032	**.864**	.031	**.860**	.052	**.799**	.050	**.819**
20	30	-.001	.932	-.001	.941	.008	.943	.015	.946	.028	**.900**	.033	**.899**	.040	**.880**	.031	**.896**
	50	.002	.929	-.003	.938	.010	.930	.016	.942	.027	**.898**	.036	**.883**	.049	**.787**	.041	**.870**
	100	.002	.940	.001	.938	.009	**.924**	.011	.934	.030	**.872**	.036	**.868**	.049	**.739**	.054	**.765**
	200	.000	.934	-.003	.928	.010	.935	.012	**.909**	.027	**.876**	.033	**.856**	.049	**.684**	.049	**.726**
50	30	-.003	.936	-.004	.940	.010	.954	.011	.946	.027	**.877**	.032	**.869**	.036	**.835**	.030	**.865**
	50	-.002	.947	-.001	.956	.008	**.924**	.017	**.919**	.031	**.855**	.039	**.822**	.049	**.642**	.046	**.723**
	100	.000	.932	.000	.953	.008	.934	.011	.935	.030	**.823**	.035	**.803**	.050	**.542**	.051	**.589**
	200	.001	.942	-.001	.944	.009	.925	.011	**.917**	.028	**.801**	.033	**.780**	.048	**.443**	.049	**.477**

Note. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 6: Meta-analysis with correction of measurement error using the r-based approach: mean correlation estimates and coverage probabilities when the tau-equivalence assumption is NOT met with 12 continuous items on scales of X and Y.

NS	$\bar{N}$	$\bar{\rho} = 0$				$\bar{\rho} = .1$				$\bar{\rho} = .3$				$\bar{\rho} = .5$			
		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed		T-Norm		T-Skewed	
		Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
									Homogeneous case								
5	30	.000	.964	-.005	.969	.002	.955	.009	.957	.010	.952	.021	.932	.015	.949	.019	**.901**
	50	-.002	.959	.001	.968	.001	.962	.011	.968	.005	.960	.021	.929	.010	.960	.014	**.904**
	100	-.001	.963	.003	.962	.002	.960	.007	.946	.006	.969	.014	.931	.011	.946	.015	**.897**
	200	-.001	.966	-.001	.967	.003	.969	.003	.955	.006	.963	.011	.936	.009	.958	.017	**.880**
20	30	.002	.959	.000	.956	.001	.957	.012	.959	.006	.949	.020	**.923**	.010	.951	.020	**.905**
	50	.001	.956	.001	.962	.001	.963	.008	.944	.007	.954	.016	.932	.011	**.924**	.018	**.895**
	100	.002	.953	-.001	.963	.004	.946	.006	.954	.008	.935	.013	**.916**	.012	**.919**	.015	**.890**
	200	.000	.961	.000	.947	.003	.957	.004	.958	.007	.936	.012	**.913**	.011	**.896**	.014	**.879**
50	30	-.001	.956	.001	.956	.002	.964	.009	.949	.006	.954	.025	**.872**	.013	.930	.018	**.882**
	50	-.001	.963	-.001	.963	.003	.949	.007	.942	.006	.947	.017	**.896**	.011	**.920**	.017	**.875**
	100	.001	.965	.000	.956	.003	.946	.006	.938	.007	.935	.012	**.882**	.011	**.877**	.016	**.842**
	200	.000	.952	.000	.965	.002	.955	.003	.947	.007	**.924**	.010	**.871**	.012	**.738**	.014	**.821**
									Heterogeneous case								
5	30	.005	**.896**	-.006	.926	.001	.929	.006	**.916**	.007	**.921**	.020	**.908**	.011	.932	.013	**.903**
	50	-.002	**.916**	.005	**.878**	-.001	**.910**	.008	**.901**	.005	**.894**	.015	**.908**	.012	**.892**	.012	**.892**
	100	-.001	**.880**	-.001	.926	.007	**.877**	.003	**.890**	.011	**.870**	.012	**.899**	.012	**.871**	.017	**.866**
	200	.004	**.867**	-.001	**.890**	.000	**.883**	.006	**.865**	.005	**.898**	.009	**.885**	.012	**.851**	.008	**.881**
20	30	.000	.948	.002	.928	.003	.930	.007	.933	.009	.931	.021	.928	.011	.932	.014	**.915**
	50	.001	.935	.001	.927	.005	.933	.005	**.920**	.007	**.936**	.014	**.918**	.010	**.914**	.014	**.913**
	100	-.001	.930	-.001	.928	-.001	.929	.006	.944	.006	**.916**	.011	**.922**	.010	**.905**	.013	**.906**
	200	.000	.932	.000	.926	.001	.942	.004	.935	.009	**.927**	.011	**.907**	.012	**.903**	.017	**.882**
50	30	-.001	.948	.000	.935	.001	.938	.009	.928	.005	.941	.021	**.916**	.010	**.909**	.017	**.904**
	50	-.002	.939	.001	.946	.004	.936	.006	.930	.004	.926	.016	**.920**	.010	**.913**	.017	**.908**
	100	-.001	.938	-.002	.943	.002	.935	.002	.951	.007	.933	.010	.933	.011	**.898**	.013	**.916**
	200	.000	.949	-.001	.954	.002	.943	.004	.950	.007	**.923**	.008	.933	.011	**.889**	.013	**.915**

Note. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 7: Type I error rate of homogeneity test using the r-based approach when the true scores are skewed.

		$\bar{\rho} = 0$				$\bar{\rho} = .5$			
		$J = 4$		$J = 12$		$J = 4$		$J = 12$	
NS	$\bar{N}$	Tau	NTau	Tau	NTau	Tau	NTau	Tau	NTau
5	30	.028	.046	.054	.062	.070	.049	**.128**	**.110**
	50	.053	.045	.055	.049	**.079**	**.077**	**.236**	**.193**
	100	.052	.059	.048	.052	**.141**	**.113**	**.229**	**.249**
	200	.045	.042	.055	.050	**.170**	**.140**	**.285**	**.280**
20	30	**.023**	**.024**	.048	.050	**.082**	.073	**.436**	**.369**
	50	.044	.031	.065	.061	**.161**	**.137**	**.482**	**.499**
	100	.061	.047	.060	.068	**.297**	**.253**	**.601**	**.591**
	200	.058	.060	.050	.062	**.358**	**.329**	**.628**	**.642**
50	30	**.023**	**.021**	.068	.061	**.107**	.037	**.649**	**.663**
	50	.045	.031	.058	.058	**.319**	**.169**	**.773**	**.800**
	100	.053	.038	.063	.059	**.481**	**.451**	**.908**	**.906**
	200	.053	.057	.062	.064	**.598**	**.604**	**.918**	**.905**

Note. Type I error rate not in the range of [.025, .075] is highlighted. Tau: tau-equivalence is met; NTau: tau-equivalence is not met.

Table 8: Bias and coverage probability for the mean effect size based on scale scores of coarsely categorized items when the number of categories is 2, $J = 12$, $\bar{N} = 200$ and $NS = 50$.

		Tau equivalent						Non-tau equivalent					
		Homogeneous case											
Dist of Items	$\bar{\rho}$	r-based approach			z-based approach			r-based approach			z-based approach		
		T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed
		Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias
All Sym	0	.000	.964	.000	.000	.954	.000	.947	.940	.000	.972	.954	.000	.962	.944	.000	.946	.000	.944
	.1	-.002	.953	.025	-.002	.947	.026	.947	**.571**	-.006	.933	**.220**	-.005	**.919**	**.210**	.043	**.210**	.043	**.210**
	.3	-.008	**.913**	.049	-.007	**.905**	.050	**.049**	**.358**	-.017	**.724**	**.001**	-.016	**.737**	**.001**	.078	**.001**	.078	**.001**
	.5	-.010	**.893**	.029	-.008	**.889**	.031	**.358**	**.068**	-.023	**.414**	**.067**	-.021	**.477**	**.068**	.048	**.068**	.048	**.068**
All RSke	0	.000	.971	-.001	.000	.959	-.001	.943	**.758**	.000	.965	**.749**	-.001	.959	**.616**	-.001	.947	-.001	.947
	.1	-.011	**.883**	-.018	-.011	**.871**	-.017	**.758**	**.625**	-.018	**.749**	**.616**	-.022	**.616**	**.625**	-.017	**.751**	-.022	**.625**
	.3	-.028	**.445**	-.024	-.027	**.481**	-.023	**.606**	**.628**	-.043	**.122**	**.581**	-.024	**.581**	**.628**	-.041	**.176**	-.022	**.628**
	.5	-.034	**.265**	-.018	-.032	**.326**	-.017	**.730**	**.744**	-.051	**.033**	**.677**	-.019	**.677**	**.744**	-.050	**.039**	-.017	**.728**
Sym+RSke	0	-.001	.942	.000	-.001	.932	.000	.942	**.909**	.001	.972	**.875**	.000	.953	**.746**	.001	.961	.000	.946
	.1	-.003	.939	.006	-.003	.929	.007	**.909**	**.643**	-.013	**.875**	**.334**	-.019	**.746**	**.762**	-.012	**.860**	-.018	**.762**
	.3	-.007	.929	.021	-.006	**.924**	.022	**.643**	**.799**	-.033	**.334**	**.705**	-.019	**.705**	**.744**	-.031	**.379**	-.018	**.744**
	.5	-.008	**.892**	.015	-.006	**.907**	.016	**.799**	**.815**	-.040	**.136**	**.770**	-.017	**.770**	**.815**	-.038	**.184**	-.015	**.815**
LSke+RSke	0	.000	.969	.000	.000	.948	.000	.955	**.792**	.000	.950	**.851**	.000	.954	**.008**	.000	.931	.000	.948
	.1	.002	.964	.017	.002	.949	.018	**.792**	**.470**	-.012	**.851**	**.406**	.082	**.008**	**.023**	-.012	**.834**	.084	**.023**
	.3	.004	.933	.036	.005	**.915**	.038	**.269**	**.186**	-.031	**.406**	**.186**	.112	**.000**	**.000**	-.030	**.449**	.116	**.000**
	.5	.006	.930	.029	.008	**.884**	.031	**.470**	**.051**	-.038	**.186**	**.051**	.063	**.013**	**.051**	-.036	**.262**	.066	**.051**
		Heterogeneous case																	
All Sym	0	.000	.945	.000	.000	.946	.000	.946	**.675**	.000	.937	**.886**	.005	.926	**.824**	.000	.940	.005	.931
	.1	-.002	.942	.017	-.002	.941	.017	**.917**	**.745**	-.006	.950	**.828**	.033	**.824**	**.824**	-.006	.950	.032	**.824**
	.3	-.007	.937	.038	-.006	.932	.038	**.675**	**.922**	-.015	**.886**	**.828**	.060	**.341**	**.358**	-.014	**.890**	.059	**.358**
	.5	-.011	**.917**	.025	-.010	**.922**	.026	**.745**	**.870**	-.021	**.828**	**.379**	.041	**.379**	**.403**	-.020	**.846**	.042	**.403**
All RSke	0	.001	**.923**	.011	.002	**.924**	.012	**.919**	**.910**	.004	.934	**.912**	.015	**.912**	**.915**	.005	.930	.015	**.915**
	.1	-.009	.926	-.005	-.009	.930	-.004	.951	**.865**	-.013	**.910**	**.623**	-.006	.934	**.860**	-.012	**.906**	-.005	.931
	.3	-.025	**.818**	-.020	-.024	**.826**	-.019	**.865**	**.726**	-.039	**.623**	**.384**	-.021	**.860**	**.873**	-.037	**.634**	-.020	**.873**
	.5	-.031	**.695**	-.018	-.029	**.726**	-.017	**.870**	**.867**	-.049	**.384**	**.867**	-.020	**.859**	**.867**	-.047	**.409**	-.019	**.867**
Sym+RSke	0	.000	.950	.006	.000	.954	.007	.943	**.921**	.003	.941	**.891**	.013	**.891**	**.889**	.003	.936	.014	**.889**
	.1	-.002	.948	.008	-.002	.948	.009	**.921**	**.907**	-.011	.925	**.761**	-.005	.940	.949	-.010	.926	-.004	.949
	.3	-.008	.938	.016	-.007	.937	.016	**.907**	**.916**	-.029	**.761**	**.591**	-.018	**.881**	**.889**	-.028	**.767**	-.017	**.889**
	.5	-.008	.936	.012	-.006	.937	.012	**.916**	**.862**	-.038	**.591**	**.862**	-.020	**.854**	**.862**	-.036	**.627**	-.019	**.862**
LSke+RSke	0	.001	.941	.000	.001	.936	.000	.943	**.917**	.003	.932	**.909**	-.022	**.909**	**.909**	.004	.935	-.022	**.909**
	.1	.003	.942	.014	.003	.944	.014	**.917**	**.788**	-.010	**.913**	**.801**	.040	**.801**	**.823**	-.009	**.912**	.036	**.823**
	.3	.004	.927	.029	.005	.929	.029	**.788**	**.758**	-.027	**.755**	**.113**	.093	**.113**	**.154**	-.027	**.779**	.091	**.154**
	.5	.007	**.915**	.026	.008	**.915**	.026	**.758**	**.247**	-.036	**.594**	**.165**	.060	**.165**	**.247**	-.035	**.623**	.060	**.247**

Note. Dist of Items: Distribution of items; All Sym: All symmetric; All RSke: All right skewed; Sym+RSke: Half of the items are symmetric, and the other half are right skewed; LSym+RSke: Half of the items are left skewed, and the other are right skewed. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 9: Bias and coverage probability for the mean effect size based on scale scores of coarsely categorized items when the number of categories is 5, $J = 12$, $\bar{N} = 200$ and $NS = 50$.

		Tau equivalent						Non-tau equivalent					
		Homogeneous case											
Dist of Items	$\bar{\rho}$	r-based approach			z-based approach			r-based approach			z-based approach		
		T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed
		Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias
All Sym	0	.001	.949	.000	.001	.936	.000	.001	.958	.000	.001	.948	.000
	.1	.000	.961	.015	.000	.951	.015	.001	.965	.033	.001	.953	.034
	.3	-.001	.954	.030	.000	.929	.031	.005	.934	.059	.006	**.907**	.060
	.5	-.001	.949	.020	.001	.928	.022	.009	**.858**	.043	.011	**.794**	.044
All RSke	0	.000	.954	.000	.000	.946	.000	.000	.949	.001	.000	.936	.001
	.1	-.004	.941	-.001	-.004	.934	-.001	-.003	.942	.014	-.003	.934	.014
	.3	-.009	**.870**	.006	-.008	**.876**	.007	-.008	**.916**	.035	**.195**	-.006	**.921**
	.5	-.010	**.824**	.004	-.009	**.838**	.005	-.006	.925	.029	**.307**	-.005	.935
Sym+RSke	0	.000	.957	.000	.000	.946	.000	.000	.953	.000	.962	.000	.948
	.1	-.001	.949	.006	-.001	.937	.006	.002	.948	.030	**.310**	.003	**.924**
	.3	-.002	.950	.017	-.001	.945	.018	.004	.945	.054	**.010**	.005	.929
	.5	-.003	.948	.013	-.001	.939	.014	.008	**.884**	.039	**.094**	.010	**.810**
LSke+RSke	0	.000	.954	.001	.000	.937	.001	.000	.959	.000	.951	.000	.953
	.1	.000	.959	.019	.000	.946	.019	-.002	.961	.058	**.020**	-.001	.940
	.3	.000	.950	.040	.001	.938	.041	-.004	.949	.092	**.000**	-.003	.939
	.5	.000	.960	.028	.001	.948	.030	-.001	.954	.057	**.000**	.000	.939
		Heterogeneous case											
All Sym	0	.000	.948	-.001	.000	.946	-.001	.001	.953	-.004	.942	.001	.952
	.1	.000	.951	.011	.001	.954	.011	.002	.948	.022	**.880**	.003	.950
	.3	-.001	.938	.025	.000	.940	.025	.004	.933	.049	**.477**	.005	.928
	.5	.000	.953	.018	.001	.952	.019	.009	**.907**	.038	**.472**	.011	**.901**
All RSke	0	.001	.939	.006	.001	.939	.007	.001	.929	.010	**.920**	.001	.933
	.1	-.003	.942	.003	-.003	.942	.004	-.001	.944	.016	**.908**	-.001	.946
	.3	-.009	.943	.003	-.008	.946	.004	-.006	.932	.029	**.759**	-.005	.930
	.5	-.010	.925	.002	-.009	.925	.003	-.006	.948	.026	**.712**	-.004	.951
Sym+RSke	0	.000	.952	.004	.000	.955	.005	.000	.949	-.004	.945	.000	.946
	.1	-.002	.946	.006	-.002	.948	.007	.000	.947	.019	**.885**	.000	.946
	.3	-.003	.947	.014	-.002	.946	.014	.004	.931	.044	**.556**	.005	.932
	.5	-.003	.946	.010	-.002	.951	.011	.007	**.924**	.034	**.562**	.008	**.924**
LSke+RSke	0	.001	.942	.000	.001	.943	.000	.001	.954	-.008	.944	.001	.955
	.1	.000	.939	.014	.000	.938	.014	-.001	.949	.033	**.813**	.000	.953
	.3	.000	.943	.031	.001	.943	.031	-.003	.935	.075	**.169**	-.002	.936
	.5	.000	.944	.024	.001	.942	.025	-.001	.953	.051	**.177**	.000	.951

		Tau equivalent						Non-tau equivalent					
		Homogeneous case											
Dist of Items	$\bar{\rho}$	r-based approach			z-based approach			r-based approach			z-based approach		
		T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed
		Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias
All Sym	0	.000	.948	-.001	.000	.946	-.001	.001	.953	-.004	.942	.001	.952
	.1	.000	.951	.011	.001	.954	.011	.002	.948	.022	**.880**	.003	.950
	.3	-.001	.938	.025	.000	.940	.025	.004	.933	.049	**.477**	.005	.928
	.5	.000	.953	.018	.001	.952	.019	.009	**.907**	.038	**.472**	.011	**.901**
All RSke	0	.001	.939	.006	.001	.939	.007	.001	.929	.010	**.920**	.001	.933
	.1	-.003	.942	.003	-.003	.942	.004	-.001	.944	.016	**.908**	-.001	.946
	.3	-.009	.943	.003	-.008	.946	.004	-.006	.932	.029	**.759**	-.005	.930
	.5	-.010	.925	.002	-.009	.925	.003	-.006	.948	.026	**.712**	-.004	.951
Sym+RSke	0	.000	.952	.004	.000	.955	.005	.000	.949	-.004	.945	.000	.946
	.1	-.002	.946	.006	-.002	.948	.007	.000	.947	.019	**.885**	.000	.946
	.3	-.003	.947	.014	-.002	.946	.014	.004	.931	.044	**.556**	.005	.932
	.5	-.003	.946	.010	-.002	.951	.011	.007	**.924**	.034	**.562**	.008	**.924**
LSke+RSke	0	.001	.942	.000	.001	.943	.000	.001	.954	-.008	.944	.001	.955
	.1	.000	.939	.014	.000	.938	.014	-.001	.949	.033	**.813**	.000	.953
	.3	.000	.943	.031	.001	.943	.031	-.003	.935	.075	**.169**	-.002	.936
	.5	.000	.944	.024	.001	.942	.025	-.001	.953	.051	**.177**	.000	.951

		Tau equivalent						Non-tau equivalent					
		Homogeneous case											
Dist of Items	$\bar{\rho}$	r-based approach			z-based approach			r-based approach			z-based approach		
		T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed	T-Norm		T-Skewed
		Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias	Bias	CP	Bias
All Sym	0	.000	.948	-.001	.000	.946	-.001	.001	.953	-.004	.942	.001	.952
	.1	.000	.951	.011	.001	.954	.011	.002	.948	.022	**.880**	.003	.950
	.3	-.001	.938	.025	.000	.940	.025	.004	.933	.049	**.477**	.005	.928
	.5	.000	.953	.018	.001	.952	.019	.009	**.907**	.038	**.472**	.011	**.901**
All RSke	0	.001	.939	.006	.001	.939	.007	.001	.929	.010	**.920**	.001	.933
	.1	-.003	.942	.003	-.003	.942	.004	-.001	.944	.016	**.908**	-.001	.946
	.3	-.009	.943	.003	-.008	.946	.004	-.006	.932	.029	**.759**	-.005	.930
	.5	-.010	.925	.002	-.009	.925	.003	-.006	.948	.026	**.712**	-.004	.951
Sym+RSke	0	.000	.952	.004	.000	.955	.005	.000	.949	-.004	.945	.000	.946
	.1	-.002	.946	.006	-.002	.948	.007	.000	.947	.019	**.885**	.000	.946
	.3	-.003	.947	.014	-.002	.946	.014	.004	.931	.044	**.556**	.005	.932
	.5	-.003	.946	.010	-.002	.951	.011	.007	**.924**	.034	**.562**	.008	**.924**
LSke+RSke	0	.001	.942	.000	.001	.943	.000	.001	.954	-.008	.944	.001	.955
	.1	.000	.939	.014	.000	.938	.014	-.001	.949	.033	**.813**	.000	.953
	.3	.000	.943	.031	.001	.943	.031	-.003	.935	.075	**.169**	-.002	.936
	.5	.000	.944	.024	.001	.942	.025	-.001	.953	.051	**.177**	.000	.951

Note. Dist of Items: Distribution of items; All Sym: All symmetric; All RSke: All right skewed; Sym+RSke: Half of the items are symmetric, and the other half are right skewed; LSym+RSke: Half of the items are left skewed, and the other are right skewed. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed; CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted.

Table 10: Type I error rate of homogeneity test using the r-based approach with coarsely categorized items when $J = 12$, $\bar{N} = 200$ and NS=50.

		Tau-equivalence				Non tau-equivalence			
		T-Norm		T-Skewed		T-Norm		T-Skewed	
Dist of Items	$\bar{\rho}$	Cat2	Cat5	Cat2	Cat5	Cat2	Cat5	Cat2	Cat5
All Sym	0	.049	.050	.067	.048	.050	.044	.042	.048
	.1	.050	.033	.033	.057	.042	.046	.040	.036
	.3	.043	.041	.035	**.133**	.047	.037	.043	.075
	.5	.026	**.013**	.059	**.327**	.048	**.022**	.056	**.224**
All RSke	0	.059	.061	.049	.056	.049	.042	.049	.050
	.1	**.087**	.054	**.116**	.069	**.095**	.057	**.105**	.047
	.3	**.180**	**.079**	**.426**	**.302**	**.211**	**.084**	**.416**	**.191**
	.5	**.245**	.054	**.818**	**.727**	**.328**	**.081**	**.750**	**.507**
Sym+RSke	0	.053	.048	.043	.056	.062	.047	.051	.043
	.1	.051	.061	.060	.050	.070	.051	.071	.039
	.3	.050	.035	**.136**	**.226**	**.139**	.040	**.291**	**.094**
	.5	.037	.029	**.254**	**.537**	**.161**	**.017**	**.530**	**.281**
LSke+RSke	0	.055	.038	.055	.053	.060	.046	.055	.050
	.1	.051	.051	.059	.037	.070	.059	.032	.029
	.3	.027	.037	.075	**.095**	**.131**	.051	.028	.037
	.5	**.021**	.028	**.132**	**.194**	**.169**	.057	.029	.042

Note. Dist of Items: Distribution of items; All Sym: All symmetric; All RSke: All right skewed; Sym+RSke: Half of the items are symmetric, and the other half are right skewed; LSym+RSke: Half of the items are left skewed, and the other are right skewed. T-Norm: Underlying true scores for X and Y are normal; T-Skewed: underlying true scores for X and Y are skewed. Cat2: 2-category items; Cat5: 5-category items. Type I error rate not in the range of [.025, .075] is highlighted.

Table 11: Bias and coverage probability using the z-based approach with correction of sampling error variance when NS = 50, J=12, and true scores are normal.

		Tau-equivalence						Non Tau-equivalence					
	$\bar{N}$	$\bar{\rho}=0$		$\bar{\rho}=.1$		$\bar{\rho}=.3$		$\bar{\rho}=0$		$\bar{\rho}=.1$		$\bar{\rho}=.3$	
		Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP	Bias	CP
		Homogeneous case											
C1	30	.000	.979	-.011	.959	-.036	.821	-.045	.558	.000	.983	-.011	.973	-.030	.883
	50	-.001	.978	-.008	.960	-.020	.866	-.030	.691	-.001	.978	-.005	.964	-.019	.908
	100	.000	.967	-.004	.961	-.011	.902	-.016	.807	.001	.971	-.001	.957	-.003	.953
	200	.000	.950	-.002	.968	-.005	.937	-.008	.886	.000	.953	.000	.962	.002	.968
		Heterogeneous case											
	30	.001	.971	-.019	.909	-.043	.744	-.062	.435	-.001	.957	-.014	.927	-.042	.755
	50	.003	.935	-.012	.923	-.029	.829	-.043	.605	-.001	.939	-.007	.940	-.022	.860
	100	.001	.942	-.004	.938	-.013	.917	-.022	.815	-.001	.944	-.003	.934	-.008	.934
	200	-.001	.950	-.002	.926	-.007	.932	-.013	.880	.000	.948	-.001	.942	-.002	.935
		Homogeneous case											
C2	30	.000	.948	.004	.948	.012	.927	.015	**.907**	-.001	.953	.006	.955	.020	**.904**
	50	-.001	.962	.001	.959	.006	.942	.009	**.920**	-.001	.957	.006	.940	.014	**.921**
	100	.000	.958	.001	.957	.002	.945	.005	.957	.001	.963	.004	.945	.010	**.899**
	200	.000	.945	.001	.963	.002	.952	.003	.949	.000	.952	.003	.950	.009	**.900**
		Heterogeneous case											
	30	.002	.955	.002	.931	.011	.936	.015	**.916**	-.002	.946	.005	.938	.018	**.922**
	50	.003	.938	.002	.944	.006	.930	.008	.938	-.002	.939	.006	.936	.011	**.922**
	100	.001	.937	.002	.935	.004	.940	.004	.932	-.001	.937	.003	.939	.011	.930
	200	-.001	.950	.001	.925	.003	.933	.002	.934	.000	.947	.002	.942	.008	**.919**

		Non Tau-equivalence		
	$\bar{N}$	$\bar{\rho}=.5$		
		Bias	CP	
		Homogeneous case		
C1	30	-.044	.662	
	50	-.021	.860	
	100	-.005	.956	
	200	.005	.926	
		Heterogeneous case		
	30	-.061	.488	
	50	-.033	.733	
	100	-.012	.902	
	200	.000	.937	
		Homogeneous case		
C2	30	.033	**.778**	
	50	.022	**.803**	
	100	.017	**.786**	
	200	.014	**.644**	
		Heterogeneous case		
	30	.029	**.826**	
	50	.021	**.844**	
	100	.016	**.866**	
	200	.014	**.867**	

Note. CP: Coverage probability. Coverage probability not in the range of [.925, .975] is highlighted. C1: The z-based approach using Equation 20 for correcting sampling error variance; C2: The z-based approach using Equation 21 for

Table 12: Type I error rate of homogeneity test using the z-based approach with correction of sampling error variance when NS = 50, J=12, and true scores are normal.

	$\bar{N}$	Tau-equivalence				Non Tau-equivalence			
		$\bar{\rho}=0$	$\bar{\rho}=.1$	$\bar{\rho}=.3$	$\bar{\rho}=.5$	$\bar{\rho}=0$	$\bar{\rho}=.1$	$\bar{\rho}=.3$	$\bar{\rho}=.5$
C1	30	**.000**	**.000**	**.000**	**.004**	**.000**	**.000**	**.000**	**.000**
	50	**.002**	**.000**	**.004**	**.008**	**.002**	**.000**	**.001**	**.004**
	100	**.012**	**.020**	**.011**	**.017**	**.014**	**.014**	**.016**	**.018**
	200	.028	**.024**	.031	**.023**	.025	**.017**	.027	**.016**
C2	30	**.172**	**.185**	**.202**	**.164**	**.217**	**.212**	**.214**	**.200**
	50	**.120**	**.102**	**.096**	**.084**	**.125**	**.118**	**.114**	**.096**
	100	**.081**	**.082**	.059	.048	.065	.071	.065	.044
	200	.058	.062	.059	.033	.056	.061	.053	.030

Note. Type I error rate not in the range of [.025, .075] is highlighted. C1: The z-based approach using Equation 20 for correcting sampling error variance; C2: The z-based approach using Equation 21 for correcting sampling error variance.

Figure Captions

Figure 1. Type I error rate with continuous items using r-based approach when true scores are normal.

Figure 2. Type I error rate with continuous items using z-based approach when NS=5.

Figure 3. Distribution of observed data

Figure (a). Symmetry

Figure (b). Asymmetry

Figure 4. Scale score distribution when the number of categories is 2

Figure (a). All the items are symmetric

Figure (b). All the items are right skewed

Figure (c). Half symmetric and half right skewed

Figure (d). Half right skewed and half left skewed

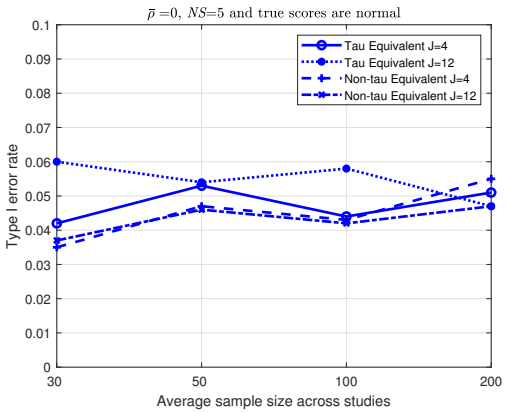
Figure 5. Scale score distribution when the number of categories is 5

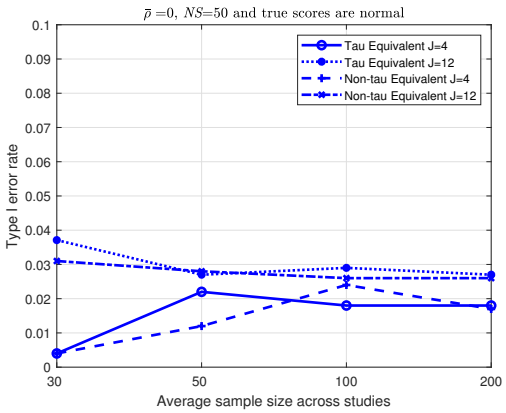
Figure (a). All the items are symmetric

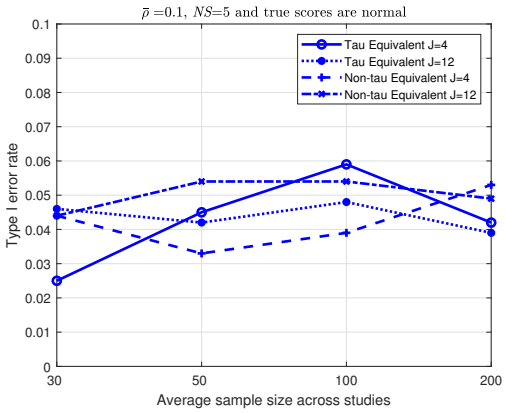
Figure (b). All the items are right skewed

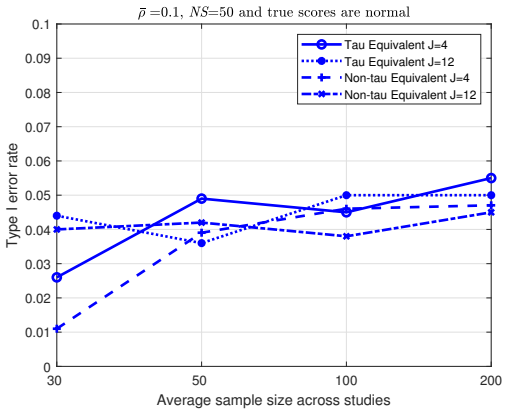
Figure (c). Half symmetric and half right skewed

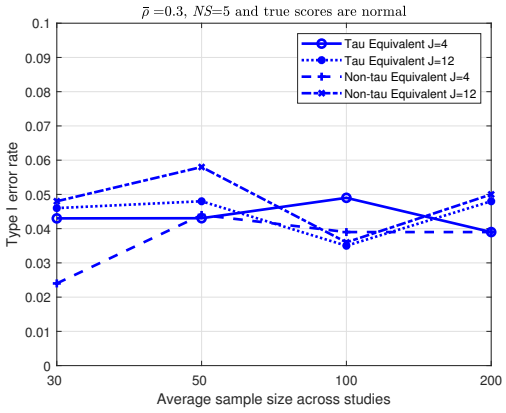
Figure (d). Half right skewed and half left skewed

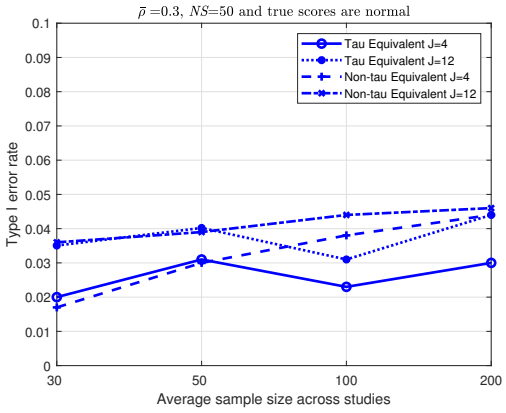

$\bar{\rho}$ =0, NS=5 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

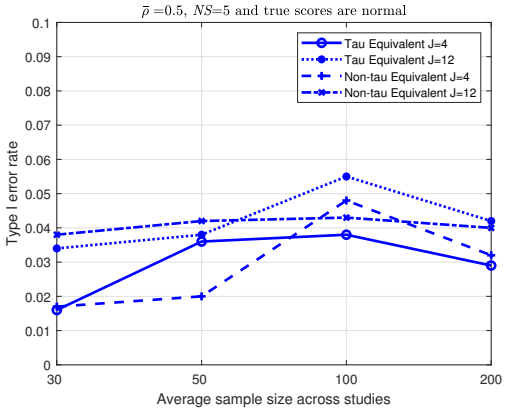

$\bar{\rho}$ =0, NS=50 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

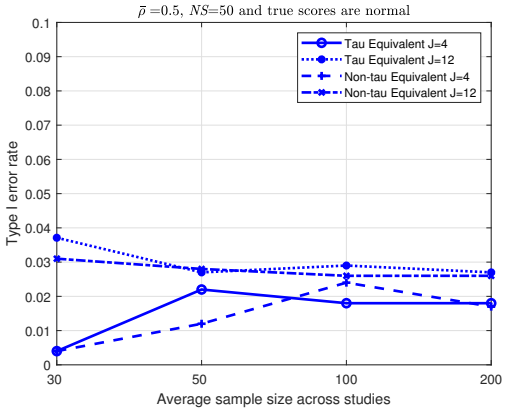

$\bar{\rho}$ =0.1, NS=5 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

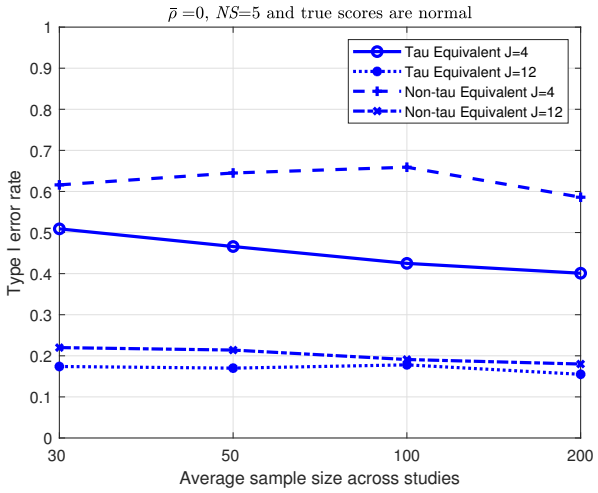

$\bar{\rho}$ =0.1, NS=50 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

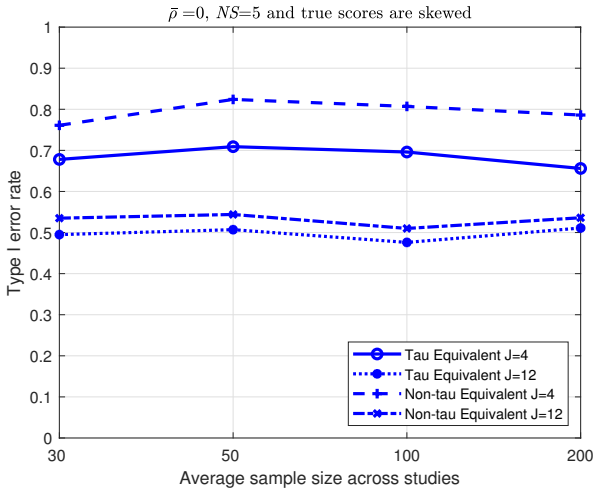

$\bar{\rho}$ =0.3, NS=5 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

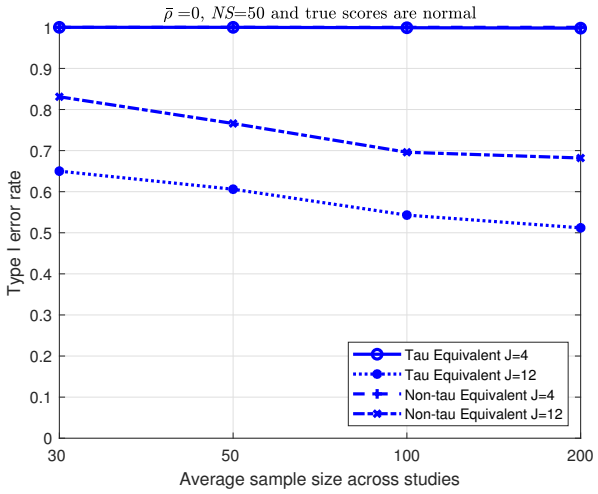

$\bar{\rho}$ =0.3, NS=50 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

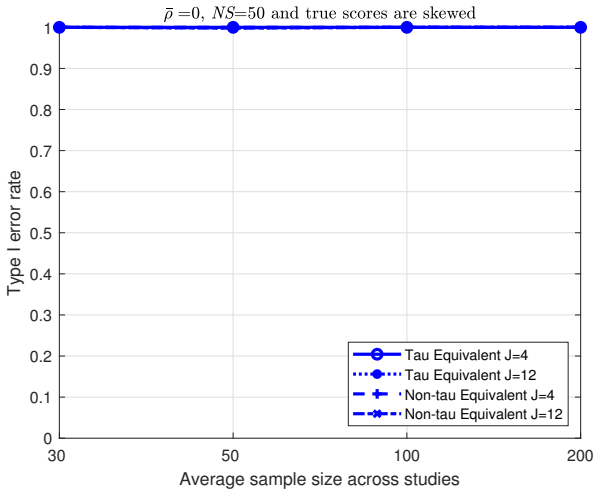

$\bar{\rho}$ =0.5, NS=5 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

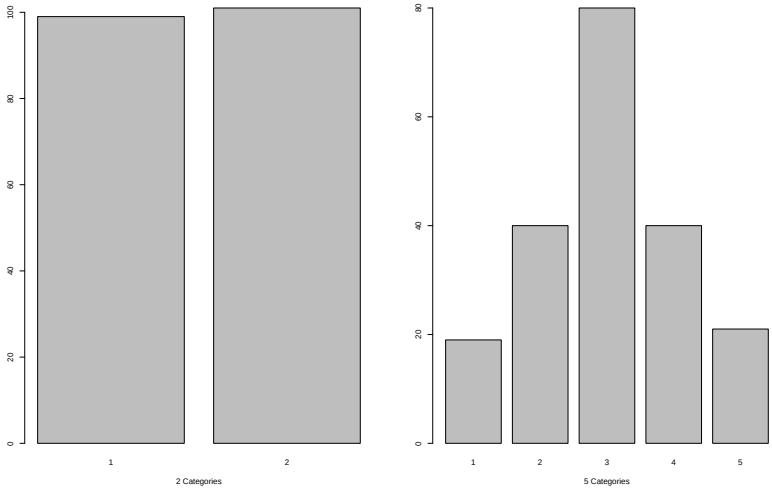

$\bar{\rho}$ =0.5, NS=50 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies


$\bar{\rho}$ =0, NS=5 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

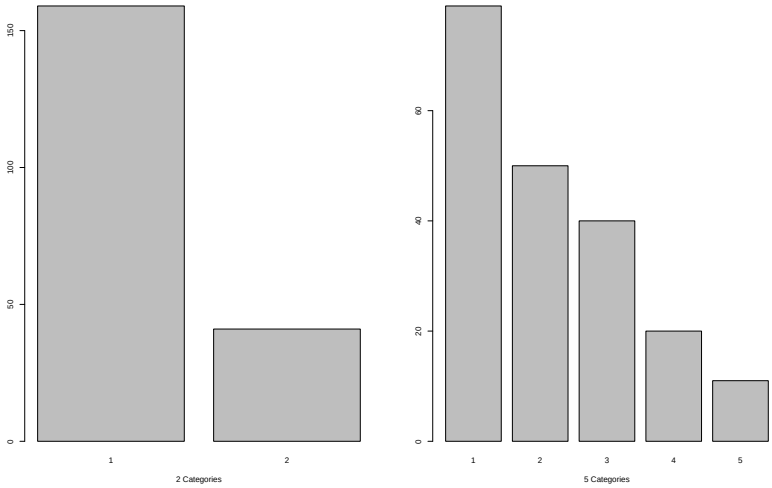

$\bar{\rho}$ =0, NS=5 and true scores are skewed
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies


$\bar{\rho}$ =0, NS=50 and true scores are normal
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies

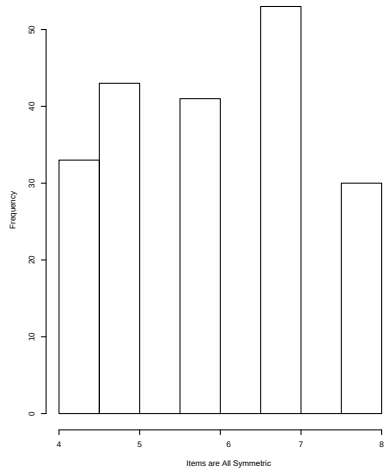

$\bar{\rho}$ =0, NS=50 and true scores are skewed
Tau Equivalent J=4
Tau Equivalent J=12
Non-tau Equivalent J=4
Non-tau Equivalent J=12
Type I error rate
Average sample size across studies



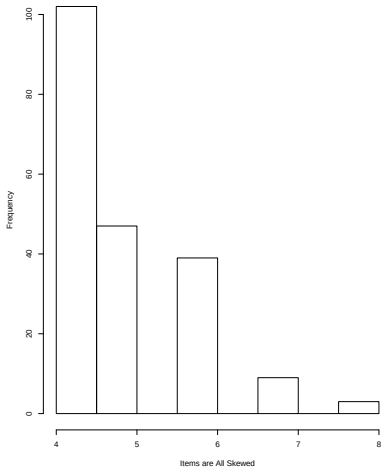
(a) Symmetry



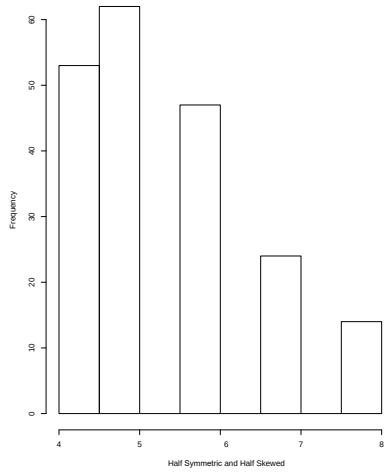
(b) Asymmetry



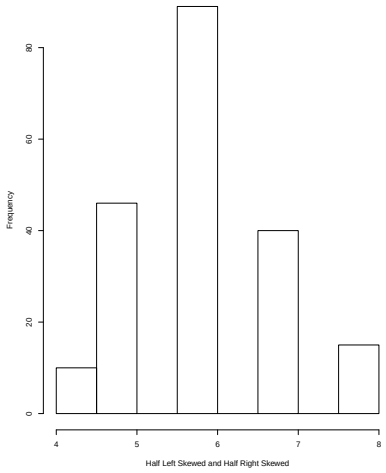
(a) All the items are symmetric



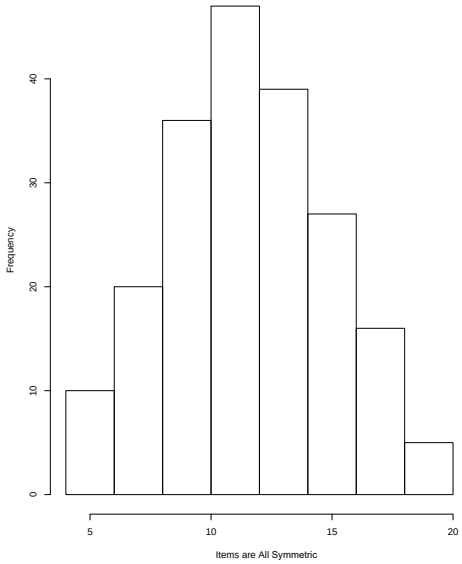
(b) All the items are right skewed



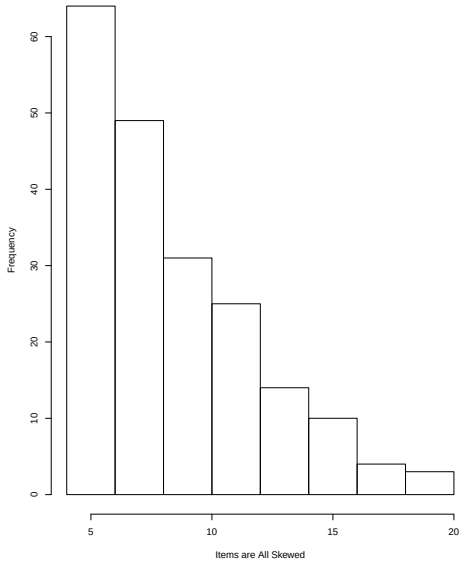
(c) Half symmetric and half right skewed



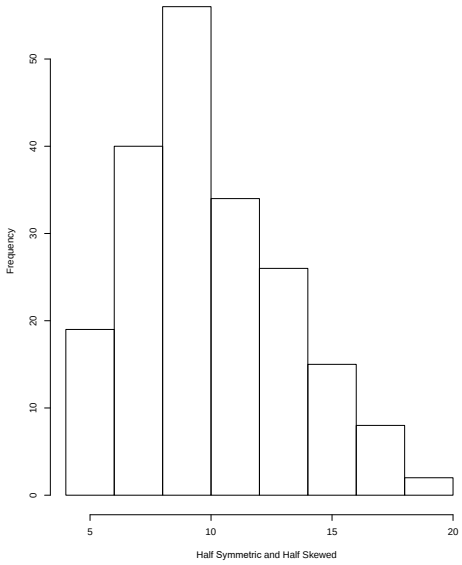
(d) Half right skewed and half left skewed



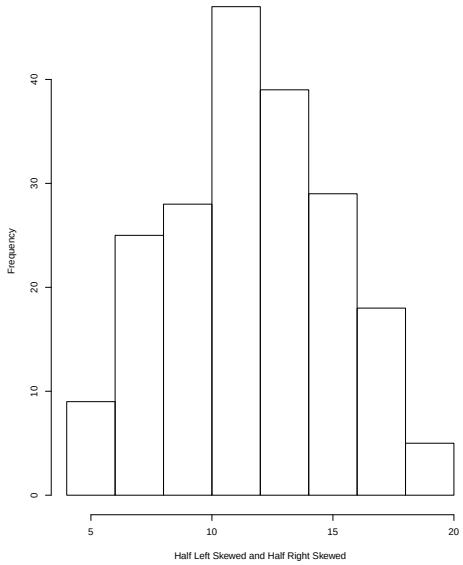
(a) All the items are symmetric



(b) All the items are right skewed



(c) Half symmetric and half right skewed



(d) Half right skewed and half left skewed