MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 4, 711–730

OPEN ACCESS

Correcting for Differences in Measurement Unreliability in Meta-Analysis of Variances

Katrin Jansen and Steffen Nestler

University of Münster, Münster, Germany

ABSTRACT

There is a growing interest of researchers in meta-analytic methods for comparing variances as a means to answer questions on between-group differences in variability. When measurements are fallible, however, the variance of an outcome reflects both the variance of the true scores and the error variance. Consequently, effect sizes based on variances, such as the log variability ratio (lnVR) or the log coefficient of variation ratio (lnCVR), may thus not only reflect between-group differences in the true-score variances but also differences in measurement reliability. In this article, we derive formulas to correct the lnVR and lnCVR and their sampling variances for between-group differences in reliability and evaluate their performance in simulation studies. We find that when the goal is to meta-analyze differences between the true-score variances and reliability differs between groups, our proposed corrections lead to accurate estimates of effect sizes and sampling variances in single studies, accurate estimates of the average effect and the between-study variance in random-effects meta-analysis, and adequate type I error rates for the significance test of the average effect. We discuss how to deal with problems arising from missing or imprecise group-specific reliability estimates in meta-analytic data sets and identify questions for further methodological research.

KEYWORDS

Meta-analysis; variability; reliability; effect sizes for variances

Introduction

Meta-analysis is typically used to summarize the available evidence on the difference between the average outcomes of two (or more) groups. Conducting such a meta-analysis usually entails calculating effect sizes, such as mean differences, standardized mean differences, or ratios of means (Borenstein et al., 2021, Chapter 4) for each study. These effect sizes are then summarized across all available studies to evaluate (1) whether the average outcome of the two groups differs in an average study, and if so, to what extent, and (2) whether there is between-study variability in the differences between the groups' average outcomes, and if so, whether this variability can be explained by certain moderators. Such a *meta-analysis of means* is often used to evaluate the efficacy of a treatment (e.g., of psychotherapy, e.g., see Althobaiti et al., 2020; Cristea et al., 2017; Cuijpers et al., 2011), or to investigate mean differences between naturally occurring groups (e.g., differences in executive functioning in bilingual vs. non-bilingual children, Gunnerud et al., 2020, or gender differences in scholastic achievement, Voyer & Voyer, 2014).

In recent years, the focus in the behavioral sciences and other disciplines has shifted from examining average outcomes to also examining the variability of outcomes. In clinical psychology and medicine, for example, inter-individual differences in treatment effects are investigated by comparing the variances of the treatment and control group outcomes at post-treatment (Imbens & Rubin, 2015; Mills et al., 2021; Salditt et al., 2024). Similarly, a higher variability of a certain characteristic in patients compared to healthy controls may indicate the existence of different patient subtypes that may respond differently to treatment (e.g., Brugger et al., 2020; Osimo et al., 2020). Questions related to between-group differences in variability also arise in research on personality and individual differences, and concern, for example, age or gender differences in the inter-individual variability of personality traits (e.g., Mõttus et al., 2016) or cognitive ability (e.g., Taylor & Barbot, 2021). There are various different approaches to test for

© 2025 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

variability differences in a single study. Comprehensive reviews of these approaches, their data requirements, and assumptions were provided by Mills et al. (2021) and Nestler and Salditt (2024). Because the power of these tests is low unless sample sizes are large (i.e., 250 subjects per group; Nestler & Salditt, 2024), it is of interest to examine variability differences meta-analytically to enhance power. This can be achieved by using effect sizes that are based on variances instead of on means, such as the lnVR and the lnCVR (Nakagawa et al., 2015; Senior et al., 2020). As in a meta-analysis of means, these effect sizes can be employed to test (1) whether groups differ in terms of their variability in an average study and if so, to what extent, and (2), whether group differences in variability vary between studies and if so, whether this between-study variability can be explained by certain moderators. Such a *meta-analysis of variances* was, for example, used to summarize gender differences in the variability of academic grades (O'Dea et al., 2018) or divergent thinking (Abdulla Alabbasi et al., 2025; Taylor et al., 2024), variability differences between patients with schizophrenia and healthy controls in terms of their striatal dopaminergic function (Brugger et al., 2020), and heterogeneity of treatment effects in pharmacological and psychological treatments of depression (Kaiser et al., 2022; Plöderl & Hengartner, 2019), post-traumatic stress disorder (Herzog & Kaiser, 2022), and borderline personality disorder (Kaiser & Herzog, 2023).

In meta-analyses, the effect sizes that are summarized are calculated based on the *observed* values of the outcomes. These values are affected by measurement error and this unreliability can distort meta-analytic results. This was shown in particular for meta-analysis of correlation coefficients, leading to the development of several methods to correct correlation coefficients for unreliability (see, e.g. Ke & Tong, 2023; Raju et al., 1991; Schmidt & Hunter, 2015). Extending this research, we demonstrate here that if the outcome measure on which variability is compared between groups is less reliable in one group than in the other, the estimates of the lnVR and the lnCVR may also be contaminated by differences in measurement error variability. Thus, when researchers are interested in variability differences of the true scores, between-group differences in reliability may lead to erroneous conclusions in a meta-analysis of variances if these reliability differences are not accounted for. We therefore propose corrections for the estimators of the lnVR and the lnCVR and their sampling variances which address this issue, and we evaluate the proposed formulas in two simulation studies.

Our proposed corrections require that group-specific reliability estimates are available for each individual study. This is often not the case in meta-analytic data sets since primary studies fail to report group-specific reliabilities. Even if they do, reliability estimates may be based on sample sizes that are too small to achieve an acceptable precision. We therefore suggest to use representative reliability estimates from external sources, such as large confirmatory factor analysis (CFA) studies when correcting the lnVR or the lnCVR for unreliability. A similar approach was used previously by Ke and Tong (2023) who examined unreliability corrections for correlation coefficients. We elaborate on alternative approaches in the discussion.

The remainder of this article is structured as follows: In the following section, we first describe how the lnVR and the lnCVR and their sampling variances are estimated, second, how estimation may be affected by between-group differences in reliability, and finally, how to correct for these reliability differences. In the subsequent section, we describe two Monte Carlo simulation studies that we conducted to evaluate the proposed corrections. Then, we show an application of our method to an illustrative example. In the final section, we discuss the theoretical and practical implications of our results, and elaborate on avenues for future research.

Meta-analysis of variances: effect sizes and reliability

In the following, we assume that we have observed an outcome in two independent groups, where X_1 is the observed outcome in group 1 and X_2 is the observed outcome in group 2. We further assume that both outcomes are normally distributed with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively. On a population level, we define the lnVR as

$$\text{lnVR} = \ln\left(\frac{\sigma_1}{\sigma_2}\right), \tag{1}$$

and the lnCVR as

$$\text{lnCVR} = \ln\left(\frac{\sigma_1/\mu_1}{\sigma_2/\mu_2}\right). \tag{2}$$

Note that lnCVR can be written as $\text{lnCVR} = \text{lnVR} - \text{lnRR}$, where lnRR is the log response ratio (i.e., $\text{lnRR} = \ln(\mu_1/\mu_2)$). Nakagawa et al. (2015) proposed to estimate the lnVR by

$$\widehat{\text{lnVR}} = \ln\left(\frac{s_1}{s_2}\right) + \frac{1}{2}\left(\frac{1}{n_1 - 1} - \frac{1}{n_2 - 1}\right), \tag{3}$$

where s_1 (s_2) is the observed standard deviation in group 1 (group 2), and n_1 (n_2) are the respective sample sizes. The second term in Equation (3) is a bias correction, whose influence is strongest when the sample sizes in the two groups are small and dissimilar. To estimate lnCVR, Senior et al. (2020) proposed to use

$$\widehat{\ln\mathrm{CVR}} = \ln\left(\frac{s_1/\bar{x}_1}{s_2/\bar{x}_2}\right) + \frac{1}{2}\left(\frac{1}{n_1-1} - \frac{1}{n_2-1}\right) + \frac{1}{2}\left(\frac{s_2^2}{n_2\bar{x}_2^2} - \frac{s_1^2}{n_1\bar{x}_1^2}\right), \tag{4}$$

where $\bar{x}_1$ and $\bar{x}_2$ denote the sample means. The second term of Equation (4) is the bias correction for the lnVR and the third term is a bias correction for the lnRR. The estimators of lnVR and lnRR are asymptotically normally distributed (cf. Hedges et al., 1999; Raudenbush & Bryk, 1987), which follows from applying the Delta method. Since the difference of two normally distributed variables also follows a normal distribution, $\widehat{\ln\mathrm{CVR}}$ is asymptotically normally distributed, too.

By dividing each group's standard deviation by its mean, the lnCVR allows to control for a mean-variance relationship. In particular, it is suited for situations in which variability is expected to increase as the mean increases because it allows to examine whether there are between-group differences in variability beyond those that arise from a between-group difference in means. Such a mean-variance relationship can arise, for example, in variables that are bounded from below. If the variance is unrelated to the mean, the lnVR should be preferred over the lnCVR because the latter would spuriously correct for differences in means. Finally, using the lnCVR requires that the outcome being examined can take only positive values, whereas using the lnVR does not.

For the sampling variances of lnVR and lnCVR, Senior et al. (2020) proposed the estimators

$$s^2(\widehat{\ln\mathrm{VR}}) = \frac{1}{2}\left(\frac{1}{n_1-1} + \frac{1}{(n_1-1)^2} + \frac{1}{n_2-1} + \frac{1}{(n_2-1)^2}\right) = \frac{1}{2}\left(\frac{n_1}{(n_1-1)^2} + \frac{n_2}{(n_2-1)^2}\right) \tag{5}$$

and

$$s^2(\widehat{\ln\mathrm{CVR}}) = \frac{s_1^2}{n_1\bar{x}_1^2} + \frac{s_1^4}{2n_1^2\bar{x}_1^4} + \frac{n_1}{2(n_1-1)^2} + \frac{s_2^2}{n_2\bar{x}_2^2} + \frac{s_2^4}{2n_2^2\bar{x}_2^4} + \frac{n_2}{2(n_2-1)^2}, \tag{6}$$

both of which are based on a second-order Taylor expansion and were found to be more accurate than estimators based on the first-order Taylor expansion (Senior et al., 2020).

Consequences of ignoring unreliability

When defining the estimators for the lnVR and the lnCVR in Equations (3) and (4), it is implicitly assumed that X_1 and X_2 are not affected by measurement error. However, this assumption is unlikely to hold in practice (Carroll et al., 2006; Hernan & Robins, 2024; McDonald, 1999). To examine how measurement error affects lnVR and lnCVR, we presume an additive measurement error model (Carroll et al., 2006; Lord & Novick, 1968) for the values in group 1,

$$X_1 = T_1 + \varepsilon_1, \tag{7}$$

where T_1 are the true scores and ε_1 the error terms. In group 2, the values are given by

$$X_2 = T_2 + \varepsilon_2, \tag{8}$$

where T_2 and ε_2 are the true scores and the error terms, respectively. From these definitions, it follows that

$$\mathrm{Var}(X_1) = \sigma_1^2 + \sigma_{\varepsilon_1}^2 = \frac{\sigma_1^2}{\sigma_1^2/(\sigma_1^2+\sigma_{\varepsilon_1}^2)} = \frac{\sigma_1^2}{\mathrm{Rel}(X_1)}, \tag{9}$$

and

$$\mathrm{Var}(X_2) = \sigma_2^2 + \sigma_{\varepsilon_2}^2 = \frac{\sigma_2^2}{\sigma_2^2/(\sigma_2^2+\sigma_{\varepsilon_2}^2)} = \frac{\sigma_2^2}{\mathrm{Rel}(X_2)}. \tag{10}$$

Here, σ_1^2 (σ_2^2) is the variance of the true scores in group 1 (group 2), $\sigma_{\varepsilon_1}^2$ ($\sigma_{\varepsilon_2}^2$) denotes the variance of the error terms, and $\mathrm{Rel}(X_1)$ ($\mathrm{Rel}(X_2)$) is the reliability of the observed variable in group 1 (group 2).

Equations (9) and (10) imply that when researchers are interested in lnVR and lnCVR of the true-score variables, as they typically are when they conduct a meta-analysis of variances, the effect size should be based on the standard deviations of the true-score variables $\sigma_1 = \mathrm{SD}(X_1)\sqrt{\mathrm{Rel}(X_1)}$ and $\sigma_2 = \mathrm{SD}(X_2)\sqrt{\mathrm{Rel}(X_2)}$. Thus, the lnVR of the true-score variables is

$$\begin{aligned}\ln\mathrm{VR}_T &= \ln\left(\frac{\mathrm{SD}(X_1)\cdot\sqrt{\mathrm{Rel}(X_1)}}{\mathrm{SD}(X_2)\cdot\sqrt{\mathrm{Rel}(X_2)}}\right)\\ &= \ln\left(\frac{\mathrm{SD}(X_1)}{\mathrm{SD}(X_2)}\right) + \ln\left(\sqrt{\frac{\mathrm{Rel}(X_1)}{\mathrm{Rel}(X_2)}}\right),\end{aligned} \tag{11}$$

where for the sake of uniqueness, we define the solution of the square root to be positive. Similarly, since $\ln\text{CVR} = \ln\text{VR} - \ln\text{RR}$, the lnCVR for the true-score variables is

$$\ln\text{CVR}_T = \ln\left(\frac{\text{SD}(X_1)}{\text{SD}(X_2)}\right) + \ln\left(\sqrt{\frac{\text{Rel}(X_1)}{\text{Rel}(X_2)}}\right) - \ln\left(\frac{\mu_1}{\mu_2}\right), \quad (12)$$

where we used that the lnRR is unaffected by unreliability because the expected value of the observed variable equals that of the true-score variable.

Between-group differences in reliability

Equations (11) and (12) show that when $\text{Rel}(X_1) = \text{Rel}(X_2)$, lnVR and lnCVR of the true-score variables equal lnVR and lnCVR, respectively, of the observed variables (cf. Equations (1) and (2)). However, when $\text{Rel}(X_1) > \text{Rel}(X_2)$, the effect sizes based on the true-score variables are larger than those based on the observed variables, and when $\text{Rel}(X_1) < \text{Rel}(X_2)$, the effect sizes based on the true-score variables are smaller than those based on the observed variables. Thus, conclusions on between-group differences in variability may be erroneous when the lnVR or the lnCVR are calculated based on the observed outcomes, at least when the two groups *differ* in their reliability.

How likely is it that there are between-group differences in reliability? Research on the psychometric properties of several psychological scales revealed that the internal consistency of the scales varies considerably (Aslan et al., 2022; Badenes-Ribera et al., 2023; Cabedo-Peris et al., 2021; Cerri et al., 2023; Demir et al., 2024; Esparza-Reig et al., 2021; Gisbert-Pérez et al., 2022; Yin & Fan, 2000) and that part of this heterogeneity can be attributed to differences in sample characteristics, such as age (Aslan et al., 2022; Bru-Luna et al., 2021). Furthermore, in the clinical setting, research on the Beck Depression Inventory-II (BDI-II) showed that internal consistency is larger for patients under remission as compared to acutely depressive patients (Kühner et al., 2007) and that the internal consistency of the BDI-II at admission to a psychiatric hospital and at discharge is also different (Keller et al., 2022).[1] In summary, then, between-group differences in reliability may exist in some contexts in which meta-analyses of variances are carried out, and these differences should be taken into account to reach valid conclusions.

Correcting for between-group differences in reliability

Equations (11) and (12) imply that a simple way to correct the effect sizes for between-group differences in reliability is to add $\ln\left(\sqrt{\text{Rel}(X_1)/\text{Rel}(X_2)}\right)$ to the formulas to compute lnVR and lnCVR, respectively, and replace the population reliabilities by their estimated counterparts. Specifically, the corrected estimating equations are

$$\widehat{\ln\text{VR}}_T = \ln\left(\frac{s_1}{s_2}\right) + \ln\left(\sqrt{\frac{\widehat{\text{Rel}}(X_1)}{\widehat{\text{Rel}}(X_2)}}\right) + \frac{1}{2}\left(\frac{1}{n_1 - 1} - \frac{1}{n_2 - 1}\right). \quad (13)$$

and

$$\widehat{\ln\text{CVR}}_T = \ln\left(\frac{s_1/\bar{x}_1}{s_2/\bar{x}_2}\right) + \ln\left(\sqrt{\frac{\widehat{\text{Rel}}(X_1)}{\widehat{\text{Rel}}(X_2)}}\right) + \frac{1}{2}\left(\frac{1}{n_1 - 1} - \frac{1}{n_2 - 1}\right) + \frac{1}{2}\left(\frac{s_2^2}{n_2\bar{x}_2^2} - \frac{s_1^2}{n_1\bar{x}_1^2}\right), \quad (14)$$

where $\widehat{\text{Rel}}(X_1)$ and $\widehat{\text{Rel}}(X_2)$ denote estimators of the reliabilities of X_1 and X_2, respectively. In addition, the estimators of the sampling variances have to be adapted by accounting for the additional uncertainty that arises from the estimation of $\ln\left(\sqrt{\widehat{\text{Rel}}(X_1)/\widehat{\text{Rel}}(X_2)}\right)$. Using the Delta method, we obtain

$$\text{Var}\left(\ln\sqrt{\frac{\widehat{\text{Rel}}(X_1)}{\widehat{\text{Rel}}(X_2)}}\right) \approx \frac{\text{Var}(\widehat{\text{Rel}}(X_1))}{4n_{1,R}\widehat{\text{Rel}}(X_1)^2} + \frac{\text{Var}(\widehat{\text{Rel}}(X_2))}{4n_{2,R}\widehat{\text{Rel}}(X_2)^2}, \quad (15)$$

where $n_{1,R}$ and $n_{2,R}$ are the sample sizes that the reliability estimates in group 1 and group 2 are based on, respectively. In applications, the samples that are used to calculate the group-specific reliability coefficients and their standard errors may differ from those which are used to obtain the effect size estimates that will be summarized in the meta-analysis because if a study

[1]Here, we use internal consistency as a term for reliability, which is consistent with the use of this term in the cited papers. However, strictly speaking, internal consistency (e.g., Cronbach's alpha) is only a valid measure of reliability when the assumptions of an essentially tau-equivalent measurement model hold in the sample.

fails to report reliability estimates, it can be necessary to impute them from external sources (see below). Therefore, the sample sizes in Equation (15) may differ from those in the preceding equations. The sampling variances of the two groups' reliabilities in Equation (15) can be estimated based on the Delta method (for a derivation of a sampling variance estimator for Cronbach's alpha, see van Zyl et al., 2000).

Taking into account the variance of $\ln\left(\sqrt{\widehat{\mathrm{Rel}}(X_1)/\widehat{\mathrm{Rel}}(X_2)}\right)$, the sampling variances of the corrected $\widehat{\ln\mathrm{VR}}_T$ and $\widehat{\ln\mathrm{CVR}}_T$ can be estimated with

$$s^2(\widehat{\ln\mathrm{VR}}_T) = \frac{1}{2}\left(\frac{n_1}{(n_1-1)^2}+\frac{n_2}{(n_2-1)^2}\right) + \frac{\mathrm{Var}(\widehat{\mathrm{Rel}}(X_1))}{4n_{1,R}\widehat{\mathrm{Rel}}(X_1)^2}+\frac{\mathrm{Var}(\widehat{\mathrm{Rel}}(X_2))}{4n_{2,R}\widehat{\mathrm{Rel}}(X_2)^2} \quad (16)$$

and

$$s^2(\widehat{\ln\mathrm{CVR}}_T) = \frac{s_1^2}{n_1\bar{x}_1^2}+\frac{s_1^4}{2n_1^2\bar{x}_1^4}+\frac{n_1}{2(n_1-1)^2}+\frac{s_2^2}{n_2\bar{x}_2^2} + \frac{s_2^4}{2n_2^2\bar{x}_2^4}+\frac{n_2}{2(n_2-1)^2}+\frac{\mathrm{Var}(\widehat{\mathrm{Rel}}(X_1))}{4n_{1,R}\widehat{\mathrm{Rel}}(X_1)^2} + \frac{\mathrm{Var}(\widehat{\mathrm{Rel}}(X_2))}{4n_{2,R}\widehat{\mathrm{Rel}}(X_2)^2}. \quad (17)$$

Obtaining estimates of the group-specific reliability coefficients and dealing with missing or imprecise values

The application of Equations (13) and (14) requires that estimates of the reliability and their standard errors are available for both groups. All possible reliability coefficients can be included in the formulas, such as internal consistency coefficients (e.g., Cronbach's alpha or ω; see McDonald, 1999), test-retest reliability coefficients, or parallel test reliability coefficients (Lord & Novick, 1968). The choice of the type of reliability coefficient should depend on the psychometric properties of the scale. This is particularly relevant because different reliability coefficients make different assumptions regarding the measurement model, and if these assumptions do not hold, reliability estimates can be biased (Graham, 2006; Green & Yang, 2009). We are not aware of any meta-analyses of variances that have gathered information on reliability from primary studies. However, research on corrections for unreliability in meta-analysis of correlations and standardized mean differences showed that studies most often report Cronbach's alpha (Wiernik & Dahlke, 2020; Zhang, 2024), which requires that an essentially tau-equivalent measurement model holds. If primary studies report group-specific correlation (or covariance) matrices at the item level, these can be used to compute reliability estimates using a CFA approach.

In a meta-analytic context, estimates of the group-specific reliability coefficients and their standard errors would have to be obtained for each individual study. This will often not be possible, as group-specific reliability coefficients or correlation matrices are seldom reported in primary studies. The meta-analytic estimate of the pooled lnVR corrected for between-group differences in reliability is

$$\hat{\mu}(\ln\mathrm{VR}) = \sum_{i=1}^{k} w_i^* \widehat{\ln\mathrm{VR}}_{Ti}, \quad (18)$$

where k is the number of studies, $w_i^* = w_i/(\sum_{i=1}^{k} w_i)$ with $w_i = 1/(s^2(\widehat{\ln\mathrm{VR}}_{Ti}) + \hat{\tau}^2)$ (assuming we conduct a random-effects meta-analysis where $\hat{\tau}^2$ is an estimate of the between-study variance) and $\widehat{\ln\mathrm{VR}}_{Ti}$ is the estimated (and corrected) lnVR obtained from the ith study. Based on Equation (13), we can rewrite formula (18) as

$$\hat{\mu}(\ln\mathrm{VR}) = \sum_{i=1}^{k} w_i^* \widehat{\ln\mathrm{VR}}_i + \sum_{i=1}^{k} w_i^* \ln\left(\sqrt{\widehat{\mathrm{Rel}}_i(X_1)/\widehat{\mathrm{Rel}}_i(X_2)}\right), \quad (19)$$

where $\widehat{\ln\mathrm{VR}}_i$ is the *uncorrected* estimate of the lnVR obtained from the ith study. The second part of Equation (19) is a weighted average of the study-specific estimates of the log-square root reliability ratio. If reliability estimates from a considerable number of studies are missing, one option is to replace the study-specific reliability estimates by representative estimates obtained from external sources. If it is not possible to obtain reliability estimates either from the primary studies included in the meta-analysis or from external sources, we suggest to conduct an additional CFA study before conducting the meta-analysis, as was suggested in the context of correcting correlation coefficients for unreliability (see Ke & Tong, 2023).

Even if group-specific reliability estimates are available from all studies, such estimates, and in particular their standard errors, may be imprecise when sample sizes are small (Kline, 2016; Wolf et al., 2013). This

applies irrespective of whether reliability estimates were directly reported or obtained from group-specific correlation matrices, and can distort the estimation of the average effect. Furthermore, the standard error of the average effect, which is estimated by $1/\sqrt{\sum_{i=1}^{k} w_i}$, depends not only on the precision of the study-specific estimates of lnVR but also on the precision of the log-square root reliability ratios (cf. Equation (16)). Therefore, relying on imprecise reliability estimates will reduce the precision of the average effect and result in a low power. We therefore suggest to give preference to reliability estimates from external sources (that were obtained in larger samples) in this situation, too. These considerations apply analogously to meta-analysis of the lnCVR.

Monte Carlo simulations

We conducted two simulation studies to investigate the performance of the suggested corrections, because, to the best of our knowledge, their performance has not yet been examined. In the first simulation, we evaluated the performance of our formulas with respect to the estimation of the study-specific effect size and sampling variance in a *single* study. This pre-study was conducted to complement the findings by Senior et al. (2020) on the performance of different estimators for the uncorrected lnCVR and its sampling variance in a setting with perfect reliability. In the second, main simulation, we examined the performance of the estimators for the corrected lnVR and lnCVR in a *meta-analytic* setting.

Both simulation studies were conducted in R (R Core Team, 2022) using the `metafor` package (Viechtbauer, 2010) for effect size calculation and for conducting the meta-analyses. Note that the effect size and sampling variance estimators that are implemented in `metafor` are based on a first-order Taylor expansion as suggested by Nakagawa et al. (2015). Since Senior et al. (2020) recommended to use the estimators based on the second-order Taylor expansion, we added the respective terms to the estimates obtained from `metafor` where necessary. In both simulation studies, group-specific reliability estimates along with their standard errors were obtained from an external CFA rather than from the individual primary studies. The CFA was conducted using the `lavaan` package (Rosseel, 2012). The first simulation was run on the high-performance computing cluster PALMA II (https://www.uni-muenster.de/ZIV/Technik/Server/HPC.html) at the University of Münster.

Pre-study: single-study setting

In our pre-study, we compared the performance of the corrected estimators for $\ln VR_T$ and $\ln CVR_T$ to that of the uncorrected estimators. In our simulation, reliability estimates were not obtained from the simulated primary studies themselves, but from an external CFA study that was simulated in addition to the primary study data. This procedure was chosen to mirror the realistic scenario that group-specific reliability estimates will often be unavailable in primary studies or that primary studies are often too small to estimate group-specific reliabilities with sufficient precision. Across simulation conditions, we varied the sample size of the CFA study and evaluated whether larger sample sizes were associated with a better performance.

In addition to the performance of the effect size estimators, we evaluated the performance of the corrected sampling variance estimators. Using Equations (16) and (17) requires that standard errors of the reliability coefficients have been obtained for both groups. However, since these estimates may not always be available, e.g., when using reliability estimates from preexisting CFA studies, we also examined under which conditions it is safe to use the uncorrected sampling variances (see Equations (5) and (6)) together with the corrected effect size estimators.

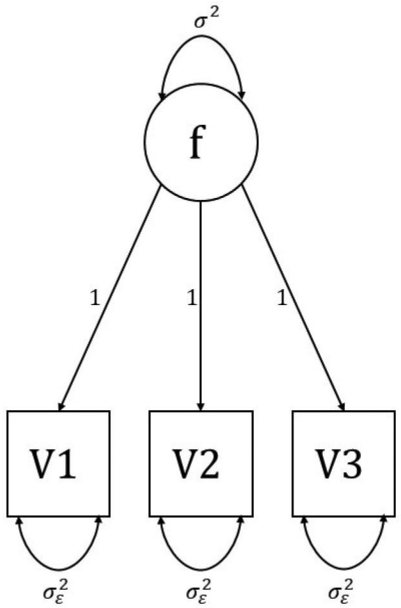


Figure 1. CFA model from which the group-specific data were generated in the simulation study with equal factor loadings of 1 for all three variables V_1–V_3, equal error variances σ^2_ε, and factor variance σ^2.

Table 1. Values of the simulation parameters in the pre-study.

Parameter	Values
$\ln VR_T$	−0.5, −0.25, 0, 0.25, 0.5
lnRR	−0.5, −0.25, 0, 0.25, 0.5
n_2	8, 16, 42
Rel_1	0.7, 0.75, 0.8, 0.85, 0.9

Data generation

In each replication of a simulation condition, the data for the two groups were each simulated from a one-factor model with three items (see Figure 1). Each item had a factor loading of 1 and the error variance σ_ε^2 of each item was set in such a way that the sum score of the three items reached a pre-specified reliability. Specifically, the error variance of a single item can be computed with $\sigma_\varepsilon^2 = \sigma^2 \cdot (1 - \text{Rel}(V))/\text{Rel}(V)$, where $\text{Rel}(V)$ is the reliability of a single item which can be obtained using the Spearman-Brown formula (see Lord & Novick, 1968). For group 2, the reliability of the sum score composed of the three items was set to $\text{Rel}_2 = 0.9$ in all simulation conditions. Also, the mean of the latent variable was set to $\mu_2 = 50$ and the standard deviation to $\sigma_2 = 10$. The values for μ_1 and σ_1 were obtained from $\mu_1 = \mu_2 \cdot \exp(\text{lnRR})$ and $\sigma_1 = \sigma_2 \cdot \exp(\ln VR_T)$. The values of lnRR and lnVR, as well as the sample sizes, depended on the simulation condition (see Table 1). The values of these parameters were chosen to be similar to those used in the simulation study by Senior et al. (2020) to facilitate comparison with their results. Specifically, we considered the same values for the sample sizes (8, 16, 42) and similar ranges for the lnVR and the lnRR (−0.5 to 0.5, respectively). In addition, we varied the reliability of the outcome in group 1 (see Table 1 again) with values ranging from 0.70 to 0.90. These values were chosen because a reliability below 0.7 is usually deemed unacceptable in applied research; hence, reliabilities of scales used in primary studies will typically exceed this value. As Equations (11) and (12) indicate, lnVR and lnCVR are distorted by the *ratio* of reliabilities. By setting reliability to a high value in group 2 (i.e., 0.9) and considering only lower or comparable values for group 1, our simulation conditions therefore cover a large range of reliability ratios (0.78 to 1). Because the (square root of the) reliability ratio enters Equations (11) and (12) on the log scale, swapping the reliabilities of the two groups only affects the sign. Hence, it is not necessary to also consider conditions with a lower reliability in group 2. Data were generated by sampling observations from a multivariate normal distribution with expectations and variance-covariance matrix as implied by the model shown in Figure 1. In all conditions, we generated balanced sample sizes, that is, we set $n_1 = n_2$.[2] Finally, means and standard deviations that were needed for calculating effect sizes and sampling variances in a replication were obtained by calculating the mean and standard deviation of the sum scores for each group.

We used a CFA approach to obtain group-specific reliability estimates and their standard errors. To this end, we used the same approach as for the individual studies to generate CFA data in each simulation replication, but with a larger sample size. Specifically, we drew a sample of 100, 250, or 500 persons per group, fitted the respective CFA model to the data (see Figure 1), and used the CFA results to compute group-specific reliability estimates and their standard errors. A sample size of 100 persons is minimal to obtain reliable estimates (see Kline, 2016; Wolf et al., 2013), while 500 is almost optimal in this respect.

In sum, simulation conditions were determined by fully crossing all parameter values, resulting in a total number of $5 \times 5 \times 3 \times 5 \times 3 = 1,125$ simulation conditions in the pre-study. In each condition, we generated 100,000 simulation replications.

Performance measures

We used the bias of the estimates to evaluate the performance of the effect size estimators. Bias was calculated as $\hat{\theta}_r - \theta$, where $\hat{\theta}_r$ is the estimate obtained for the r-th simulation replication ($r = 1, \ldots, 100,000$) and θ is the true value of the respective effect size (i.e., $\ln VR_T$ or $\ln CVR_T$). We used the relative bias to evaluate the performance of the estimators for the sampling variances:

$$\text{bias}(s_r^2(\hat{\theta})) = \frac{s_r^2(\hat{\theta}) - \widehat{\text{Var}}(\hat{\theta})}{\widehat{\text{Var}}(\hat{\theta})}, \tag{20}$$

where $s_r^2(\hat{\theta})$ is an estimate of the sampling variance obtained for the estimator $\hat{\theta}$ for the r-th simulation replication and $\widehat{\text{Var}}(\hat{\theta})$ is the empirical variance of the corrected effect size estimator (i.e., $\widehat{\ln VR_T}$ or $\widehat{\ln CVR_T}$) across simulation replications. Relative bias was computed separately for uncorrected and corrected sampling variance estimates. With respect to relative bias of the sampling variances, we consider absolute biases of $<5\%$ as negligible, absolute biases between 5 and 10% as moderate, and absolute biases of $>10\%$ as substantial.

[2]We did not consider unbalanced sample sizes because we would not expect that imbalance would affect the simulation results. This is because the impact of the two groups' sample sizes on the sampling variances is independent of each other (see Equations (5), (6), (16), and (17)).

Main simulation: meta-analytic setting

The main simulation was done to evaluate the performance of the proposed corrections in a meta-analytic setting. We compared three procedures: (a) Neither correcting effect size estimates nor sampling variances for unreliability (i.e., using Equations (3) and (4) for effect size estimation and (5) and (6) for sampling variance estimation), (b) correcting only effect size estimates (i.e., using Equations (13) and (14) for effect size estimation and (5) and (6) for sampling variance estimation), and (c) correcting both effect size estimates and sampling variance estimates (i.e., using Equations (13) and (14) for effect size estimation and (16) and (17) for sampling variance estimation). Based on the findings from the pre-study (see below), we would expect procedure (c) to perform best. We included procedure (b) to evaluate whether reasonably accurate results can be obtained when no information on the uncertainty of the reliability estimates used in the correction is available. We expect that this will often be the case because primary studies frequently fail to report this information (Kelley & Pornprasertmanit, 2016). For the estimation of the between-study variance τ^2, we used the restricted maximum likelihood estimator (Raudenbush, 2009; Viechtbauer, 2005). Standard errors were calculated using the Hartung-Knapp-Sidik-Jonkman method (Hartung & Knapp, 2001; Sidik & Jonkman, 2002).

Data generation

In each simulation replication, data for k studies were generated. The true lnVR for each study i, lnVR_{Ti} $(i = 1, \ldots, k)$ was drawn from a normal distribution with mean $\mu(\text{lnVR}_T)$ and variance $\tau^2(\text{lnVR}_T)$. In addition, the true lnRR for each study, lnRR_i was drawn from a normal distribution with mean 0 and variance $\tau^2(\text{lnRR})$. Accordingly, the between-study variance of the lnCVR was $\tau^2(\text{lnCVR}_T) = \tau^2(\text{lnVR}_T) + \tau^2(\text{lnRR})$. We generated both lnVR_{Ti} and lnRR_i from normal distributions because then, lnCVR_{Ti} is also normally distributed. Thus, all data were simulated in accordance with the assumptions of the random-effects meta-analysis model, avoiding model misspecification as a potential additional source of bias in the simulation.

For each individual study, we used the same data-generating model as in the pre-study (cf. Figure 1). That is, values for group 2 were $\mu_2 = 50$, $\sigma_2 = 10$ and $\text{Rel}_2 = 0.9$ for all studies in all simulation conditions, and the mean and the standard deviation of the true-score variable in group 1 were obtained with $\mu_{1i} = \mu_2 \cdot \exp(\text{lnRR}_i)$ and $\sigma_{1i} = \sigma_2 \cdot \exp(\text{lnVR}_{\tau,i})$, respectively. For each study, the sample size of group 2 was drawn from a discrete uniform distribution with lower bound $n_2 - 5$ and upper bound $n_2 + 5$. Again, the same sample size was used for group 1 of the respective study.

Table 2. Values of the simulation parameters in the main simulation.

Parameter	Values
Number of studies k	15, 25
Heterogeneity of the lnVR $\tau^2(\text{lnVR}_T)$	0, 0.05
Heterogeneity of the lnRR $\tau^2(\text{lnRR})$	0, 0.05
Average lnVR $\mu(\text{lnVR}_T)$	−0.25, 0, 0.25
Sample size n_2 (n_1)	16, 42
Reliability in group 1 Rel_1	0.7, 0.75, 0.8, 0.85, 0.9

Table 2 shows the levels of the factors that we varied in this simulation. The values for reliability in group 1, the sample size, and the average lnVR were selected based on the pre-study. We considered small and large numbers of studies (15 and 25, respectively). With regard to heterogeneity, we considered conditions with a between-study variance of zero or 0.05 for both lnVR and lnRR. Using the same values for lnVR and lnRR facilitates comparing the influence of heterogeneity in lnVR *vs.* lnRR on the estimation of lnCVR. We deem a between-study variance of 0.05 to be relatively large because a prediction interval for lnVR with $\tau^2 = 0.05$ spans a considerable range.[3] Simulation conditions were obtained by fully crossing these factors, resulting in a total number of $2 \times 2 \times 2 \times 3 \times 2 \times 5 = 240$ simulation conditions. We generated 1,000 replications per condition.

Per simulation replication, one additional set of CFA data was simulated in the same way as an individual study, but with a sample size of 100 per group. From these data, reliability estimates and their standard errors were obtained and then used to estimate the corrected lnVR, the corrected lnCVR, and their sampling variances in each of the k individual studies.

Performance measures

Performance was evaluated in terms of bias in the estimation of the average effect. It was calculated with $\hat{\mu}(\theta)_r - \mu(\theta)$ for each effect size θ (i.e., lnVR_T and lnCVR_T) and simulation replication r $(r = 1, \ldots, 1,000)$. In addition, we evaluated the type I

[3]For a pooled lnVR of zero, the prediction interval ($\mu \pm 1.96\tau$, i.e., not taking into account the standard error) is approximately $(-0.44, 0.44)$. Exponentiating the bounds yields a prediction interval of $(0.65, 1.55)$ on the VR scale. In consequence, the prediction interval covers situations in which the standard deviation in group 1 is between two thirds of and one and a half times the standard deviation in group 2.

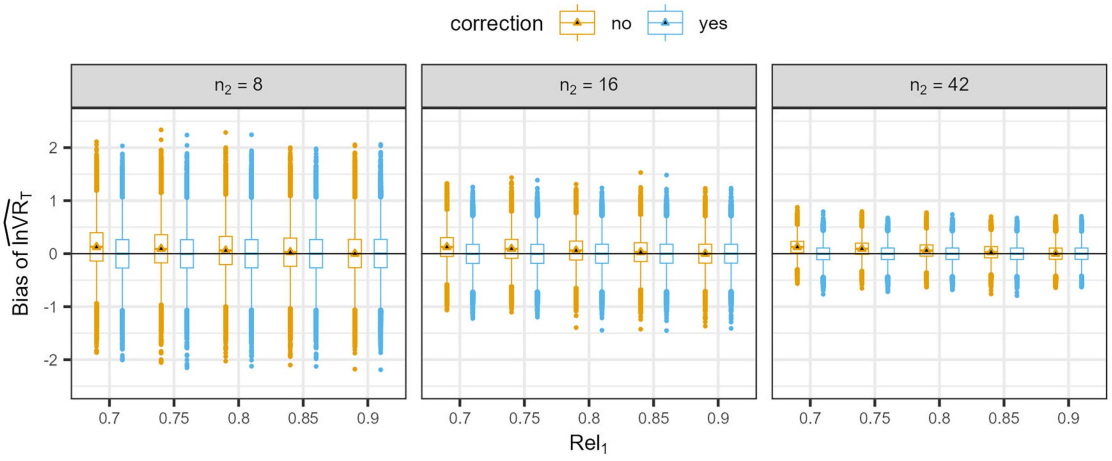
error rate and the power of the average effect size test. The type I error rate was obtained as the percentage of replications in a simulation condition with $\mu(\theta) = 0$ in which the t-test indicated that the average effect was significantly different from zero. The power was calculated by computing the percentage of simulation replications in conditions with $\mu(\theta) \neq 0$ in which the t-test indicated that the average effect was significantly different from zero. Finally, we evaluated the performance of the different estimators in terms of bias in the estimation of the between-study variance, calculated as $\hat{\tau}_r^2(\theta) - \tau^2(\theta)$. We use boxplots to visualize biases of $\hat{\mu}(\theta)$ and $\hat{\tau}^2$. Each boxplot depicts a single condition and therefore shows the distribution of bias for this particular condition across simulation replications.

Results of the simulation studies

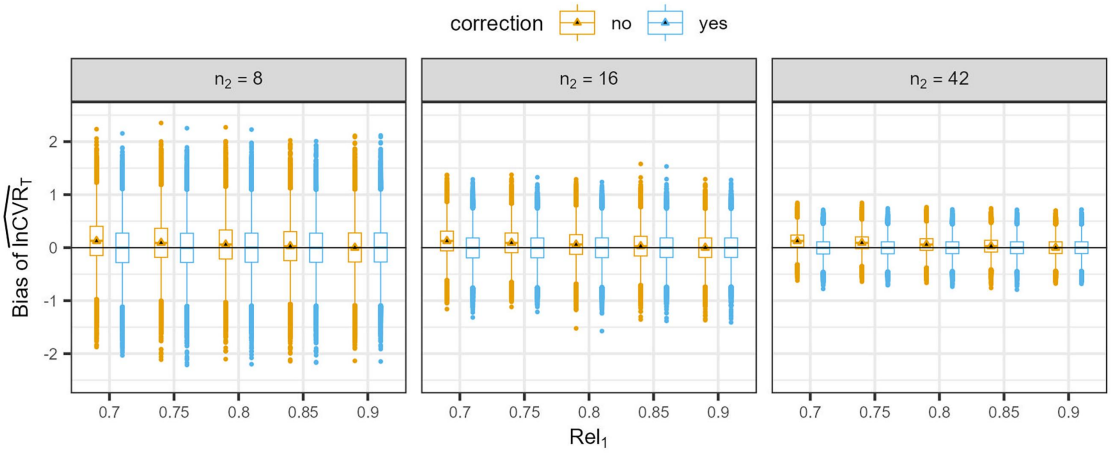
Results of the pre-study

Bias of the effect size estimators

Bias of the effect size estimators was very similar across simulation conditions that differed in terms of lnRR, $\ln VR_T$, or the sample size of the CFA study. In Figure 2, we therefore present the results for $\ln RR = 0$, $\ln VR_T = 0$ and $n_{2,R} = 100$ to exemplify these patterns. Figures for all remaining conditions are available at https://osf.io/nkbdm. Figure 2a shows boxplots of the bias of the corrected and uncorrected estimators for $\ln VR_T$, while Figure 2b shows boxplots of the bias of the corrected and uncorrected estimators for $\ln CVR_T$. For both effect size measures, median bias of the uncorrected



(a) Bias of the Estimators for $\ln VR_T$



(b) Bias of the Estimators for $\ln CVR_T$

Figure 2. Bias of the effect size estimators in conditions with $\ln VR_T = 0$ and $\ln RR = 0$ for a sample size in the CFA study of $n_{2,R} = 100$.

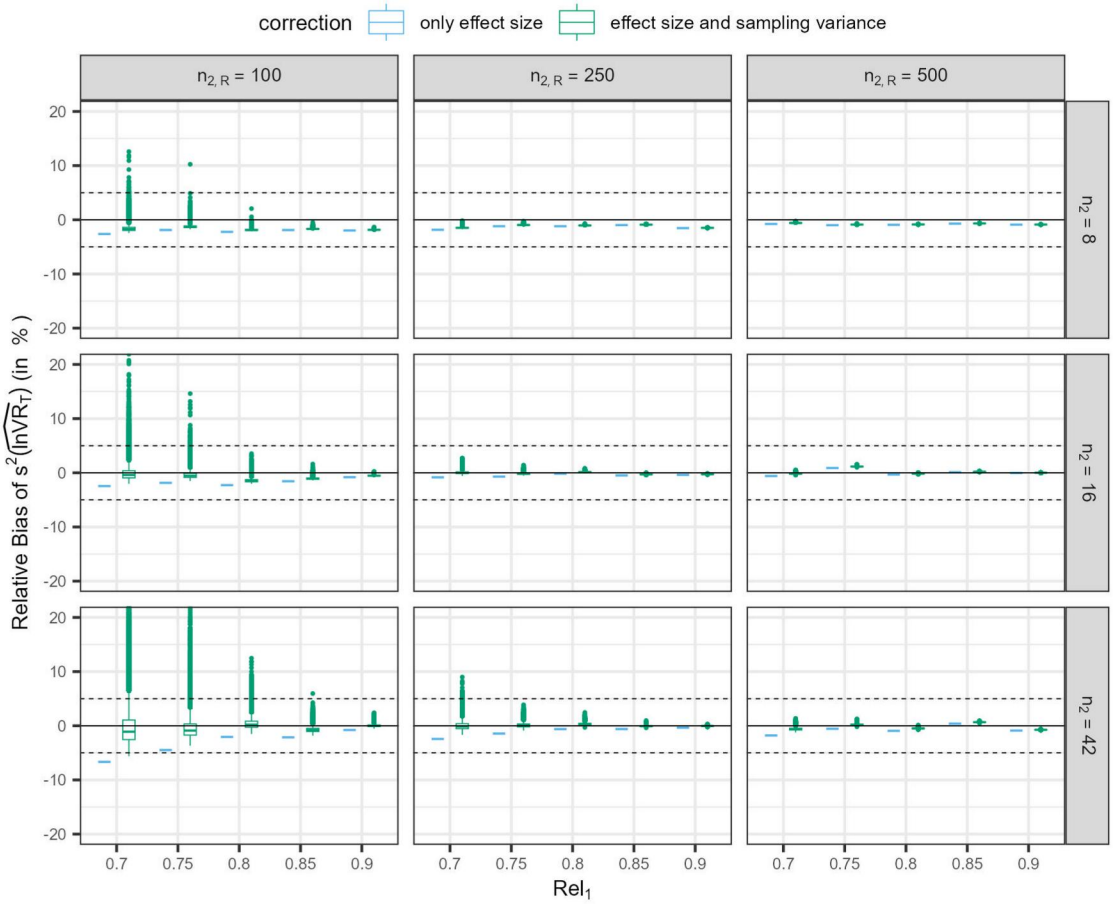


Figure 3. Relative bias of the sampling variance estimators for $\widehat{\ln VR}_T$ in conditions with $\ln VR_T = 0$ and $\ln RR = 0$ for different study-specific sample sizes (vertical panels) and different sample sizes of the additional CFA study (horizontal panels).

estimators only depended on the magnitude of $\ln\left(\sqrt{\widehat{\mathrm{Rel}}(X_1)/\widehat{\mathrm{Rel}}(X_2)}\right)$, and therefore, was largest for the conditions in which $\mathrm{Rel}(X_1)$ differed most from $\mathrm{Rel}(X_2)$. As expected, median bias of the corrected estimators was virtually zero in all conditions.

Relative bias of the sampling variance estimators

Relative bias of the sampling variance estimators exhibited similar patterns across simulation conditions that differed in terms of lnRR and $\ln VR_T$. Figures 3 and 4 show the results for conditions with $\ln VR_T = 0$ and $\ln RR = 0$ to exemplify these patterns. Figures for all further conditions are available at the OSF repository.

For both effect size measures, the uncorrected estimators of the sampling variances tended to underestimate the sampling variance of $\widehat{\ln VR}_T$ and $\widehat{\ln CVR}_T$ more often as compared to the corrected estimators. Nevertheless, both the uncorrected and the corrected sampling variance estimator had a negligible bias for both effect size measures in most simulation conditions. Exceptions of this were conditions with a small sample size of the CFA study, large sample sizes within the studies, and large reliability differences between groups (cf. the lower left panels in Figures 3 and 4): In these conditions, bias of the uncorrected sampling variance estimator tended to be moderate. For the corrected sampling variance estimator of the lnVR, bias was close to zero in the vast majority of simulation replications for all simulation conditions, but in conditions with a smaller sample size for the CFA, there tended to be more replications in which the sampling variance was overestimated. Note that since the uncorrected sampling variance estimator of the lnVR is only based on the sample sizes within the studies (cf. Equation (5)), and as these were held constant across simulation replications, there was no variability of bias across simulation replications. For the corrected sampling variance estimator of the lnCVR, bias was negligible in almost all simulation conditions, with exceptions only in very few extreme conditions (i.e., conditions with $\ln VR_T = 0.5$, $\ln RR = -0.5$, $n_2 = 8$, $n_{2,R} \in \{100, 250\}$, $\mathrm{Rel}(X_1) = 0.7$, and $\ln VR_T = 0.5$, $\ln RR = -0.5$, $n_2 = 8$, $n_{2,R} = 500$, $\mathrm{Rel}(X_1) = 0.85$).

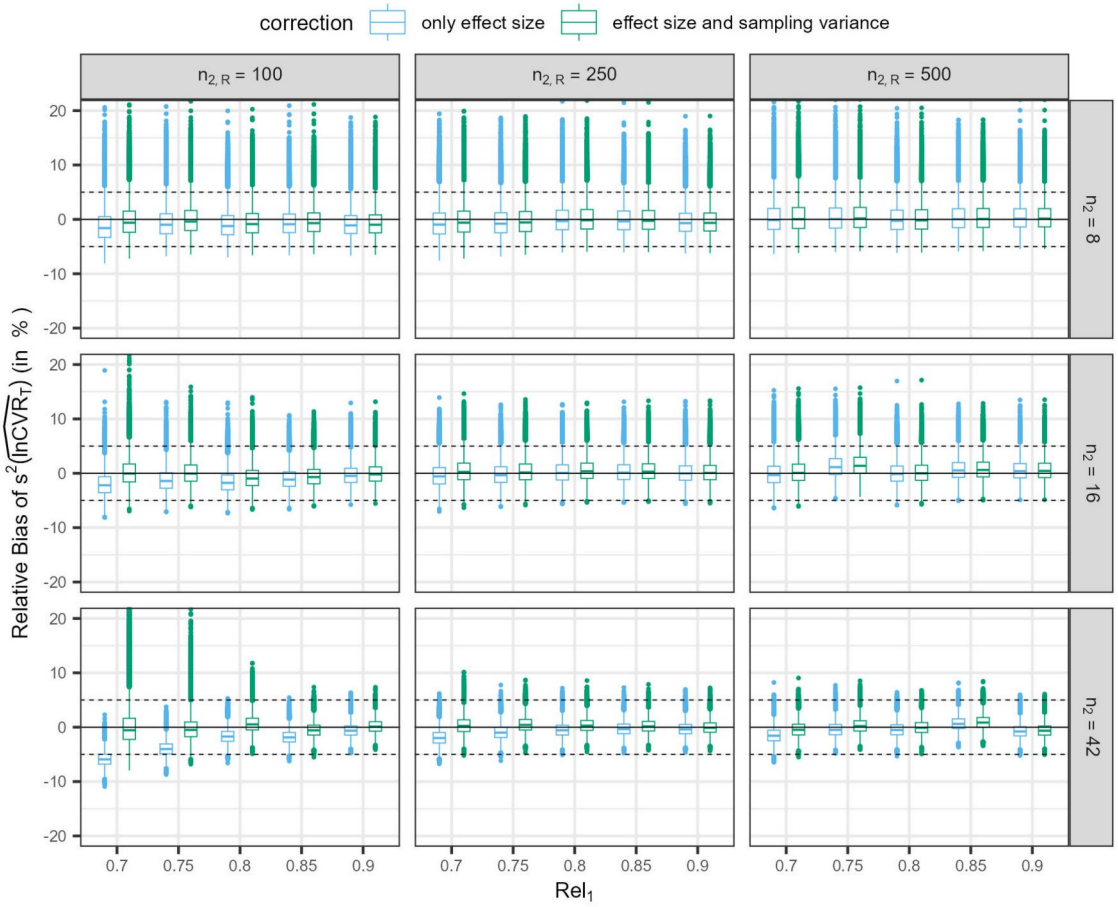


Figure 4. Relative bias of the sampling variance estimators for $\widehat{\ln CVR}_T$ in conditions with $\ln VR_T = 0$ and $\ln RR = 0$ for different study-specific sample sizes (vertical panels) and different sample sizes of the additional CFA study (horizontal panels).

In summary, we have found that the correction formulas reduce bias, both in the effect sizes and in the sampling variances. These results indicate that the corrections should also work in the meta-analytic setting, and the main simulation was done to examine this.

Results of the main simulation

The general patterns found in the results of the main simulation study were similar for the two different effect size measures. Therefore, we focus on the results for lnVR in this section. Figures showing the results for lnCVR are available at the OSF repository.

Bias in the estimation of the average effect

Figure 5 illustrates the results with respect to bias in the estimation of $\mu(\ln VR_T)$ when (a) neither correcting effect sizes estimates nor sampling variance estimates for differences in reliability, (b) correcting only effect size estimates, and (c) correcting both effect size estimates and sampling variance estimates. As the results were similar for different values of $\mu(\ln VR_T)$, we only show the results for conditions with $\mu(\ln VR_T) = 0$.

As expected, when neither correcting effect size estimates nor sampling variance estimates, bias depended on the magnitude of $\ln\left(\sqrt{\widehat{Rel}(X_1)/\widehat{Rel}(X_2)}\right)$, and hence, was larger the more $Rel(X_1)$ differed from $Rel(X_2)$. When the effect size estimates were corrected for reliability differences, median bias was virtually zero in all simulation conditions. It did not make a difference whether sampling variances were also corrected or not, which is a result of the unbiasedness of effect size estimation in the individual studies (cf. Figure 2).

Type I error rates

Figure 6 shows the empirical type I error rates for the test of the hypothesis $\mu(\ln VR_T) \neq 0$.

When using uncorrected estimators for both effect sizes and sampling variances, type I error rates were mainly driven by bias, and therefore increased as the between-group difference in reliabilities increased. In addition, there was a pronounced increase in the type

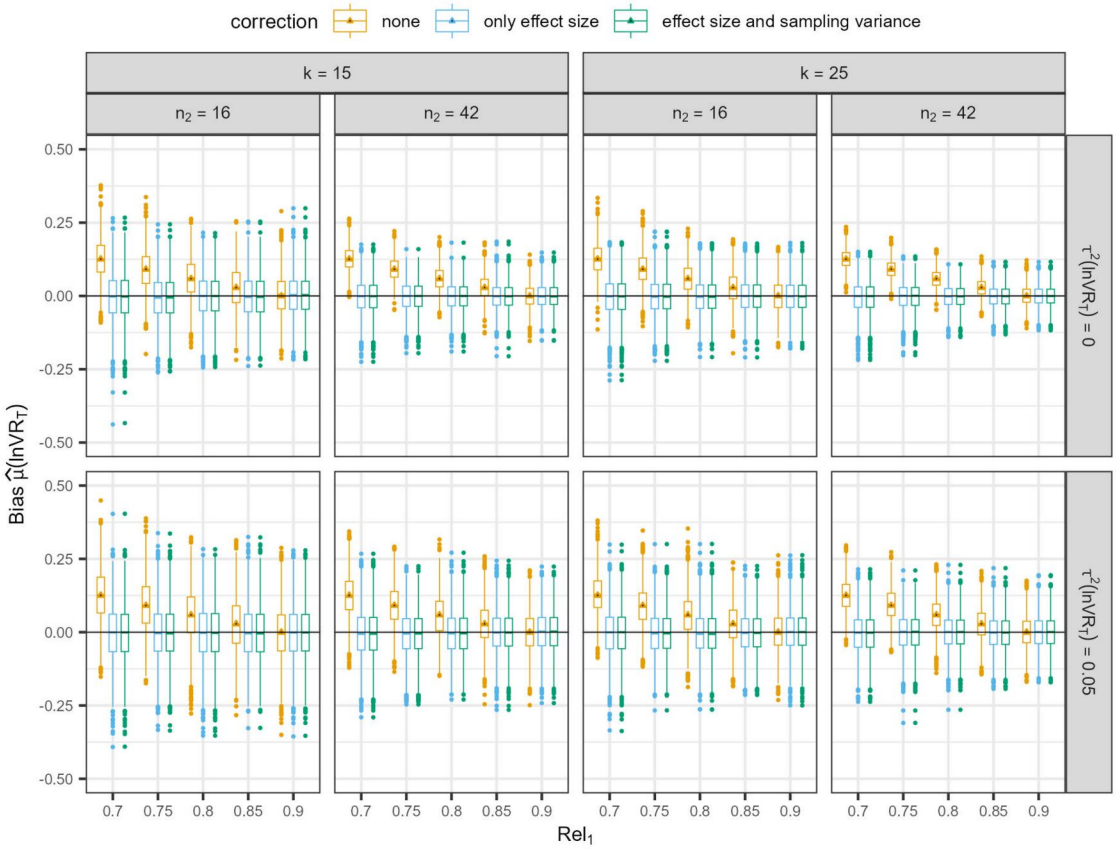


Figure 5. Bias in the estimation of the average effect $\mu(\ln VR_T)$ in conditions with $\mu(\ln VR_T) = 0$ for different numbers of studies and sample sizes (nested horizontal panels) and different levels of between-study heterogeneity (vertical panels).

I error rates for larger sample sizes and larger numbers of studies, as well as for smaller between-study variances. Type I error rates were similar for the two correction procedures. In general, they were considerably smaller than the type I error rates when not using a correction, but were slightly inflated for larger sample sizes, larger numbers of studies, smaller between-study variances, and larger reliability differences. The increase in type I error rates for larger reliability differences and larger sample sizes can be explained by a larger downward bias of the sampling variances in this situation (cf. Figures 3 and 4 from the pre-study). The effect of this downward bias is greater for larger numbers of studies because the biases of the sampling variances add up in the estimation of the standard error of the average effect. We return to the issue of inflated type I error rates in the discussion.

Power

Figure 7 shows the empirical power for the test of the hypothesis $\mu(\ln VR_T) \neq 0$ for the three procedures. Here, we only show the results for $\ln VR = -0.25$, because similar to type I error rates, power was mainly driven by bias when no correction was used. In the conditions considered in our simulation, there was an upward bias due to $\mathrm{Rel}(X_1)$ being smaller than $\mathrm{Rel}(X_2)$, and hence it is trivial that power will be larger for the uncorrected procedure than when using corrections for $\ln VR_T > 0$.

The results depicted in Figure 7 show that power can be compromised to a considerable extent when neither correcting effect size estimates nor sampling variance estimates for differences in reliability. The low power mainly resulted from bias and was thus smaller for larger reliability differences. In addition, smaller sample sizes, smaller numbers of studies, and larger heterogeneity were associated with a lower power. Power of the two correction procedures was similar and was mainly driven by the sample size, the number of studies, and the between-study variance.

Bias in the estimation of the between-study variance

Figure 8 illustrates the results with respect to bias in the estimation of $\tau^2(\ln VR_T)$ for the three procedures. As the results were similar for different values of $\mu(\ln VR_T)$, we only show the results for conditions with $\mu(\ln VR_T) = 0$.

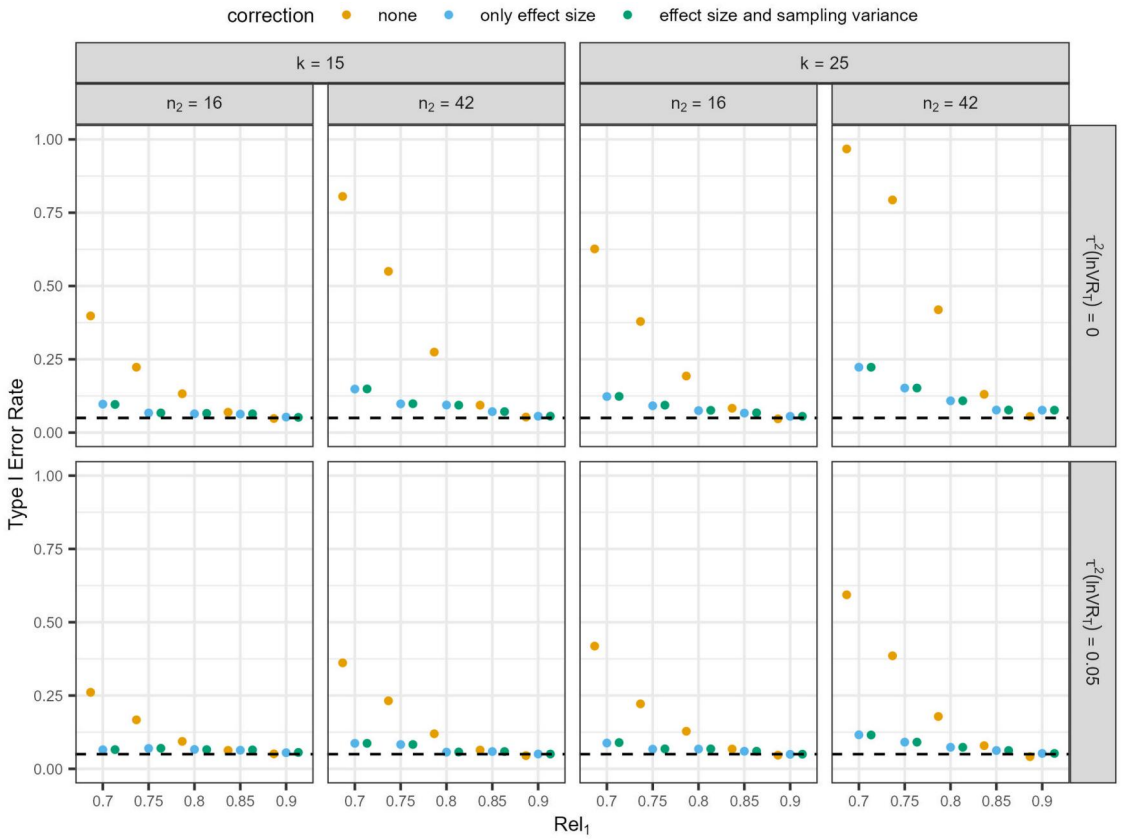


Figure 6. Type I error rates for the test of the hypothesis $\mu(\ln VR_T) \neq 0$ for different numbers of studies and sample sizes (nested horizontal panels) and different levels of between-study heterogeneity (vertical panels).

For all three procedures, bias was small for the majority of simulation replications for all simulation conditions. In conditions with $\tau^2(\ln VR_T) = 0$, bias tended to be slightly smaller when both the effect size estimates and the sampling variances were corrected for differences in reliability, as compared to the two other procedures. In conditions with $\tau^2(\ln VR_T) = 0.05$, correcting both effect size estimates and sampling variances sometimes resulted in a small negative median bias, in particular when sample sizes within studies were small. For the two other procedures, median bias was virtually zero for all simulation conditions.

Illustrative example

We now show how to apply our method in practice by using it to reanalyze data from a meta-analysis by Kim et al. (2019). In their meta-analysis, they considered differences in variability in self- and other-reports of personality. The full data from this meta-analysis are available at https://osf.io/snfjx/. The data and code used in our reanalysis are available in the OSF repository of this paper. Here, we use a subset of these data comprising studies which were published in peer-reviewed journals and in which other-reports were obtained from family members or friends. We further restricted our reanalysis to personality assessments of the Big Five (i.e., openness, conscientiousness, extraversion, agreeableness or neuroticism), which were obtained with the Big Five Inventory (BFI), the Five-Factor Personality Inventory (FFPI), the International Personality Item Pool (IPIP), the IPIP-NEO, the NEO Five-Factor Inventory (NEO-FFI), the NEO Personality Inventory-Revised (NEO-PI-R), or the Ten Item Personality Inventory (TIPI). This selection was made with the intention of providing a succinct example that still allows us to illustrate some considerations that are important when applying the correction for unreliability (see below). For each primary study, we tried to gather information on the reliability of self- and other-reports from its published manuscript. Those studies that reported reliabilities used Cronbach's alpha as their reliability measure. Although this may be suboptimal for some personality scales, we used these values when applying the correction because it was the only information on reliability available from the studies, and conducting an

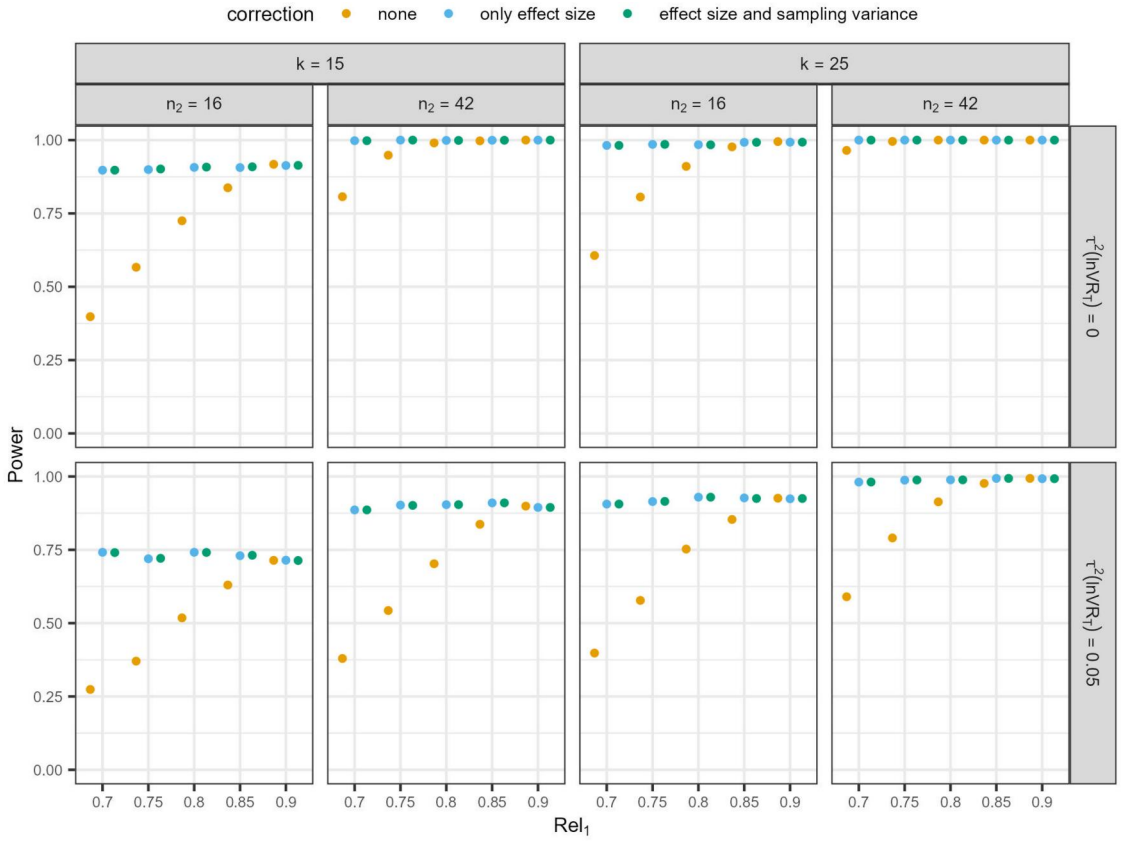


Figure 7. Power of the test of the hypothesis $\mu(\ln VR_T) \neq 0$ in conditions with $\mu(\ln VR_T) = -0.25$ for different numbers of studies and sample sizes (nested horizontal panels) and different levels of between-study heterogeneity (vertical panels).

additional CFA study to determine the best fitting measurement model for each scale would not have been feasible.

We conducted separate meta-analyses for each personality trait. As an effect size, we used the lnVR because we did not expect standard deviations to increase with means. The proposed correction for unreliability (cf. Equation (13)) was applied using study-specific reliabilities if these were reported and the sample size was 100 or larger. When this was not the case, we imputed the log-square root reliability ratio required for the correction. We compared two approaches: In the first approach, we pooled the log-square root reliability ratios from the studies that reported reliabilities separately for each trait. The second approach was similar, but log-square root reliability ratios were pooled separately for each trait *and* personality scale. Because no study using the NEO-FFI reported reliabilities, the imputed log-square root reliability ratio for this scale was based on the reliabilities from studies using the NEO-PI-R. Before pooling, we adjusted these reliabilities for scale length using the inverse Spearman-Brown formula because the NEO-FFI is a short version of the NEO-PI-R. Pooled log-square root reliability ratios were obtained as a sample size-weighted average because information on sampling variability was unavailable. For the same reason, we only corrected effect sizes for unreliability, but not their sampling variances (i.e., we used procedure (b) from the main simulation). Self- and other-reports of personality are correlated, therefore we computed the sampling variance for the lnVR based on the formula for dependent samples (see Senior et al., 2020) using the correlations for self- and other-reports obtained from a large meta-analysis (Connolly et al., 2007, openness: 0.59, conscientiousness: 0.56, extraversion: 0.62, agreeableness: 0.46, neuroticism: 0.51). Because some studies reported effect sizes for multiple samples, we conducted three-level meta-analyses. Table 3 contains the results of these meta-analyses per trait, along with information of the numbers of studies and samples that were included in the analysis.

The results illustrate that the correction mainly had an effect on the analyses comparing the variability of self- and other-reports of conscientiousness and agreeableness. For these two traits, the estimated average lnVR was somewhat (although not much) smaller when accounting for unreliability. For all further traits, results were similar regardless of whether or

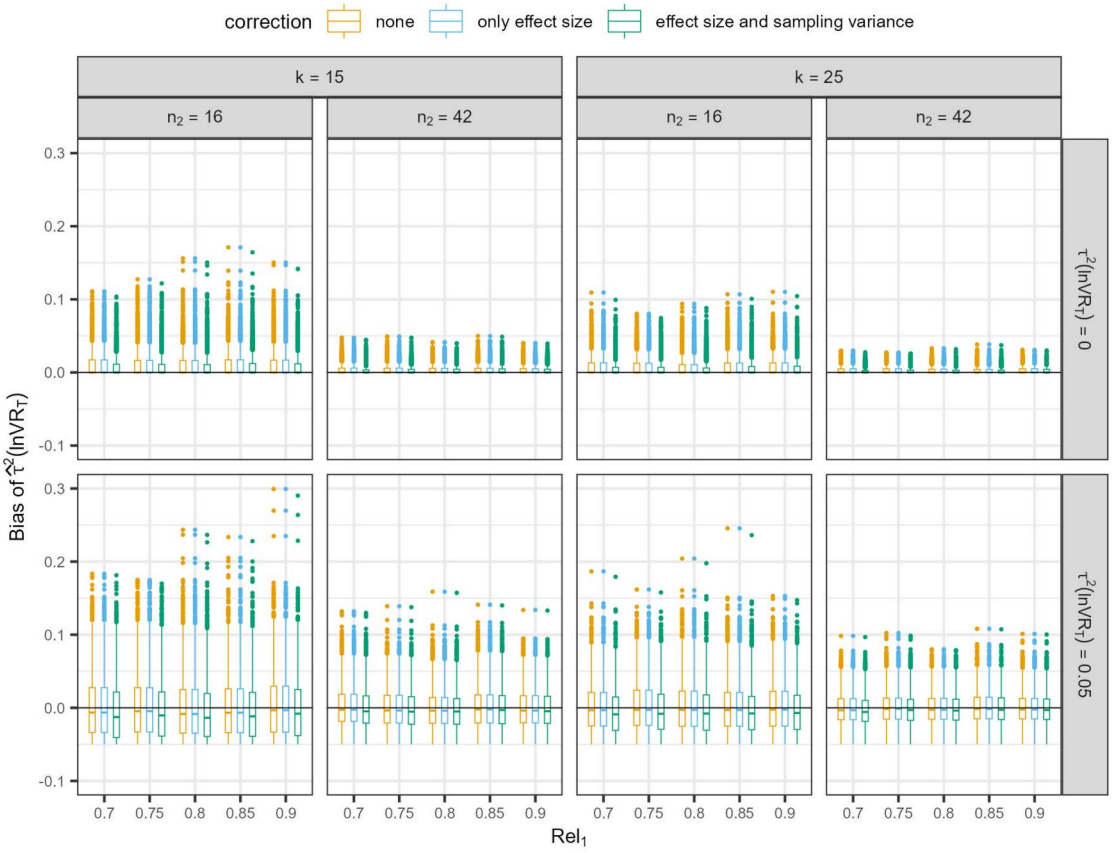


Figure 8. Bias in the estimation of $\tau^2(\ln VR_T)$ in conditions with $\mu(\ln VR_T) = 0$ for different numbers of studies and sample sizes (nested horizontal panels) and different levels of between-study heterogeneity (vertical panels).

Table 3. Results of the meta-analyses for the illustrative example.

Trait	*c* (*k*)	Method	c_{rel}	$\hat{\mu}$	95% CI	τ
Openness	28 (20)	No correction		0.00	[−0.04, 0.05]	0.07
		Approach 1	13	0.01	[−0.04, 0.05]	0.08
		Approach 2	1–5	0.01	[−0.03, 0.06]	0.08
Conscientiousness	20 (32)	No correction		−0.12	[−0.16, −0.09]	0.06
		Approach 1	8	−0.15	[−0.18, −0.11]	0.06
		Approach 2	1–2	−0.15	[−0.18, −0.11]	0.06
Extraversion	27 (19)	No correction		−0.02	[−0.05, 0.01]	0.04
		Approach 1	8	−0.03	[−0.06, 0.00]	0.04
		Approach 2	1–2	−0.03	[−0.07, 0.00]	0.05
Agreeableness	28 (20)	No correction		−0.17	[−0.22, −0.12]	0.09
		Approach 1	8	−0.20	[−0.25, −0.15]	0.10
		Approach 2	1–2	−0.21	[−0.26, −0.15]	0.10
Neuroticism	30 (21)	No correction		0.03	[−0.02, 0.07]	0.09
		Approach 1	10	0.02	[−0.02, 0.07]	0.09
		Approach 2	1–2	0.03	[−0.02, 0.07]	0.08

c: number of comparisons/samples; *k*: number of studies; c_{rel}: number of samples on which the correction was based (Approach 1) or range of the number of samples on which the correction was based (Approach 2, range across scales).

not unreliability was accounted for. Notably, results for the two imputation approaches (imputation per trait *vs.* imputation per trait and scale) were similar for all traits. This implies that it might not be necessary to impute scale-specific reliabilities in this example. We return to the issue of heterogeneous reliabilities in the discussion.

Discussion

Meta-analysis of variances is becoming increasingly popular, hence discussing the impact of measurement error on its results and conclusions is highly relevant. In this article, we demonstrated how between-group differences in reliability affect the estimation of the lnVR and the lnCVR, the two main effect sizes used in meta-analysis of variances. Specifically, we showed that the magnitude of bias of the estimators of lnVR and lnCVR depends on the ratio of the group-specific reliabilities. Based on this result, we proposed a simple correction for the estimators of lnVR and lnCVR that only requires that adequate estimates of the group-specific reliabilities are available.

Two simulations were done to examine the adequacy of the proposed corrections. As expected, we found that each correction removed bias caused by between-group differences in reliability in the study-specific effect size, enabled (almost) unbiased estimation of the average effect in meta-analysis of variances, and also led to more reliable statistical inferences. As each correction introduces additional uncertainty in the estimation of lnVR and lnCVR by relying on estimates of the group-specific reliabilities, we also

examined whether it is necessary to account for this uncertainty in the estimation of the sampling variances of lnVR and lnCVR. We found that accounting for the additional uncertainty indeed usually leads to more accurate sampling variance estimates in individual studies. However, in meta-analysis, the estimation of the average effect, the type I error rate, and the power appeared to be almost unaffected by the additional consideration of the uncertainty. Hence, when reliability estimates from sufficiently large studies are available (as was the case in our simulation study), it seems to be sufficient to correct the effect size estimates for between-group differences in reliability and to use the (standard, uncorrected) formulas proposed by Senior et al. (2020) to estimate their sampling variances.

Simulation results regarding type I error rates revealed that failing to account for unreliability is more detrimental in situations without heterogeneity. A further exploration of our simulation results showed that the standard errors of the average effect were more biased in these conditions than in conditions with heterogeneity, probably because sampling variances of the individual studies were underestimated and the between-study variance was usually accurately estimated to be zero. Of note, when accounting for unreliability, type I errors were still notably inflated in conditions without heterogeneity when the between-group difference in reliabilities was large. The results from our pre-study indicate that this could be countered by imputing reliability estimates from a larger CFA study, because then, sampling variances will be less biased.

Recommendations

In our simulation studies, we presupposed that group-specific reliability estimates from an (external) CFA study with at least 100 subjects per group are available. Our illustrative example shows how to apply the correction when some (sufficiently large) studies report reliabilities while some do not. If information on reliabilities is completely unavailable, researchers may consider making an informed guess on the reliability ratio, for instance, based on studies that examine the reliabilities of similar outcome measures for the groups of interest, or from studies that examine the reliabilities of the outcome measure of interest in groups that are similar to those that shall be examined. It may be advisable to carry out a sensitivity analysis using different reliability ratios in this situation. If making an informed guess is not possible, a potential solution is to conduct an additional CFA study before the meta-analysis to obtain a reliable estimate of the reliability ratio (see Ke & Tong, 2023).

We also presupposed that the true reliabilities underlying the CFA data coincide with the true reliabilities underlying the individual study data pooled in the meta-analysis. We underscore that imputing reliability estimates obtained from samples that are not truly representative of those included in the meta-analysis can induce bias. In practice, it is possible that the true reliabilities not only vary between groups, but also across studies, for example, because they used different scales, assessed different populations, or applied different types of interventions. It is therefore important to consider whether different reliability estimates need to be imputed for different types of studies. In our illustrative example, we showcase how to assess whether accounting for the use of different scales has an impact on the correction. We found that imputing different reliabilities for different scales did not alter the results much, although we cannot rule out that other factors not considered by us would have had an impact. While not accounting for heterogeneous reliabilities does not necessarily induce bias in the average effect (given the imputed reliabilities are similar to the average reliabilities), it can lead to an overestimation of the between-study variance to the extent to which the (true) log-square root reliability ratio varies between studies. To this end, artifact distribution approaches (Hunter & Schmidt, 2004) might provide a further option to correct the average effect for unreliability and simultaneously account for heterogeneous reliabilities. However, they require that a sufficiently large number of (large) studies report group-specific reliability estimates. In summary, these considerations underline the importance of research on between-group differences in reliability, as knowing the conditions under which they occur and being able to estimate their magnitude is a prerequisite for the successful application of reliability corrections, such as those we propose. Furthermore, they emphasize the importance of reporting group-specific reliability estimates in primary studies that are sufficiently large, as these can then be used in meta-analyses of variances.

The formulas that we proposed for correcting the estimators of lnVR and lnCVR for differences in reliability can also be applied when the groups are not independent, as shown in our illustrative example. However, accounting for the additional uncertainty arising from the estimation of the reliability ratio when estimating the sampling variances of lnVR and lnCVR would require that an estimate of the sampling

covariance of the group-specific reliabilities is available. This will usually not be the case. In this respect, the finding from our simulation that correcting the sampling variance might not be necessary as long as the reliability estimates are based on sufficiently large sample sizes is reassuring.

Practical implications and future directions

In our simulation, we made the assumption that the outcomes are normally distributed. An interesting avenue for future research could be to examine the performance of the proposed correction formulas in non-normal data. In this case, the correlation between the logarithmized mean and logarithmized standard deviation has to be accounted for when estimating the sampling variance of the lnCVR (Nakagawa et al., 2015). Further research is needed to evaluate whether it is required to correct this correlation for unreliability to accurately estimate the sampling variance of the lnCVR in non-normally distributed data.

Beyond that, we relied on meta-analytic methods which assume that effects are normally distributed between studies. We would like to emphasize that the proposed correction methods can be used in combination with any estimator for the between-study variance τ^2 (in particular, also with those that do not require the assumption of a normal distribution between studies), as well as together with meta-analytic models that assume non-normal random-effects distributions (for an overview, see Panagiotopoulou et al., 2024). Because our proposed corrections yielded almost unbiased study-specific effect size estimates and sampling variances (as the results of the first simulation study showed), we would expect them to perform well in combination with these methods, too, given their assumptions hold.

Furthermore, we only evaluated our proposed formulas with regard to the estimation of the average effect and the between-study variance in a setting without moderators. As meta-analyses often include moderator analyses (Tipton et al., 2019), it would be interesting to consider in future studies how correcting for differences in unreliability affects the estimation of the coefficients in meta-regression.

Although several reliability generalization studies have revealed heterogeneity in reliability estimates (Aslan et al., 2022; Badenes-Ribera et al., 2023; Cabedo-Peris et al., 2021; Cerri et al., 2023; Demir et al., 2024; Esparza-Reig et al., 2021; Gisbert-Pérez et al., 2022; Yin & Fan, 2000), it remains largely unclear which factors affect the reliability of outcome measures. Therefore, it is impossible to state to what extent not accounting for differences in reliability may have affected the conclusions obtained in existing meta-analyses of variances. Future reliability generalization studies should thus focus on moderator analyses to shed light on the magnitude of between-group differences in reliability. In addition, we recommend that primary studies that examine two or more groups report group-specific reliability estimates for their outcomes. This would enable the use of the formulas proposed here in future meta-analyses of variances.

In the present study, we focused on corrections for unreliability in *meta-analysis*. It should be noted that the corrected estimators that we proposed for the lnVR and lnCVR along with their sampling variances can also be used to construct tests for variability differences in a *single* study. We think that examining these but also alternative ways to account for unreliability in tests for variability differences (e.g., using multiple group structural equation modeling, cf. Tucker-Drob, 2011) is an interesting avenue for future research.

Conclusion

There is a growing interest in meta-analysis of variances, also because they can be easily conducted by reusing existing data from meta-analyses of means. In this article, we have shown that the results obtained in meta-analysis of variances are sensitive to between-group differences in reliability. The results of our simulation studies underline the importance of gathering information on between-group differences in reliability before conducting a meta-analysis that uses the lnVR or lnCVR as an effect size. When such information is available, corrections for between-group differences in reliability are easy to apply and lead to reliable meta-analytic results.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Abdulla Alabbasi, M., Thompson, T. L., Runco, M. A., Alansari, L. A., & Ayoub, A. E. A. (2025). Gender differences in creative potential: A meta-analysis of mean differences and variability. *Psychology of Aesthetics, Creativity, and the Arts*.

Althobaiti, S., Kazantzis, N., Ofori-Asenso, R., Romero, L., Fisher, J., Mills, K. E., & Liew, D. (2020). Efficacy of interpersonal psychotherapy for post-traumatic stress disorder: A systematic review and meta-analysis. *Journal of Affective Disorders*, *264*, 286–294. https://doi.org/10.1016/j.jad.2019.12.021

Aslan, Ö. Ş., Göcen, S., & Şen, S. (2022). Reliability generalization meta-analysis of mathematics anxiety scale for primary school students. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(2), 117–133. https://doi.org/10.21031/epod.1119308

Badenes-Ribera, L., Duro-García, C., López-Ibáñez, C., Martí-Vilar, M., & Sánchez-Meca, J. (2023). The Adult Prosocialness Behavior Scale: A reliability generalization meta-analysis. *International Journal of Behavioral Development*, *47*(1), 59–71. https://doi.org/10.1177/01650254221128280

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). John Wiley & Sons.

Brugger, S. P., Angelescu, I., Abi-Dargham, A., Mizrahi, R., Shahrezaei, V., & Howes, O. D. (2020). Heterogeneity of striatal dopamine function in schizophrenia: Meta-analysis of variance. *Biological Psychiatry*, *87*(3), 215–224. https://doi.org/10.1016/j.biopsych.2019.07.008

Bru-Luna, L. M., Martí-Vilar, M., Merino-Soto, C., & Livia, J. (2021). Reliability generalization study of the Person-Centered Care Assessment Tool. *Frontiers in Psychology*, *12*, 712582. https://doi.org/10.3389/fpsyg.2021.712582

Cabedo-Peris, J., Martí-Vilar, M., Merino-Soto, C., & Ortiz-Morán, M. (2021). Basic Empathy Scale: A systematic review and reliability generalization meta-analysis. *Healthcare*, *10*(1), 29. https://doi.org/10.3390/healthcare10010029

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman & Hall/CRC.

Cerri, L. Q., Justo, M. C., Clemente, V., Gomes, A. A., Pereira, A. S., & Marques, D. R. (2023). Insomnia Severity Index: A reliability generalisation meta-analysis. *Journal of Sleep Research*, *32*(4), e13835. https://doi.org/10.1111/jsr.13835

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment*, *15*(1), 110–117. https://doi.org/10.1111/j.1468-2389.2007.00371.x

Cristea, I. A., Gentili, C., Cotet, C. D., Palomba, D., Barbui, C., & Cuijpers, P. (2017). Efficacy of psychotherapies for borderline personality disorder: A systematic review and meta-analysis. *JAMA Psychiatry*, *74*(4), 319–328. https://doi.org/10.1001/jamapsychiatry.2016.4287

Cuijpers, P., Clignet, F., Van Meijel, B., Van Straten, A., Li, J., & Andersson, G. (2011). Psychological treatment of depression in inpatients: A systematic review and meta-analysis. *Clinical Psychology Review*, *31*(3), 353–360. https://doi.org/10.1016/j.cpr.2011.01.002

Demir, E., Öz, S., Aral, N., & Gürsoy, F. (2024). A reliability generalization meta-analysis of the Mother-To-Infant Bonding Scale. *Psychological Reports*, *127*(1), 447–464. https://doi.org/10.1177/00332941221114413

Esparza-Reig, J., Guillén-Riquelme, A., Martí-Vilar, M., & González-Sala, F. (2021). A reliability generalization meta-analysis of the South Oaks Gambling Screen (SOGS). *Psicothema*, *33*(3), 490–499. https://doi.org/10.7334/psicothema2020.449

Gisbert-Pérez, J., Martí-Vilar, M., Merino-Soto, C., & Vallejos-Flores, M. (2022). Reliability generalization meta-analysis of Internet Gaming Disorder Scale. *Healthcare*, *10*(10), 1992. https://doi.org/10.3390/healthcare10101992

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*(6), 930–944. https://doi.org/10.1177/0013164406288165

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Gunnerud, H. L., Ten Braak, D., Reikerås, E. K. L., Donolato, E., & Melby-Lervåg, M. (2020). Is bilingualism related to a cognitive advantage in children? A systematic review and meta-analysis. *Psychological Bulletin*, *146*(12), 1059–1083. https://doi.org/10.1037/bul0000301

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary

outcome. *Statistics in Medicine*, *20*(24), 3875–3889. https://doi.org/10.1002/sim.1009

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156. https://doi.org/10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2

Hernan, M. A., & Robins, J. M. (2024). *Causal inference: What if* (1st ed.). Chapman & Hall/CRC.

Herzog, P., & Kaiser, T. (2022). Is it worth it to personalize the treatment of PTSD? – A variance-ratio meta-analysis and estimation of treatment effect heterogeneity in RCTs of PTSD. *Journal of Anxiety Disorders*, *91*, 102611. https://doi.org/10.1016/j.janxdis.2022.102611

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

Imbens, G., & Rubin, D. B. (2015). *Causal inference for statistics, social and biomedical sciences: An introduction*. Cambridge University Press.

Kaiser, T., & Herzog, P. (2023). Is personalized treatment selection a promising avenue in BPD research? A meta-regression estimating treatment effect heterogeneity in RCTs of BPD. *Journal of Consulting and Clinical Psychology*, *91*(3), 165–170. https://doi.org/10.1037/ccp0000803

Kaiser, T., Volkmann, C., Volkmann, A., Karyotaki, E., Cuijpers, P., & Brakemeier, E.-L. (2022). Heterogeneity of treatment effects in trials on psychotherapy of depression. *Clinical Psychology: Science and Practice*, *29*(3), 294–303. https://doi.org/10.1037/cps0000079

Ke, Z., & Tong, X. (2023). Correcting for the multiplicative and additive effects of measurement unreliability in meta-analysis of correlations. *Psychological Methods*, *28*(1), 21–38. https://doi.org/10.1037/met0000396

Keller, F., Kühner, C., Alexandrowicz, R. W., Voderholzer, U., Meule, A., Fegert, J. M., & Hautzinger, M. (2022). Zur Messqualität des Beck-Depressionsinventars (BDI-II) in unterschiedlichen klinischen Stichproben: Eine Item-Response-Theorie Analyse. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *51*(3–4), 234–246.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. https://doi.org/10.1037/a0040086

Kim, H., Di Domenico, S. I., & Connelly, B. S. (2019). Self-other agreement in personality reports: A meta-analytic comparison of self- and informant-report means. *Psychological Science*, *30*(1), 129–138. https://doi.org/10.1177/0956797618810000

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Kühner, C., Bürger, C., Keller, F., & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II): Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt*, *78*(6), 651–656. https://doi.org/10.1007/s00115-006-2098-7

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Publishers.

Mills, H. L., Higgins, J. P. T., Morris, R. W., Kessler, D., Heron, J., Wiles, N., Davey Smith, G., & Tilling, K. (2021). Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between trial arms. *Epidemiology*, *32*(6), 846–854. https://doi.org/10.1097/EDE.0000000000001401

Mõttus, R., Allik, J., Hřebíčková, M., Kööts–Ausmees, L., & Realo, A. (2016). Age differences in the variance of personality characteristics. *European Journal of Personality*, *30*(1), 4–11. https://doi.org/10.1002/per.2036

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, *6*(2), 143–152. https://doi.org/10.1111/2041-210X.12309

Nestler, S., & Salditt, M. (2024). Comparing type 1 and type 2 error rates of different tests for heterogeneous treatment effects. *Behavior Research Methods*, *56*(7), 6582–6597. https://doi.org/10.3758/s13428-024-02371-x

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications*, *9*(1), 3777. https://doi.org/10.1038/s41467-018-06292-0

Osimo, E. F., Pillinger, T., Rodriguez, I. M., Khandaker, G. M., Pariante, C. M., & Howes, O. D. (2020). Inflammatory markers in depression: A meta-analysis of mean differences and variability in 5,166 patients and 5,083 controls. *Brain, Behavior, and Immunity*, *87*, 901–909. https://doi.org/10.1016/j.bbi.2020.02.010

Panagiotopoulou, K., Evrenoglou, T., Schmid, C. H., Metelli, S., & Chaimani, A. (2024). Meta-analysis models relaxing the random effects normality assumption: Methodological systematic review and simulation study. *arXiv*. https://doi.org/10.48550/arXiv.2412.12945

Plöderl, M., & Hengartner, M. P. (2019). What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. *BMJ Open*, *9*(12), e034816. https://doi.org/10.1136/bmjopen-2019-034816

R Core Team (2022). *R: A language and environment for statistical computing*. https://www.R-project.org/

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*(3), 432–446. https://doi.org/10.1037/0021-9010.76.3.432

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–316). Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, *12*(3), 241–269. https://doi.org/10.3102/10769986012003241

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Salditt, M., Eckes, T., & Nestler, S. (2024). A tutorial introduction to heterogeneous treatment effect estimation with meta-learners. *Administration and Policy in Mental Health*, *51*(5), 650–673. https://doi.org/10.1007/s10488-023-01303-9

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. SAGE Publications, Ltd.

Senior, A. M., Viechtbauer, W., & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods*, *11*(4), 553–567. https://doi.org/10.1002/jrsm.1423

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, *21*(21), 3153–3159. https://doi.org/10.1002/sim.1262

Taylor, C. L., & Barbot, B. (2021). Gender differences in creativity: Examining the greater male variability hypothesis in different domains and tasks. *Personality and Individual Differences*, *174*, 110661. https://doi.org/10.1016/j.paid.2021.110661

Taylor, C. L., Said-Metwaly, S., Camarda, A., & Barbot, B. (2024). Gender differences and variability in creative ability: A systematic review and meta-analysis of the greater male variability hypothesis in creativity. *Journal of Personality and Social Psychology*, *126*(6), 1161–1179. https://doi.org/10.1037/pspp0000484

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Tucker-Drob, E. M. (2011). Individual differences methods for randomized experiments. *Psychological Methods*, *16*(3), 298–318. https://doi.org/10.1037/a0023349

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*(3), 271–280. https://doi.org/10.1007/BF02296146

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123. https://doi.org/10.1177/2515245919885611

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201–223. https://doi.org/10.1177/00131640021970466

Zhang, Q. (2024). Meta-analysis of correlation coefficients: A cautionary tale on treating measurement error. *Psychological Methods*, *29*(2), 308–330. https://doi.org/10.1037/met0000498