Handling Effect Size Dependency in Meta-Analysis

Daniel F. Gucciardi[1], Robin L.J. Lines[1], and Nikos Ntoumanis[2,3,4]

[1] *Curtin School of Allied Health, Curtin University*

[2] *Department of Sports Science and Clinical Biomechanics, University of Southern Denmark*

[3] *Curtin School of Population Health, Curtin University*

[4] *School of Health and Welfare, Halmstad University*

Author Notes

*Address correspondence to Daniel Gucciardi, Curtin School of Allied Health, Curtin University, GPO Box U1987, Perth, Australia, 6845. Email: daniel.f.gucciardi@gmail.com

Gucciardi, D.F., Lines, R.L.J., & Ntoumanis, N. (in press). Handling effect size dependency in meta-analysis. *International Review of Sport and Exercise Psychology*. doi: 10.1080/1750984X.2021.1946835

Abstract

The statistical synthesis of quantitative effects within primary studies via meta-analysis is an important analytical technique in the scientific toolkit of modern researchers. As with any scientific method or technique, knowledge of the weaknesses that might render findings limited or potentially erroneous as well as strategies by which to mitigate these biases is essential for high-quality scientific evidence. In this paper, we focus on one prevalent consideration for meta-analytical investigations, namely dependency among effects. We provide readers with a non-technical introduction to and overview of statistical solutions for handling dependent effects for their efforts to integrate evidence within primary studies. This goal is achieved via a series of seven reflective questions that scholars might consider when planning and executing a meta-analysis in which some degree of dependency among effect sizes from primary studies may exist. We also provide an example application of the recommendations with real-world data, including an analytical script that readers can adapt for their own purposes.

Keywords: effect sizes; multivariate meta-analysis; open science; research synthesis; robust variance estimation; systematic review; three-level meta-analysis

Handling Effect Size Dependency in Meta-Analysis

Systematic reviews of the literature, particularly of experimental or interventional research in the lab or field, and the statistical integration of individual effects from eligible primary studies via meta-analysis are typically considered the highest form of evidence for theory, practice, and policy (Chan & Arvey, 2012; NHMRC, 2009). As with any research method or statistical analysis, however, systematic reviews and meta-analyses are not immune to potential biases (e.g., inadequate search strategy, absence of methodological quality assessments) that may render the findings of such studies limited or potentially erroneous (Shea et al., 2017). It is therefore essential that scholars consider these potential sources of bias in the search, data extraction, and data analysis when planning and executing a systematic review and meta-analysis. There exists several high-quality guidelines (e.g., (Guyatt et al., 2008; Liberati et al., 2009; Moher et al., 2009; Steel et al., 2020) and tutorials (e.g., Field & Gillett, 2010; Moreau & Gamble, 2020; Quintana, 2015) for readers interested in these considerations, including the excellent contribution of Hagger and Chatzisarantis to this special issue. In this article, we focus on one key consideration for the analytical stage, namely the handling of dependent effect sizes because non-independence of data points is a core assumption of most meta-analytical techniques (Cheung, 2014; for a general review of statistical non-independence, see Forstmeier et al., 2017). Our objective is to provide readers with a non-technical introduction to and overview of statistical solutions for handling dependent effects for their efforts to integrate evidence within primary studies.

Meta-Analysis and Non-Independent Effects

At its core, meta-analysis enables researchers to cumulate and statistically summarize quantitative evidence across large bodies of work to provide answers to important questions. As relatively broad examples, researchers might be interested in clarifying the magnitude of an association between two variables, the effectiveness of interventions designed to change human behaviour, the psychometric properties of questionnaires, or the veracity of theoretical models including several factors and their underlying propositions (e.g., mediation pathways). Key strengths of meta-analysis include the incorporation of evidence drawn from multiple primary studies and the differential weighting of effects from each study to minimise sampling error and maximise precision in the summarised effect; estimation of the amount of heterogeneity in the overall effect and factors that (partly) explain this variation; adjustments for artefacts that may bias the mean effect (e.g., measurement error, range restrictions); and tests of complex models that remain untested or are impractical to test in individual primary studies (e.g., large samples requires for statistical power; Borenstein et al., 2009; Hunter & Schmidt, 2004). For these reasons, meta-analysis represents an important analytical technique in the scientific toolkit of modern researchers who are inspired by a positivist (i.e., there exists an objective, true reality that can be known and examined objectively) or post-positivist (i.e., there exists an objective, true reality that can only be estimated rather than known perfectly) paradigmatic position on science (Creswell & Poth, 2016)[1].

Among the potential sources of bias for meta-analytical investigations, dependency among effects represents a key consideration because it is a widespread problem and handling it in a suboptimal fashion violates a core assumption of most statistical techniques (Cheung, 2014; Forstmeier et al., 2017; Jackson et al., 2011). Violating such assumptions ultimately negatively affects the quality of the meta-analytic output for knowledge accumulation (e.g., ‘garbage in, garbage out’; Sharpe & Poets, 2020). Briefly, effects are considered ‘dependent’ when they are related in some way. Dependency in effects can arise within the primary studies from which statistical evidence is drawn or can be introduced by the meta-analyst (Cheung, 2014). With regard to primary studies, researchers may utilise

[1] A detailed analysis of the philosophical foundations of meta-analysis is beyond the scope of the current article. Interested readers are referred elsewhere for critiques (Stegenga, 2009, 2011). Given the numerous subjective decisions that are applied in the design and execution of meta-analyses, we prioritise a post-positivist paradigmatic perspective.

multiple outcomes (e.g., indices of psychological well-being and ill-being) or multiple measures of the same outcome (e.g., speed and accuracy of performance) to assess the effectiveness of an intervention; compare two or more different types of interventions (e.g., biofeedback versus cognitive restructuring) or adaptions of the same intervention (e.g., frequency of biofeedback sessions) against a comparator group; or assess associations among variables or differences between groups across several time points. Equally, researchers might introduce dependency via the substantive focus of the meta-analysis; for example, one might be interested in statistically summarising cross-cultural differences in the associations between two variables or the effectiveness of a certain type of intervention (e.g., mindfulness), such that studies conducted within the same country are likely to be more alike than studies from other countries (for an excellent exposition of psychological and cultural distance, see Muthukrishna et al., 2020). Treating dependent effects as independent underestimates the standard errors of the overall effect and therefore biases any statistical conclusions from the analysis (López-López et al., 2017).

If dependent effects are prevalent in the literature (Ahn et al., 2012), yet they undermine the statistical conclusions of meta-analytic findings (López-López et al., 2017), how then have scholars typically handled them in the past? Leading meta-analysts have identified several approaches, each of which comes with its own strengths and weaknesses (Borenstein et al., 2009; Cheung, 2014). First, researchers might execute several separate meta-analyses to analyse dependent effects as if they were independent. For example, an analyst might be interested in the effects of stress regulation interventions on athlete performance, yet conduct separate meta-analyses for different categories of performance (e.g., cognitive, physical) or indices of each type of performance category (e.g., speed and accuracy of decision-making for cognitive performance). This approach is practically appealing for its simplicity, yet ignores the conceptual and statistical overlap between effects thereby underestimating the standard errors that results from the dependency of the effects and inflating Type I error (Cheung, 2014).

Second, scholars might utilise an aggregation approach in which they remove dependence among multiple effects within the same study by creating an averaged or pooled estimate (Cheung & Chan, 2004; Rosenthal & Rubin, 1986). Remaining with the previous example, a researcher might combine effects obtained for speed and accuracy of decision-making into a single effect to reflect cognitive performance. This approach is advantageous because the effect sizes and their variances can be utilised within a univariate meta-analysis and precision is optimised when the effects reflect the same concept, yet this approach reduces the pool of available data (think ‘sample size’) and therefore limits statistical power, precision of the estimates, and the types of questions that can be tested via meta-analysis (Cheung, 2014). One also needs to consider the nature of the dependent effects for which they wish to aggregate as a pooled estimate, particularly with regard to the direction of effects. For example, the inverse association between speed and accuracy, otherwise known as the ‘speed-accuracy trade-off’, is ubiquitous in the behavioural sciences (for a review, see Heitz, 2014), such that their combination as a pooled estimate will inevitably be biased towards zero.

Third, analysts might prefer an elimination approach in which they select one effect among several effects for each primary study. The elimination approach might be the preferred strategy in some cases to avoid the ‘garbage in, garbage out’ criticism. For example, analysts might prefer to utilise device-based measures of physical activity rather than subjective self-reports because of the high cognitive demands required to recall different intensities accurately (Lines, Ntoumanis, et al., 2020). However, the elimination approach is subject to the same limitations as the aggregation approach (M. W.-L. Cheung, 2014). In particular, the magnitude and direction of the association between dependent effects is an important consideration; for example, randomly selecting one effect size per study when the dependent effects are negatively correlated may lead to biased results.

Finally, analysts might average effect sizes according to a shift in the unit of analysis (Cooper, 2016), that is, strategic averaging of effects. As an example, when the interest is an estimate of the overall pooled effect for the effectiveness of stress regulations, researchers might average all of the effect sizes within a single study (e.g., performance, well-being) to calculate a single effect that can be utilised within a univariate meta-analysis. Conversely, if the interest is on the effect of stress regulation interventions on performance and well-being, then meta-analysts might average the effects for multiple measures of each domain to disaggregate effects across conceptually distinct outcome categories. Nevertheless, like the aggregation method, the shifting unit of analysis approach limits the overall pool of data and conceptual information, and presumes within-study homogeneity (Cheung, 2014). With few exceptions (e.g., selecting a widely accepted ‘gold standard’ measure for assessments of intervention effectiveness), these common approaches to handling dependent effects are often suboptimal and therefore require researchers to leverage statistical techniques that account for dependency as part of the analytical framework.

Solutions for Handling Dependent Effects

Meta-analysts can utilise one of three statistical techniques to account for dependency among effect sizes extracted from primary studies in their analytical framework, namely multivariate meta-analysis (Cheung, 2013; Kalaian & Raudenbush, 1996), robust variance estimation (Hedges et al., 2010), and three-level meta-analysis (Cheung, 2014). Essentially, the decision regarding which of these options is most suitable depends on the information available to those who conduct the meta-analysis regarding the dependency among effects. When the correlations among dependent effects are available, researchers can model this dependency directly as part of the statistical model and, therefore, retain all available effects using a multivariate meta-analysis. For example, analysts interested in the usefulness of a psychological intervention may use a multivariate meta-analysis to produce summary effects separately for cognitive and physical performance obtained from the same participants while accounting for the correlation between these indices of 'performance'[2]. However, multivariate meta-analysis is largely uncommon in practice because correlations among dependent effect sizes are typically unreported in primary research (Ahn et al., 2012; Becker, 2000). For example, researchers might have reported the mean difference between two groups on subjective and objective indices of performance (e.g., Cohen's d), yet excluded the degree of association between these two effects in the reporting of their results. In cases when correlations among within-study effects are unknown, analysts can impute approximated values (e.g., based on simulations), then perform sensitivity analyses to determine the influence of the varying degrees of magnitude of the correlation (Berkey et al., 1996; Raudenbush et al., 1988; see also Riley et al., 2008). An alternative approach to accounting for dependency among effects within primary studies include multilevel models, where variance is partitioned across three levels (i.e., Level 1 = sampling variance, Level 2 = within-study variance, Level 3 = between-study variance)[3]. A third approach is robust variance estimation in which dependence is approximated via a working model of hierarchical (i.e., dependence is due to common features of a research group with independent effects) or correlated effects (i.e., dependence is due to the use of the same sample for unique effects; Tipton & Pustejovsky, 2015). Simulation studies support the utility of these two approaches. For example, three-level meta-analysis and robust variance estimation methods result in unbiased estimates, with the latter offering the best performance

[2] Network meta-analysis is another type of multivariate approach in which analysts compare multiple interventions against a common comparator group. Interested readers are referred elsewhere for overviews of network meta-analysis (Molloy et al., 2018; Salanti, 2012).

[3] Traditional univariate meta-analysis is technically multilevel in nature with participants in primary studies located at level 1 and the studies at level 2.

for small samples (López-López et al., 2017; Moeyaert et al., 2017; Park & Beretvas, 2019). Further, three-level models perform similarly to multivariate meta-analysis in terms of accuracy of parameter and standard error estimates (Van den Noortgate et al., 2013).

To what extent are dependent effects in primary research an important consideration for statistical syntheses in the field of sport and exercise psychology? Looking back at existing meta-analyses in this field offers us one perspective to answer to this question. We manually searched core peer-reviewed outlets from their inception to September 3rd 2020 to locate existing meta-analyses in the field of sport and exercise psychology, including *Journal of Sport and Exercise Psychology*; *Psychology of Sport and Exercise*; *Journal of Applied Sport Psychology*; *The Sport Psychologist*; *International Review of Sport and Exercise Psychology*; *Sport, Exercise, and Performance Psychology*; and *International Journal of Sport and Exercise Psychology*. An overview of the 39 meta-analyses identified via this search is provided in an excel file located on the Open Science Framework (https://osf.io/cqywh/). Briefly, dependent effect sizes in primary studies were an issue in 37 (~95%) of these meta-analyses because of multiple outcomes (n = 26), multiple groups (n = 2), multiple time points (n = 2), multiple outcomes and time points (n = 3), multiple groups and time points (n = 3), or multiple outcomes, groups, and time points (n = 1). Despite the prevalence of dependent effects, only two meta-analyses accounted for this dependency via robust variance estimation or a multivariate fixed effects model. This finding is unsurprising because knowledge of methods for handling dependency and software to execute these analyses were largely unavailable a decade or so ago. Common approaches to handling effect size dependency included executing separate meta-analyses (n = 12), not reporting any explicit information regarding dependency (n = 9), computing an averaged effect size (n = 8), or selecting a single effect from primary studies (n = 3). Of course, we acknowledge that we are not immune to such suboptimal practices in a past meta-analysis we conducted (Carr et al., 2019), yet have subsequently improved our knowledge of these issues and applied them in our work (Lines, Pietsch, et al., 2020). As applied scientists, we fully appreciate the challenges of remaining abreast of statistical and methodological advancements, particularly as they appear rapidly, so we hope the guidelines presented in the remainder of this article will make this information more widely known among our field.

Best Practice Guidelines for Handling Dependency among Effect Sizes in Meta-Analysis

In this section, we present a series of reflective questions that scholars might consider when planning and executing a meta-analysis in which some degree of dependency among effects sizes from primary studies may exist. These reflective questions are largely based on our own experiences completing meta-analyses, so it is important to acknowledge that the list is non-exhaustive and the application of these questions may differ depending on one's preferences. Readers are referred elsewhere for detailed expositions and guidelines for planning and executing a meta-analysis (e.g., Johnson & Hennessy, 2019; Pigott & Polanin, 2020; Rudolph et al., 2020).

Question 1: Will There be a Need to Accommodate Non-Independence in Effect Sizes?

Once researchers have decided that a meta-analysis is suitable for the purposes of their research questions, it follows that one needs to determine whether or not they have or will likely have (because of analysts' decisions) dependency in effect sizes and, if so, how best to account for this dependency. The nature of the research question is a fundamental consideration because 'multiplicity' in effects might be present in varying degrees depending on the breadth or narrowness of the primary goal of the meta-analysis (López-López et al., 2018). For example, a meta-analysis concerned with the effectiveness of biofeedback training relative to practice as usual on subjective experiences of state anxiety assessed using the relevant subscale of the revised Competitive State Anxiety Inventory-2 tool (Cox et al., 2003) is likely to have no effect size dependency because each primary study will contribute one effect to this research question. Compare this example with one where the outcome variable is broadened to focus on athletic performance; in this scenario, it is possible that performance is assessed in multiple ways within primary studies (e.g., speed and accuracy, cognitive and physical) and, therefore, there would be a requirement to accommodate dependent effects in the statistical analysis.

Analysts also might introduce dependency among effects via the primary research question. As an example, researchers might be interested in cultural considerations regarding the associations between psychological (e.g., moral identity) and social-contextual factors (e.g., coach motivational climate) and anti-doping attitudes among athletes around the globe. In this scenario, meta-analysts could incorporate country or culture as a level within the meta-analytic model. Expectations regarding dependency among effects within primary studies or introduced by the analyst are best documented as part of one's protocol registration, or acknowledged as a consideration if one is unsure of the potential for dependency in the meta-analysis (and propose the conditions under which different solutions will be considered). For example, robust variance estimation performs best when sample sizes are small, whereas multilevel models offer the greatest flexibility for interrogations of factors that moderate the pooled effect (López-López et al., 2017; Moeyaert et al., 2017). López-López and colleagues (2018) provide an informative decision tree regarding key decisions for handling dependency among effects or 'multiplicity' as part of meta-analysis that can be used to inform protocol registrations.

Question 2: What is the Effect Size of Interest?

Meta-analyses are used to statistically synthesise effect sizes reported in primary research. In scientific research, an effect represents a quantitative summary of the magnitude and direction of findings from a statistical test (Flora, 2020; Funder & Ozer, 2019; Pek & Flora, 2018). The scale of an effect is fundamentally important for the execution and interpretation of a meta-analysis (Cheung, 2015a). Unstandardized effects are typically preferred when the unit of measurement is intuitively or practically meaningful (e.g., minutes of sleep per night), whereas standardized effects are often utilised in scenarios where measurement is unintuitive (e.g., proprietary scale) or disparate (e.g., quantitative differences in assessment scales) (Borenstein et al., 2009; Pek & Flora, 2018). Broadly speaking, effects can be categorised as a summary of binary outcomes, such as pass or fail on some type of test (e.g., odds ratio); an association between two variables, such as strength and magnitude of the relation between cognitive ability and test performance (e.g., correlation coefficient); or a difference between two variables, such as the effect of an intervention on a primary outcome between experimental and control groups (e.g., Cohen’s d) (Borenstein et al., 2009). Thus, an important question for scholars to consider is which category and type of effect size is most relevant to their meta-analysis.

The answer to this question depends on both substantive and practical considerations of the meta-analysis and the existing body of literature. Perhaps most important, it is essential that the effect size used to summarise findings of primary studies is appropriately aligned with the research question of the meta-analysis. In other words, is the primary focus for your meta-analysis designed to assess some sort of binary outcome, association, or difference between groups (Borenstein et al., 2009; Cooper, 2016)? When interested in a binary outcome, for example, meta-analysts might consider effects such as an odds ratio (e.g., % increase in passing a selection test based on psycho-physiological determinants; Gucciardi et al., 2020) or relative risk (e.g., differences in the prevalence of mental ill-health symptoms between groups; Wolanin et al., 2016) to summarise the statistical results of primary studies. Other considerations for selecting the most appropriate effect size of interest include a) the comparability of effect sizes across studies, noting that statistics from different designs can be converted into a common effect size (e.g., log odds ratio can be converted to Cohen’s d;

Borenstein et al., 2009), b) availability of statistical information in primary studies to compute effect sizes and their sampling distributions, c) incorporation of the magnitude and direction of effect (e.g., R^2 represents the portion of variance explained in regression, yet is inadequate because it excludes directional information), and d) relative independence from sample size in that the effect does not increase or decrease as a function of the number of participants in the study (Borenstein et al., 2009; Cheung, 2015a). In our experience as editors, reviewers, and consumers of the sport and exercise psychology literature over the past 20 or so years, the most commonly reported effects in primary research are correlations, (standardized) mean differences, and odds ratios and therefore the ones we recommended for meta-analysts.

Question 3: How Do I Calculate Effect Sizes from Primary Studies?

The calculation of effect sizes from primary studies is one of the most critical tasks for any meta-analyst because errors at this stage can have profound effects for the overall results of the statistical synthesis and the conclusions made from those findings (Maassen et al., 2020). Fortunately, there exists numerous resources to assist novice and experienced meta-analysts with this important task, including books (e.g., Borenstein et al., 2009; Cheung, 2015; Hunter & Schmidt, 2004), journal articles (e.g., Caldwell & Vigotsky, 2020; Durlak, 2009), online calculators (Wilson, n.d.), excel calculators (e.g., Lakens, 2013), and packages within statistical software (e.g., *escalc* function of the *metafor* package in R; Viechtbauer, 2010). We have prepared an excel file for the calculation of common effects in the sport and exercise psychology literature, which is available on the Open Science Framework (https://osf.io/pa375/). At this point, it is worth highlighting that scholars would need to input as a minimum the effect size and its sampling variance to perform a meta-analysis so as to incorporate an estimate of precision into the model; the absence of the sampling variance makes an effect size essentially useless for meta-analysis because it introduces one source of bias (Cheung & Vijayakumar, 2016). The effect size of interest can be obtained in its original form (e.g., Cohen's d, r) or computed from test statistics (e.g., *t* or *F* values). Scholars are encouraged to report the methods or formulas they used to compute effect sizes from primary studies as part of their published report or supplementary material to facilitate reproducible knowledge accumulation (Maassen et al., 2020). It is often the case that information required to compute effect sizes is absent from reports on primary studies, such that analysts will need to contact corresponding authors for additional information to enable them to calculate effects. We also recommend that researchers account for this likelihood in their protocol registration by indicating their approach to requesting information from authors and decision regarding when to cease such requests (e.g., contact corresponding authors twice, with each communication separated by two weeks). All papers that are excluded at this stage are typically provided in a supplementary file with a rationale for why they have been excluded from the review (i.e., no data were provided by the authors). Of course, it is also worth acknowledging that analysts are likely to receive information from around 20-25% of requests (Polanin & Terzian, 2019).

Question 4: How Do I Structure My Data File for a Meta-Analysis that Accounts for Non-Independence in the Data?

As with any statistical methodology, it is essential that researchers structure their data file in a way that permits the incorporation of dependency among primary effects within the main analysis. Most statistical programs require that data be structured in long format, where each row represents a unique observation (effect) within a specific dependency category (e.g., one effect size extracted from a single primary study). Exceptions to this typical trend is for multivariate meta-analysis and structural equation modelling, which require that data be structured in wide format. We provide an excel template on the Open Science Framework for researchers who have calculated the unique effects manually and therefore have a numerical value for each effect and its sampling variance https://osf.io/bewn6/). There are four key elements for the data file, regardless of which approach one adopts to handle dependency in the data, namely a unique identifier for each primary study (e.g., studyID) and effect size in the dataset (e.g., esID), and the numerical values for the effect size (e.g., y) and its sampling variance (e.g., v). For robust variance estimation and multilevel meta-analysis, each row will contain one effect size and one sampling variance value; in contrast, multivariate meta-analysis requires a ‘pair’ of effects (e.g., y1 and y2) and their sampling variances (e.g., v1 and v2) per row, alongside the covariance among this pair of effects (e.g., cov). Analysts also can incorporate additional columns for variables that will be tested as moderators of heterogeneity of the pooled effect, including categorical (e.g., pre-registered protocol: 0 = no, 1 = yes) and continuous factors (e.g., average age of participants in sample). In essence, the key distinction in the data file structure between a traditional univariate meta-analysis and one containing non-independent effects is the inclusion of columns that identify this dependency (e.g., studyID and esID).

Question 5: What Statistical Program Should I Use to Conduct a Meta-Analysis That Handles Dependency Among Effects?

There exists several software programs in which researchers can execute meta-analyses. With regard to general statistical environments, analysts can utilise programs such as Stata (StataCorp, 2019), SAS (SAS Institute Inc, 2016), M*plus* (Muthén & Muthén, 2017), R (R Development Core Team, 2018), JASP (JASP Team, 2020), or jamovi (The jamovi project, 2020). Alternatively, one can use dedicated programs like Comprehensive Meta-Analysis (https://www.meta-analysis.com/), Meta-Analyst (Wallace et al., 2009), Cochrane’s Review Manager (or RevMan) (http://revman.cochrane.org/), and Meta-Essentials (Suurmond et al., 2017) to execute a meta-analysis. Scholars have compared the features of many of these programs (Bax et al., 2007). Researchers tend to make these decisions with consideration of the research question, software program capabilities, statistical capabilities, and financial resources. Analysts can implement multivariate meta-analysis via the *mvmeta* function in Stata (White, 2011) and the *metafor* (Viechtbauer, 2010) or *metasem* (Cheung, 2015b) packages in R; robust variance estimation via the *robumeta* (Fisher & Tipton, 2015) or *clubSandwich* (Pustejovsky, 2020) packages in R, or macros developed for SPSS and Stata (for a tutorial, see Tanner-Smith & Tipton, 2014); and multilevel meta-analysis via the *metafor* (Viechtbauer, 2010) or *metasem* (Cheung, 2015b) packages in R, M*plus* (Muthén & Muthén, 2017; for a tutorial, see Cheung, 2019), and PROC MIXED in SAS® (Stroup et al., 2018). We personally believe R (R Development Core Team, 2018), which is a free software environment for statistical computing and graphics, provides analysts with the greatest coverage and flexibility to execute meta-analyses using any of the approaches outlined here to handle non-independent effects. Yet, it can be a 'daunting' program for applied scientists with little to no experience with programming (for a review of R packages for meta-analysis, see Polanin et al., 2017)[4]. Nevertheless, excellent tutorials are available on conducting and interpreting the results of meta-analyses for readers unfamiliar with R (M. W.-L. Cheung, 2015b, 2019; Harrer et al., 2019; Quintana, 2015; Tanner-Smith et al., 2016). There also is an increasing trend for meta-analysts to make their R scripts available via the Open Science Framework (https://osf.io), a practice that provides readers with real-world examples of the translation of concepts into analytical code (e.g., Moreau & Chou, 2019; Steffens et al., 2019). Personally, we have learned a great deal from other researchers who have made their analytical scripts with instructional annotations freely available to the research community.

Question 6: How Do I Interpret the Results of a Three-Level Meta-Analysis?

[4] Open source programs such as Jamovi (The jamovi project, 2020) and JASP (JASP Team, 2020) have capabilities for researchers interested in conducting univariate meta-analyses with 'point and click' software. It is likely that multilevel meta-analysis will be incorporated in future updates to these programs (e.g., https://github.com/kylehamilton/MAJOR/blob/master/README.md).

You've likely spent the best part of several months screening articles to identify primary studies using your inclusion and exclusion criteria, and extracting all relevant information (and/or requesting that data from authors) to compute individual effects for your meta-analysis (Borah et al., 2017). Now comes the fun part of data analysis and interpretation! The exact nature of the data analysis output available to you will depend on the statistical software or package that you use to conduct the meta-analysis. Nevertheless, there are several elements common to most statistical programs that one needs to consider as part of the interpretation and reporting processes. We focus here on three-level meta-analysis because correlations among dependent effect sizes are typically unreported in primary research (Ahn et al., 2012; Becker, 2000) and therefore it is the framework most likely available to sport and exercise psychology researchers.

Overall pooled effect. As with a univariate meta-analysis, the first step for a multilevel meta-analysis is to determine the overall pooled effect across all eligible studies. This estimate is typically located in a section called "model results" or something similar, and is accompanied by an estimate of its variance (e.g., standard error), confidence intervals, and test statistic (e.g., *t* value) and its *p* value. An important quality check at this point is whether your output is based on the expected number of effect sizes utilised for the overall analysis (typically denoted by *k*). It is also likely that the model summary section contains one or more goodness-of-fit indices, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) that provide information regarding the degree to which the meta-analytic model fits the data; these indices can be used alongside likelihood ratio tests for comparing and selecting models in subsequent stages of the analytical process if required (interested readers are referred elsewhere for a technical discussion of these indices; Cheung, 2015a; Vrieze, 2012). The key distinction between a traditional univariate and three-level meta-analysis for the output of the overall pooled effect is the presentation of variance components distributed across level 2 (within-study variance) and level 3 (between-study variance) of the model alongside the sampling variance of individual effects (level 1); in a traditional univariate meta-analysis, only one variance component (between-study) is reported. Typically, the number of effect sizes (level 2) and studies (level 3) utilised for the variance components are presented in the output, so it is recommended that you check these values to ensure they align with your expectations from the main data file.

Heterogeneity of effects. The overall pooled effect provides a statistical summary of the body of work included in your meta-analysis, yet it provides no information on the degree to which these effects are homogenous. The question of homogeneity among effects is often a primary consideration for meta-analysts (Borenstein et al., 2017). Suffice to say, summary effects are best interpreted with caution when there is substantial heterogeneity among effects. For a traditional univariate meta-analysis, the Q test provides information on the binary decision regarding the presence or absence of heterogeneity, whereas the I^2 statistic quantifies the proportion of dispersion of effects that cannot be attributed to sampling variance alone (Higgins & Thompson, 2002; Huedo-Medina et al., 2006). We believe an intuitive way to appreciate the essence of the I^2 statistic is as an indication of “the amount of non-overlap among confidence intervals” (Borenstein et al., 2017, p. 7). Defined in this way, the I^2 statistic provides no information on the absolute amount of variability among effects; if this metric is of interest, then analysts need to compute the prediction interval so that they can make inferences regarding the 95% likelihood that an effect in future similar studies will fall between some range (Borenstein et al., 2017; IntHout et al., 2016). The key distinction between a traditional univariate and three-level meta-analysis is the decomposition of the total heterogeneity (I^2 statistic) across levels, such that there exists within-study heterogeneity (I_2^2) and between-study heterogeneity(I_3^2) or levels 2 and 3, respectively. As we have partitioned variance across two levels, it is important that we assess the degree to which this model complexity introduced by the meta-analyst is meaningful. To do so, one can utilise likelihood-ratio tests to compare the full model where the variances are freely estimated with separate models in which the within-study or between-study variance is fixed to zero (Assink & Wibbelink, 2016). However, the application of likelihood-ratio tests is somewhat complex in practice because tests of variance assume the variance parameter is zero in one of the models (i.e., the lower boundary of possible values), yet the likelihood-ratio test assumes negative values are possible making the asymptotic (χ^2) distribution no longer df = 1 (Andrews, 2001). The presence of substantial heterogeneity (*Q* test) and meaningful proportional dispersion within studies and/or between studies provides sufficient justification for the examination of potential moderators of the overall effect.

Question 7: What Factors Might Explain the Heterogeneity Among Effects?

Heterogeneity is a core consideration for psychological science because it contributes substantially to the inconsistencies in findings, often described as characteristic of the replication crisis (Kenny & Judd, 2019; McShane et al., 2019; Stanley et al., 2018). When there exists meaningful heterogeneity of the overall pooled effect, meta-analysts are advised to examine factors that might help to explain this heterogeneity. We typically approach this task as one that is focused on sensitivity analyses to assess the robustness and consistency of the findings (e.g., outliers, publication bias) and moderator analyses to examine factors that are of substantive interest or practical value (e.g., intervention dosage, number and type of behaviour change techniques; for an example, see Ntoumanis et al., 2020). Readers are referred elsewhere for best practice recommendations regarding the use of meta-regression for testing boundary conditions within the context of traditional two-level meta-analysis (Gonzalez-Mulé & Aguinis, 2018).

Outliers and influential cases. As a logical extension of traditional regression models (e.g., linear, ANOVA), it is important to check for the presence and influence of outliers and influential cases on the pooled effect within a meta-analysis (Viechtbauer & Cheung, 2010). Briefly, outliers represent statistically extreme values, whereas influential cases are those effects that alter the substantive meaning of the overall pooled effect (e.g., a small and statistically inconsequential pooled effect that becomes moderate and meaningful when an effect is excluded). Visual inspection of the individual effects grouped by a higher-level of the model (e.g., study) via a forest plot is an appropriate starting point in this regard. Meta-analysts can formally test for the presence of outliers and influential cases via residuals in relation to their standard error (e.g., > ± 1.96 bounds, noting that roughly 5% of a sample is typically expected to exceed these bounds); the identification of individual effects whose confidence interval excludes the range captured by the confidence interval of the pooled effect; Baujat plot (Baujat et al., 2002); and case deletion diagnostics (Harrer et al., 2019; Viechtbauer & Cheung, 2010).

There are two key distinctions between a univariate and multilevel meta-analysis in this regard. First, one could focus their attention on individual effects and/or individual studies in a multilevel meta-analysis rather than individual studies as is the case for a univariate meta-analysis. For example, a coach-reported measure of performance in an individual study may represent an extreme or influential effect, whereas an objective measure (e.g., competition race time) in the same study might not be an extreme or influential case. Second, one can examine the changes in the overall pooled effect alongside the partitioning of heterogeneity across within-study and between-study levels; for example, the exclusion of outliers and/or influential cases may influence heterogeneity for one level more so than the other. There are no set rules as to which tests one should include as part of an assessment of outliers and influential cases. Broadly speaking, a multifaceted approach is ideal; if multiple tests support the presence (or absence) of outliers and/or influential cases, then one has lesser or greater confidence in the robustness of the overall pooled effect than relying only on one test alone.

Meta-bias. It is well-accepted that meta-analysts need to assess and report on the extent to which the pooled effect might be over- or underestimated because of elements of the scientific process (e.g., comprehensiveness of search, quality of eligible studies; Borenstein et al., 2009; Johnson & Hennessy, 2019; Pigott & Polanin, 2020). As with outliers and influential cases, it is important to distinguish between individual effects (level 2) and studies (level 3) when assessing bias in the body of work. One useful way to think about this distinction is variation among individual effects within a study in some capacity, and variation between studies in the dataset. Consider the case of methodological quality, which can be operationalised and statistically tested across multiple levels in the analysis. Primary studies might include multiple indicators of performance outcomes to assess the effectiveness of some type of intervention. The quality with which these performance indicators are operationalised may vary within individual studies and between studies in the entire dataset. For example, individual studies might include a high-quality assessment of performance using objective metrics (e.g., race time) alongside a low-quality informant assessment of performance using a single, self-reported item (e.g., “All things considered, rate the quality of [participant’s name] performance on this task”). Equally, the study as a whole can be evaluated in terms of its methodological quality using established frameworks like the Risk of Bias 2 for randomised controlled trials (Sterne et al., 2019).

Publication bias is one of several major considerations in this regard because it encompasses scenarios where the literature published in peer-reviewed journals is systematically unrepresentative of the population of completed work (Mathur & VanderWeele, 2020; Sterling, 1959). The primary consideration for publication bias is the size and/or statistical significance of the estimated effects (Dickersin, 1990). As a first look at publication bias, meta-analysts can quantify the magnitude and meaningfulness of effect size differences via meta-regressions in which the overall effect is regressed on key factors such as sample size, publication type (peer-reviewed versus unpublished, such as PhD dissertations or full conference proceedings), and study quality (e.g. risk of bias assessments, such as RoB2[5] for randomised controlled trials; Sterne et al., 2019). Formally, there exist several approaches for examining the likelihood of publication bias in meta-analyses. Examples include visual inspection of funnel plots that include individual effects plotted against their standard error or precision (inverse of the standard error), where asymmetry is consider representative of bias (Sterne & Egger, 2001); contour-enhanced funnel plots that include lines corresponding to levels of statistical significance (Peters et al., 2008); power-enhanced funnel plots that include estimates of statistical power relative to an expected true value of the effect (Kossmeier et al., 2020); statistical test of the funnel plot via Egger's regression test (Egger et al., 1997) that has been extended to multilevel contexts where the overall pooled effect from the three-level model is regressed on some function of the standard error of effect sizes (Fernández-Castilla et al., 2021)[6]; precision-effect test-precision-effect (PET-PEESE) that involves a weighted regression of individual effects on their standard error (PET) or the sampling variance weighted by study precision (PEESE), where a positive slope indicates some degree of bias (Stanley & Doucouliagos, 2014); *p*-curve analysis to assess the evidential value of the body of work via a distribution of statistically significant findings only, where a left-skewed curve indicates possible bias and a right-skewed supports evidential value (Simonsohn et al., 2014). Simulation evidence indicates the multilevel level

[5] Interested readers are referred elsewhere for an excellent R package and Shiny web app to visualise risk of bias (McGuinness & Higgins, 2020).

[6] Ultimately, the approach one adopts to extending Egger's test beyond the traditional two-level meta-analytic framework depends on the nature of the effect sizes adopted in one's analysis. For example, there is an inherent association between correlations and their sampling variances, so one may prefer to use the inverse of the sample size as the predictor in the moderation analysis (https://stat.ethz.ch/pipermail/r-sig-meta-analysis/2020-May/002086.html).

extension of Egger's test provides an acceptable, valid test with regard to Type I error, but has limited power to detect small study effects (Fernández-Castilla et al., 2021; Rodgers & Pustejovsky, 2020). Nevertheless, it is important to bear in mind that the usefulness of the other methods or approaches for meta-analyses involving dependent effects currently remains unknown. Informed by simulation work with traditional two-level meta-analytic statistical models (Carter et al., 2019), we suggest the best approach in the meantime is one that is multifaceted in nature and occurs alongside sensitivity tests.

Substantive moderators of the pooled effect. Is a meta-analytic effect stronger for some people or in certain contexts than it is for others? Statistical interrogations of pooled effects with regard to social-contextual (e.g., active ingredients of an intervention, social norms) or individual (e.g., psychological or biological dimensions) factors provide knowledge of important building blocks for theory development, refinement, and elaboration (Gonzalez-Mulé & Aguinis, 2018; Tipton et al., 2020). From an application standpoint, knowledge of what works, for whom, and under what conditions is essential for translating scientific knowledge meaningfully into practice and policy. For example, scholars often employ different tools to assess the same concept (e.g., Lonsdale et al., 2014) or adapt existing tools with regard to their length, context, or content to suit the needs of their research (Heggestad et al., 2019; Marsh et al., 2010; Pelletier et al., 2013). Ideally, factors considered as potential moderators are informed by conceptual or empirical expectations and outlined in one's preregistered protocol; of course, we acknowledge the types and breadth of moderators tested ultimately relies on information available in primary studies. The salience of such moderators can be examined via the quantification of the percentage of variance explained by the predictors of the pooled effect across levels 2 and 3 of the statistical model (M. W.-L. Cheung, 2014). Meta-analysts might also consider the interaction among multiple moderator variables whereby one factor (e.g., delivery mode of intervention such as virtual versus face-to-face) might amplify or attenuate the effect of another variable (e.g., number of behaviour change techniques present in an intervention). The R function *metacart*, for example, leverages tree-based models to facilitate variable selection, handle non-linear associations, and retain ordering information of moderator variables (Li et al., 2020).

Tutorial on Applying Best Practice Recommendations for Handling Dependency among Effect Sizes in Meta-Analysis

To assist readers in applying the recommendations outlined in this article, we provide a brief tutorial using a subset of one of our published datasets (Lines, Pietsch, et al., 2020). The data file, annotated R script, and output files are located on the Open Science Framework (https://osf.io/bzkg2). We encourage readers to have the R script (https://osf.io/8bhzu/) and output file (https://osf.io/6awvy/) open when working through the following sections. Briefly, we identified randomised experiments via a systematic review of the literature and extracted data from primary studies to assess the effectiveness of team reflexivity interventions on collective performance outcomes and behaviours, and factors that augment their effects. We focus on performance behaviours here, which represent those cognitive, affective, and behavioural actions that precede or determine goal attainment (e.g., coordination, communication). Readers are referred to our protocol registration (https://osf.io/3bv6y/) and published article for full details on the processes related to questions 1-4 outlined above; we focus on question 5-7 here.

Structure of Data File

The data are stored in long format, where rows represent individual effect sizes and columns capture unique elements for the analyses (https://osf.io/btxek/). When creating data files for execution in R in other programs like Excel, it is essential that the same character text is applied for cells that replicate elements in the data set (e.g., the descriptive labels for the author and year in column A) because R will treat discrepancies in character text as different factors. One common mistake that we have learned through trial and error is the inclusion of a ‘space’ at the end of a descriptive label, which visually is undetectable (e.g., ‘Dyas (2018)’ is treated as different to ‘Dyas (2018) ’). Except for ‘outcome_type’ and ‘outcome_method’, all moderators reflect between study (level 3) differences.

Overall Pooled Effect

We wrote a function in R to extract the key findings to a text output file so that readers can follow this section with ease. The first section of this output file is the overall model results (see Figure 1). Under ‘model results’, we can see that the overall pooled effect was moderate in magnitude (*g* =.55, 95% CI = .32, .77, *p* < .001) and that variability in observed effects is larger than one would expect based on sampling variability (i.e., heterogeneous; *Q* (38) = 123.71, *p* < .001). The proportion of variability in effects that cannot be attributed to sampling variance (I^2 = 69.67%) is distributed roughly evenly within-studies (I_2^2 = 34.81%) and between-studies (I_3^2 = 34.86%). Although not originally reported by Lines, Pietsch, et al. (2020), we include the prediction interval in our output as an estimate of the absolute amount of variability among effects (Borenstein et al., 2017; IntHout et al., 2016); here we can see there is a 95% likelihood that an effect in future similar studies will fall between -0.36 and 1.46, which suggests that some interventions inefficacious or detrimental for performance behaviours. Finally, the log likelihood ratio tests revealed significant variance in effects within studies (level 2; LRT = 10.02, *p* = .002) and between studies (level 3; LRT = 4.28, *p* = .039), which suggests the partitioning of variance across levels is meaningful. The forest plot of effect sizes in the meta-analysis can be viewed online (https://osf.io/ambvt/).

Outliers and Influential Cases

The second section of this output file provides the results of the sensitivity analyses focused on outlier and influential cases (see Figure 2). In terms of outliers, none of the individual effects had residuals that exceeded three standard deviations. However, four individual effects – and as a result one study because it contributed only one individual effect – were flagged with a Cook's distance of three times the mean. The exclusion of these four influential cases increased the overall pooled effect by .04 (*g* =.59, 95% CI = .35, .83, *p* < .001). Collectively, these findings suggest the overall pooled effect is largely robust to outliers and influential cases. Meta-analysts might also consider a similar approach in which they compare moderator analyses with and without outlier and influential cases.

Meta-Bias

The third section of this output file provides the results of the meta-bias analyses; we considered three different elements for the purposes of this tutorial. First, we visualised potential publication bias in a traditional funnel plot as well as a sunset-enhanced funnel plot via the R package *metaviz* (Kossmeier et al., 2019) to integrate knowledge of statistical power of individual effects (Kossmeier et al., 2020). The absence of publication bias is evident when the effects sizes are scattered much like a funnel, where studies with larger samples (smaller standard error) cluster tightly around the mean effect and smaller studies (larger standard error) are highly dispersed around the mean (Lau et al., 2006). Second, we complemented a visual inspection of the funnel plot with the multilevel extension of Egger's test of symmetry in which we regressed the overall pooled effect on the standard error of individual effect sizes (Fernández-Castilla et al., 2021). Third, we examined three methodological features as moderators of the overall pooled effect, namely publication status (peer-reviewed versus dissertation) and the total number of participants (sample size) and teams (team size) in the study.

The traditional funnel plot and power-enhanced funnel plot are depicted in Figure 3. The multilevel extension of Egger's test, *F* (1,32) = 2.72, *p* = .11, suggested symmetry in the funnel plot. Visual inspection of the traditional funnel plot indicates that individual effects are roughly symmetrical and evenly distributed, with few effects falling outside the 99% confidence interval. Publication status, *F* (1,37) = 1.40, *p* = .24, sample size, *F* (1,32) = 2.08, *p* = .16, and team size, *F* (1,37) = 0.47, *p* = .50, were inconsequential predictors of the overall pooled effect. Collectively, the traditional funnel plot, multilevel extension of Egger's test, and methodological moderators suggest the meta-analytic data are influenced minimally by publication bias. However, the power-enhanced funnel plot illustrates substantial differences in the design and test combination of studies to detect a presumed true effect of *d* = .50, with the median power around 38% and no studies meeting the typically adopted 80% power. This finding is important to interpretations of the credibility of a body of evidence because statistically significant findings from low powered design and test combinations are likely to be false positives (Forstmeier et al., 2017).

Explanation of Heterogeneity Among Effects

The final section of this output file provides the results of the moderator analyses of candidate substantive factors that might alter the strength of the overall pooled effect. We implemented two different approaches to these moderator analyses in which the intercept is included (mods = ~moderator) or excluded (mods = ~moderator-1) from the statistical model. Essentially, the model with the intercept tells us if the average effect sizes for different levels of the moderator are equal to the intercept or reference value of the moderator, whereas the model that excludes the intercept tells us if the average effect sizes of the different levels of the moderator are all equal to zero. In R, the reference value (intercept) of a factor is automatically assigned to the level which occurs first numerically or alphabetically (e.g., 'active control' would be used as the reference for comparators including active control, no contact, and waitlist), unless one manually assigns the reference level prior to the model fitting element.

In the section ‘Moderator Analyses: ANOVAs’ we can see that outcome type, *F* (2,36) = 3.55, *p* = .04, and team type, *F* (1,37) = 7.81, *p* = .008, are statistically interesting moderators using an alpha of .05. For illustrative purposes, we focus on outcome type here (see Figure 4). Specifically, cognitive (*b* = .58) and behavioural outcomes (*b* = .51) are significantly higher when compared with affective outcomes (*b* = .07, *p* = .01). An examination of the output in the section ‘Moderator Analyses: Individual Moderators’ indicates that cognitive outcomes (*b* = .65, *p* <.001) but not behavioural (*b* = .58, *p* = .06) and affective outcomes (*b* = .07, *p* = .73) are statistically different from zero. Why are behavioural outcomes statistically inconsequential when the magnitude of effect is similar to cognitive outcomes? One possibility is the high standard error, which is likely due to there being three contributing individual effects to this level of the moderator, compared with 30 effects for cognitive outcomes. Finally, we can compare the degree of heterogeneity in the model when the salient moderators are included with the baseline model or overall pooled effect only (see section ‘Heterogeneity Comparisons’ in the output file). The addition of these two moderators to the baseline model, Cochran’s *Q*(38) = 123.71, *p* <.001, significantly reduced heterogeneity, yet the residual heterogeneity remained statistically meaningful, *QE*(35) = 87.97, *p* <.001. R^2 values indicate that the moderators account for approximately 26% of within-study and 29% of between-study variance.

Summary and Conclusion

Scientific knowledge is ever increasing, broadly (Bornmann & Mutz, 2015) and within the field of sport and exercise psychology (Lindahl et al., 2015). Coinciding with this growth in scientific knowledge is an increased need for and application of methods that synthesise large bodies of work in ways that summarise the current state of affairs in a specific area or on a focal topic. With significant advancements in computer and data science on the rise (Marshall & Wallace, 2019), we expect that many of the elements of the systematic review process will be optimised in years to come (e.g., machine learning in screening articles; Chai et al., 2021). With data in hand, meta-analytic techniques are ideally suited to summarise evidence across primary studies that have been quantified in some way through a summary effect. In this paper, we considered one potential source of bias that may render the findings of a meta-analysis limited or potentially erroneous, namely dependency among primary effects. We first critically evaluated common approaches to handling dependency among effect sizes in a meta-analysis, then overviewed the statistical approaches available to researchers who wish to incorporate this dependency into their statistical models. Subsequently, we presented readers with a series of reflective questions for scholars to consider when planning and executing a meta-analysis in which some degree of non-independence among effects sizes from primary studies may exist and showcased their application via a brief tutorial. As applied scientists, we adopted a non-technical approach in this paper to make it as accessible as possible, yet acknowledge that in so doing we left out some of the nuanced considerations for planning and executing a meta-analysis. Scholars interested in cultural considerations or other grouping factors (e.g., sports), for example, might need to implement a four-level meta-analysis to account for individual participants, multiple effects from within the same study, different studies, and cultural groups. With open science practices on the rise (Moreau & Gamble, 2020; Polanin et al., 2020), we envisage that statistical techniques for handling dependency among effects within meta-analyses will become more accessible over the coming years.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses. *Review of Educational Research*, *82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Andrews, D. W. K. (2001). Testing When a Parameter is on the Boundary of the Maintained Hypothesis. *Econometrica*, *69*(3), 683–734. https://doi.org/10.1111/1468-0262.00210

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Baujat, B., Mahé, C., Pignon, J.-P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine*, *21*(18), 2641–2652. https://doi.org/10.1002/sim.1221

Bax, L., Yu, L.-M., Ikeda, N., & Moons, K. G. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology*, *7*(1), 40. https://doi.org/10.1186/1471-2288-7-40

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Academic Press. https://doi.org/10.1016/B978-012691360-6/50018-5

Berkey, C. S., Anderson, J. J., & Hoaglin, D. C. (1996). Multiple-Outcome Meta-Analysis of Clinical Trials. *Statistics in Medicine*, *15*(5), 537–557. https://doi.org/10.1002/(SICI)1097-0258(19960315)15:5<537::AID-SIM176>3.0.CO;2-S

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, *7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Caldwell, A., & Vigotsky, A. D. (2020). A case against default effect sizes in sport and exercise science. *PeerJ*, *8*, e10314. https://doi.org/10.7717/peerj.10314

Carr, R. M., Prestwich, A., Kwasnicka, D., Thøgersen-Ntoumani, C., Gucciardi, D. F., Quested, E., Hall, L. H., & Ntoumanis, N. (2019). Dyadic interventions to promote physical activity and reduce sedentary behaviour: Systematic review and meta-analysis. *Health Psychology Review*, *13*(1), 91–109. https://doi.org/10.1080/17437199.2018.1532312

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chai, Kevin. E. K., Lines, R. L. J., Gucciardi, Daniel. F., & Ng, L. (2021). Research Screener: A machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews*. https://doi.org/10.1186/s13643-021-01635-3

Chan, M. E., & Arvey, R. D. (2012). Meta-Analysis and the Development of Knowledge. *Perspectives on Psychological Science*, *7*(1), 79–92. https://doi.org/10.1177/1745691611429355

Cheung, M. W.-L. (2013). Multivariate Meta-Analysis as Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 429–454. https://doi.org/10.1080/10705511.2013.797827

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. Wiley.

Cheung, M. W.-L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L. (2019). A Guide to Conducting a Meta-Analysis with Non-Independent Effect Sizes. *Neuropsychology Review*, *29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cheung, M. W.-L., & Vijayakumar, R. (2016). A Guide to Conducting a Meta-Analysis. *Neuropsychology Review*, *26*(2), 121–128. https://doi.org/10.1007/s11065-016-9319-z

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent Effect Sizes in Meta-Analysis: Incorporating the Degree of Interdependence. *Journal of Applied Psychology*, *89*(5), 780–791. https://doi.org/10.1037/0021-9010.89.5.780

Cooper, H. (2016). *Research Synthesis and Meta-Analysis: A Step-by-Step Approach* (5th ed.). Sage.

Cox, R. H., Martens, M. P., & Russell, W. D. (2003). Measuring Anxiety in Athletics: The Revised Competitive State Anxiety Inventory–2. *Journal of Sport and Exercise Psychology*, *25*(4), 519–533. https://doi.org/10.1123/jsep.25.4.519

Creswell, J. W., & Poth, C. N. (2016). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (4th ed.). Sage. https://au.sagepub.com/en-gb/oce/qualitative-inquiry-and-research-design/book246896

Dickersin, K. (1990). The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA*, *263*(10), 1385–1389. https://doi.org/10.1001/jama.1990.03440100097014

Durlak, J. A. (2009). How to Select, Calculate, and Interpret Effect Sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Noortgate, W. V. den. (2021). Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study. *The Journal of Experimental Education*, *89*(1), 125–144. https://doi.org/10.1080/00220973.2019.1582470

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665–694. https://doi.org/10.1348/000711010X502733

Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. *ArXiv:1503.02220 [Stat]*. http://arxiv.org/abs/1503.02220

Flora, D. B. (2020). Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology/Psychologie Canadienne*. https://doi.org/10.1037/cap0000218

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings - a practical guide: Avoiding false-positive findings. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gonzalez-Mulé, E., & Aguinis, H. (2018). Advancing Theory by Assessing Boundary Conditions With Metaregression: A Critical Review and Best-Practice Recommendations. *Journal of Management*, *44*(6), 2246–2273. https://doi.org/10.1177/0149206317710723

Gucciardi, D. F., Lines, R. L. J., Ducker, K. J., Peeling, P., Chapman, M. T., & Temby, P. (2020). Mental toughness as a psychological determinant of behavioral perseverance in special forces selection. *Sport, Exercise, and Performance Psychology*. https://doi.org/10.1037/spy0000208

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing Meta-Analysis in R*. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale Adaptation in Organizational Science Research: A

Review and Best-Practice Recommendations. *Journal of Management*, *45*(6), 2596–2627. https://doi.org/10.1177/0149206319850280

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150. https://doi.org/10.3389/fnins.2014.00150

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

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

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, *30*(20), 2481–2498. https://doi.org/10.1002/sim.4172

JASP Team. (2020). *JASP (Version 0.14) [Computer software]*. https://jasp-stats.org/

Johnson, B. T., & Hennessy, E. A. (2019). Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. *Social Science & Medicine*, *233*, 237–251. https://doi.org/10.1016/j.socscimed.2019.05.035

Kalaian, H. A., & Raudenbush, S. W. (1996). A Multivariate Mixed Linear Model for Meta-Analysis. *Psychological Methods*, *1*(3), 227–235. https://doi.org/10.1037/1082-989X.1.3.227

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Kossmeier, M., Tran, U. S., & Voracek, M. (2019). Visual Inference for the Funnel Plot in Meta-Analysis. *Zeitschrift Für Psychologie*, *227*(1), 83–89. https://doi.org/10.1027/2151-2604/a000358

Kossmeier, M., Tran, U. S., & Voracek, M. (2020). Power-enhanced funnel plots for meta-analysis: The sunset funnel plot. *Zeitschrift Für Psychologie*, *228*(1), 43–49. https://doi.org/10.1027/2151-2604/a000392

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ*, *333*(7568), 597–600. https://doi.org/10.1136/bmj.333.7568.597

Li, X., Dusseldorp, E., Su, X., & Meulman, J. J. (2020). Multiple moderator meta-analysis using the R-package Meta-CART. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01360-0

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ*, *339*. https://doi.org/10.1136/bmj.b2700

Lindahl, J., Stenling, A., Lindwall, M., & Colliander, C. (2015). Trends and knowledge base in sport and exercise psychology research: A bibliometric review study. *International*

Review of Sport and Exercise Psychology, *8*(1), 71–94. https://doi.org/10.1080/1750984X.2015.1019540

Lines, R. L. J., Ntoumanis, N., Thøgersen-Ntoumani, C., McVeigh, J. A., Ducker, K. J., Fletcher, D., & Gucciardi, D. F. (2020). Cross-sectional and longitudinal comparisons of self-reported and device-assessed physical activity and sedentary behaviour. *Journal of Science and Medicine in Sport*, *23*(9), 831–835. https://doi.org/10.1016/j.jsams.2020.03.004

Lines, R. L. J., Pietsch, S., Crane, M. F., Ntoumanis, N., Temby, P., Graham, S., & Gucciardi, D. F. (2020). The effectiveness of team reflexivity interventions: A systematic review and meta-analysis of randomised controlled trials. *Sport, Exercise, and Performance Psychology*. https://doi.org/10.1037/spy0000251

Lonsdale, C., Hodge, K., Hargreaves, E. A., & Ng, J. Y. Y. (2014). Comparing sport motivation scales: A response to Pelletier et al. *Psychology of Sport and Exercise*, *15*(5), 446–452. https://doi.org/10.1016/j.psychsport.2014.03.006

López-López, J. A., Noortgate, W. V. den, Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2017). Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. *Research Synthesis Methods*, *8*(4), 435–450. https://doi.org/10.1002/jrsm.1245

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods*, *9*(3), 336–351. https://doi.org/10.1002/jrsm.1310

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Marsh, H. W., Martin, A. J., & Jackson, S. (2010). Introducing a Short Version of the Physical Self Description Questionnaire: New Strategies, Short-Form Evaluative Criteria, and Applications of Factor Analyses. *Journal of Sport and Exercise Psychology*, *32*(4), 438–482. https://doi.org/10.1123/jsep.32.4.438

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*(1), 163. https://doi.org/10.1186/s13643-019-1074-9

Mathur, M. B., & VanderWeele, T. J. (2020). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, *n/a*(n/a), 1–6. https://doi.org/10.1002/jrsm.1464

McGuinness, L. A., & Higgins, J. P. T. (2020). Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods*. https://doi.org/10.1002/jrsm.1411

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-Scale Replication Projects in Contemporary Psychological Research. *The American Statistician*, *73*(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Molloy, G. J., Noone, C., Caldwell, D., Welton, N. J., & Newell, J. (2018). Network meta-analysis in health psychology and behavioural medicine: A primer. *Health Psychology Review, 12*(3), 254–270. https://doi.org/10.1080/17437199.2018.1457449

Moreau, D., & Chou, E. (2019). The Acute Effect of High-Intensity Exercise on Executive Function: A Meta-Analysis. *Perspectives on Psychological Science*, *14*(5), 734–764. https://doi.org/10.1177/1745691619850568

Moreau, D., & Gamble, B. (2020). Conducting a Meta-Analysis in the Age of Open Science: Tools, Tips, and Practical Recommendations. *Psychological Methods*. https://psyarxiv.com/t5dwg/

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/html_ug.shtml

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

NHMRC. (2009). *National Health and Medical Research Council (NHMRC) additional levels of evidence and grades for recommendations for developers of guidelines.* Commonwealth of Australia.

Ntoumanis, N., Ng, J. Y. Y., Prestwich, A., Quested, E., Hancox, J. E., Thøgersen-Ntoumani, C., Deci, E. L., Ryan, R. M., Lonsdale, C., & Williams, G. C. (2020). A meta-analysis of self-determination theory-informed intervention studies in the health domain: Effects on motivation, health behavior, physical, and psychological health. *Health Psychology Review*, 1–31. https://doi.org/10.1080/17437199.2020.1718529

Park, S., & Beretvas, S. N. (2019). Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model. *Behavior Research Methods*, *51*(1), 152–171. https://doi.org/10.3758/s13428-018-1156-y

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126

Pelletier, L. G., Rocchi, M. A., Vallerand, R. J., Deci, E. L., & Ryan, R. M. (2013). Validation of the revised sport motivation scale (SMS-II). *Psychology of Sport and Exercise*, *14*(3), 329–341. https://doi.org/10.1016/j.psychsport.2012.12.002

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Pigott, T. D., & Polanin, J. R. (2020). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Polanin, J. R., Hennessy, E. A., & Tanner-Smith, E. E. (2017). A Review of Meta-Analysis Packages in R. *Journal of Educational and Behavioral Statistics*, *42*(2), 206–242. https://doi.org/10.3102/1076998616674315

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and Reproducibility of Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

Polanin, J. R., & Terzian, M. (2019). A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. *Journal of Clinical Epidemiology*, *106*, 60–69. https://doi.org/10.1016/j.jclinepi.2018.10.006

Pustejovsky, J. (2020). *ClubSandwich: Cluster-robust (Sandwich) variance estimators with small-sample corrections. R package version 0.5.0.* https://cran.r-project.org/web/packages/clubSandwich/clubSandwich.pdf

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01549

R Development Core Team. (2018). *R: A language and environment for statistical computing*. R foundation for statistical computing Vienna, Austria.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*(1), 111–120. https://doi.org/10.1037/0033-2909.103.1.111

Riley, R. D., Thompson, J. R., & Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, *9*(1), 172–186. https://doi.org/10.1093/biostatistics/kxm023

Rodgers, M. A., & Pustejovsky, J. E. (2020). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*. https://doi.org/10.1037/met0000300

Rosenthal, R., & Rubin, D. B. (1986). Meta-Analytic Procedures for Combining Studies With Multiple Effect Sizes. *Psychological Bulletin*, *99*(3), 400–406. https://doi.org/10.1037/0033-2909.99.3.400

Rudolph, C. W., Chang, C. K., Rauvola, R. S., & Zacher, H. (2020). Meta-analysis in vocational behavior: A systematic review and recommendations for best practices. *Journal of Vocational Behavior*, *118*, 103397. https://doi.org/10.1016/j.jvb.2020.103397

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: Many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, *3*(2), 80–97. https://doi.org/10.1002/jrsm.1037

SAS Institute Inc. (2016). *SAS® 9.4 Language Reference: Concepts* (6th ed.). SAS Institute Inc.

Sharpe, D., & Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology/Psychologie Canadienne*. https://doi.org/10.1037/cap0000215

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ, 358*. https://doi.org/10.1136/bmj.j4008

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

StataCorp. (2019). *Stata Statistical Software: Release 16.* StataCorp.

Steel, P., Beugelsdijk, S., & Aguinis, H. (2020). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews. *Journal of International Business Studies*. https://doi.org/10.1057/s41267-020-00385-z

Steffens, N. K., LaRue, C. J., Haslam, C., Walter, Z. C., Cruwys, T., Munt, K. A., Haslam, S. A., Jetten, J., & Tarrant, M. (2019). Social identification-building interventions to improve health: A systematic review and meta-analysis. *Health Psychology Review*, 1–28. https://doi.org/10.1080/17437199.2019.1669481

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science*, *76*(5), 650–661. https://doi.org/10.1086/605819

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.2307/2282137

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 10.

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*. https://doi.org/10.1136/bmj.l4898

Stroup, W. W., Milliken, G. A., Claassen, E. A., & Wolfinger, R. D. (2018). *SAS® for Mixed Models: Introduction and Basic Applications.* SAS Institute Inc.

Suurmond, R., Rhee, H. van, & Hak, T. (2017). Introduction, comparison, and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods*, *8*(4), 537–553. https://doi.org/10.1002/jrsm.1260

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and spss. *Research Synthesis Methods*, *5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: A Tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5

The jamovi project. (2020). *Jamovi*. https://www.jamovi.org

Tipton, E., Bryan, C. J., & Yeager, D. S. (2020). *To Change the World, Behavioral Intervention Research Will Need to Get Serious About Heterogeneity*. https://statmodeling.stat.columbia.edu/wp-content/uploads/2020/07/ Heterogeneity-1-23-20-NHB.pdf

Tipton, E., & Pustejovsky, J. E. (2015). Small-Sample Adjustments for Tests of Moderators and Model Fit Using Robust Variance Estimation in Meta-Regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

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

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

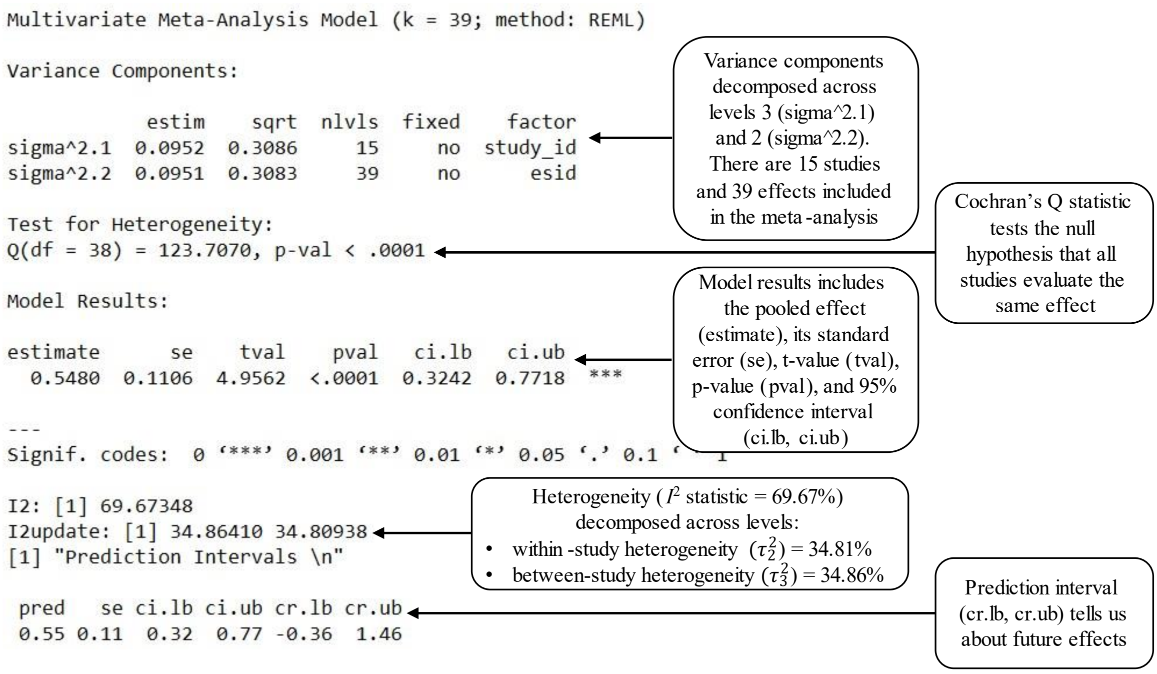
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Wallace, B. C., Schmid, C. H., Lau, J., & Trikalinos, T. A. (2009). Meta-Analyst: Software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology*, *9*(1), 80. https://doi.org/10.1186/1471-2288-9-80

White, I. R. (2011). Multivariate Random-effects Meta-regression: Updates to Mvmeta. *The Stata Journal*, *11*(2), 255–270. https://doi.org/10.1177/1536867X1101100206

Wilson, D. B. (n.d.). *Effect Size Calculator*. Retrieved 31 August 2020, from https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD-main.php

Wolanin, A., Hong, E., Marks, D., Panchoo, K., & Gross, M. (2016). Prevalence of clinically elevated depressive symptoms in college athletes and differences by gender and sport. *British Journal of Sports Medicine*, *50*(3), 167–171. https://doi.org/10.1136/bjsports-2015-095756



```
Multivariate Meta-Analysis Model (k = 39; method: REML)

Variance Components:

            estim    sqrt  nlvls  fixed    factor
sigma^2.1  0.0952  0.3086     15     no  study_id
sigma^2.2  0.0951  0.3083     39     no      esid

Test for Heterogeneity:
Q(df = 38) = 123.7070, p-val < .0001

Model Results:

estimate      se    tval    pval   ci.lb   ci.ub
  0.5480  0.1106  4.9562  <.0001  0.3242  0.7718  ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

I2: [1] 69.67348
I2update: [1] 34.86410 34.80938
[1] "Prediction Intervals \n"

 pred   se ci.lb ci.ub cr.lb cr.ub
 0.55 0.11  0.32  0.77 -0.36  1.46
```

Variance components decomposed across levels 3 (sigma^2.1) and 2 (sigma^2.2). There are 15 studies and 39 effects included in the meta -analysis

Cochran’s Q statistic tests the null hypothesis that all studies evaluate the same effect

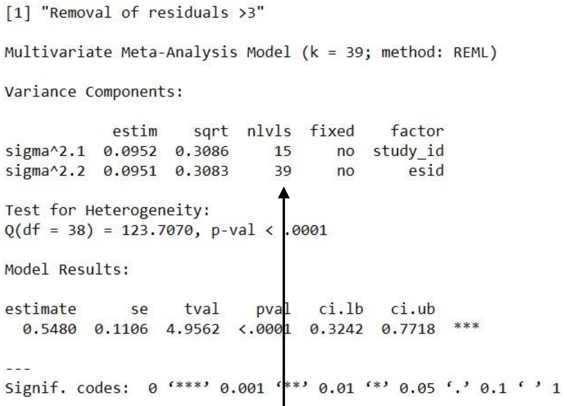
Model results includes the pooled effect (estimate), its standard error (se), t-value (tval), p-value (pval), and 95% confidence interval (ci.lb, ci.ub)

Heterogeneity (I^2 statistic = 69.67%) decomposed across levels:

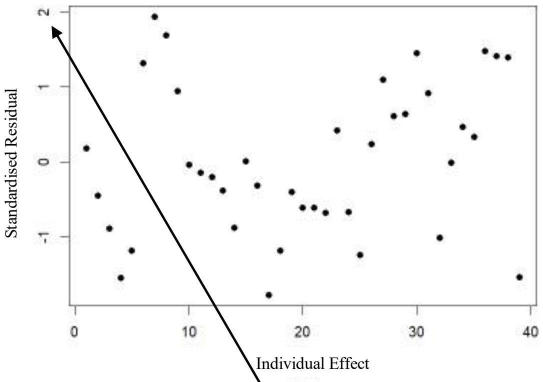
- within -study heterogeneity (τ_2^2) = 34.81%
- between-study heterogeneity (τ_3^2) = 34.86%

Prediction interval (cr.lb, cr.ub) tells us about future effects

Figure 1. Overall model R output of meta-analysis using the *metafor* package (Viechtbauer, 2010).

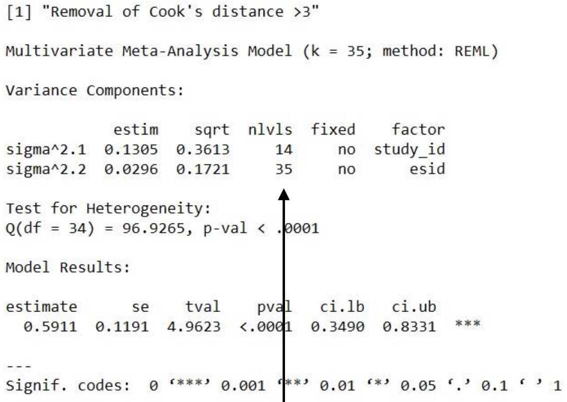


```
[1] "Removal of residuals >3"

Multivariate Meta-Analysis Model (k = 39; method: REML)

Variance Components:

            estim    sqrt  nlvls  fixed    factor
sigma^2.1  0.0952  0.3086     15     no  study_id
sigma^2.2  0.0951  0.3083     39     no      esid

Test for Heterogeneity:
Q(df = 38) = 123.7070, p-val < .0001

Model Results:

estimate      se    tval    pval   ci.lb   ci.ub
  0.5480  0.1106  4.9562  <.0001  0.3242  0.7718  ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

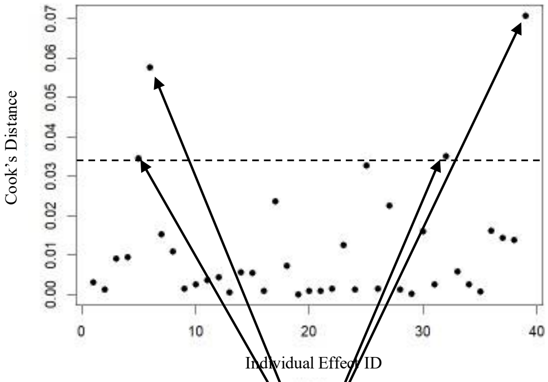


We have the same number of individual effects (level 2) and studies (level 3) as the overall model.

Effects whose standardised residuals exceed ± 3 (or the ± 1.96 bounds) are considered worthy of investigation as an outlier; none were identified here.



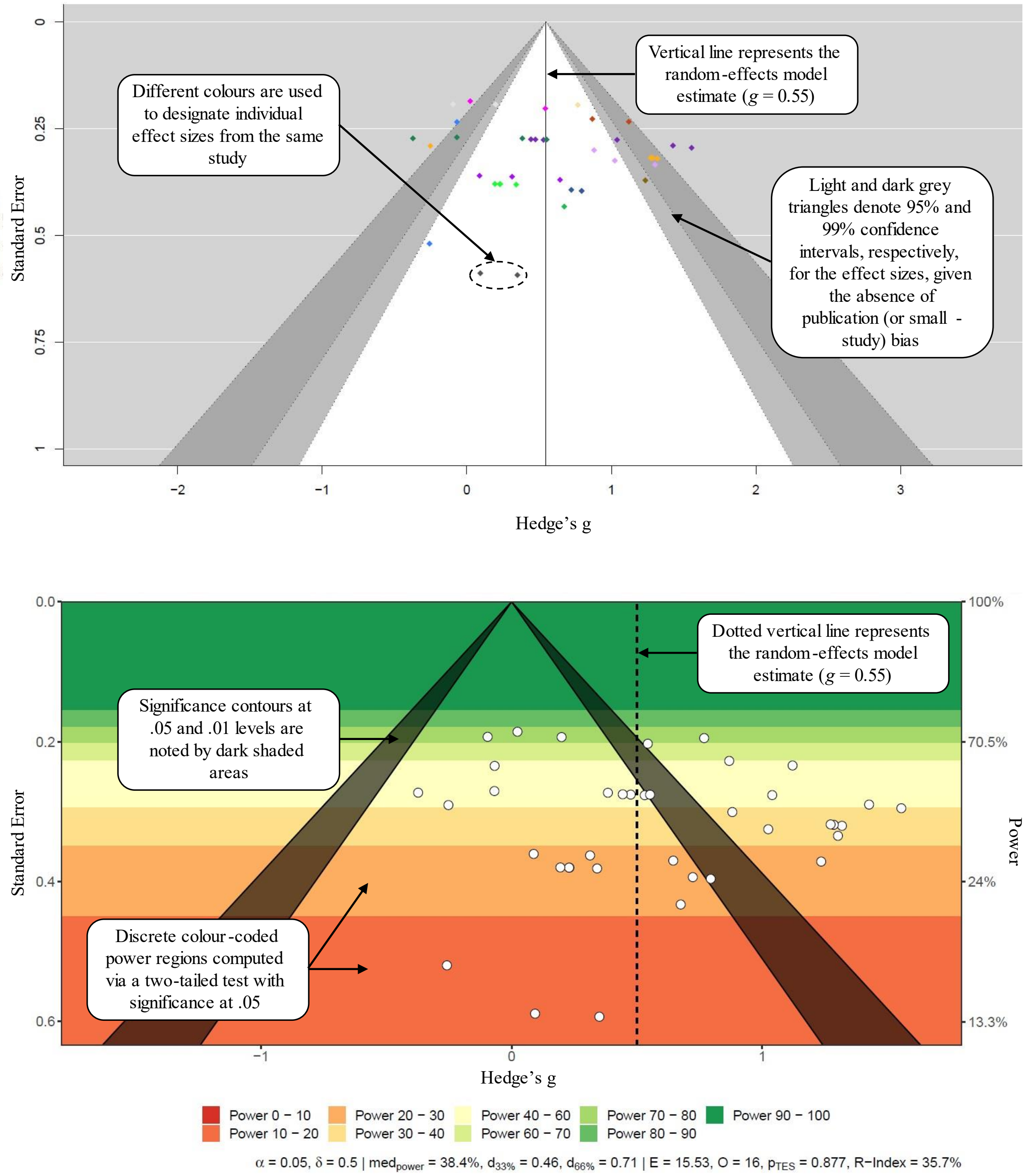
```
[1] "Removal of Cook's distance >3"

Multivariate Meta-Analysis Model (k = 35; method: REML)

Variance Components:

            estim    sqrt  nlvls  fixed    factor
sigma^2.1  0.1305  0.3613     14     no  study_id
sigma^2.2  0.0296  0.1721     35     no      esid

Test for Heterogeneity:
Q(df = 34) = 96.9265, p-val < .0001

Model Results:

estimate      se    tval    pval   ci.lb   ci.ub
  0.5911  0.1191  4.9623  <.0001  0.3490  0.8331  ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Four individual effects (level 2) and, as a result, 1 study (level 3) were removed as influential cases for the sensitivity test.

Effects with a Cook's distance more than 3 times the mean are considered worthy of investigation as an influential case; 4 individuals effects are flagged here.

Figure 2. Outlier (top) and Cook's distance (bottom) analysis output (Note: dotted line represents the threshold for Cook's distance of 3 times the mean).



Different colours are used to designate individual effect sizes from the same study

Vertical line represents the random-effects model estimate (*g* = 0.55)

Light and dark grey triangles denote 95% and 99% confidence intervals, respectively, for the effect sizes, given the absence of publication (or small - study) bias

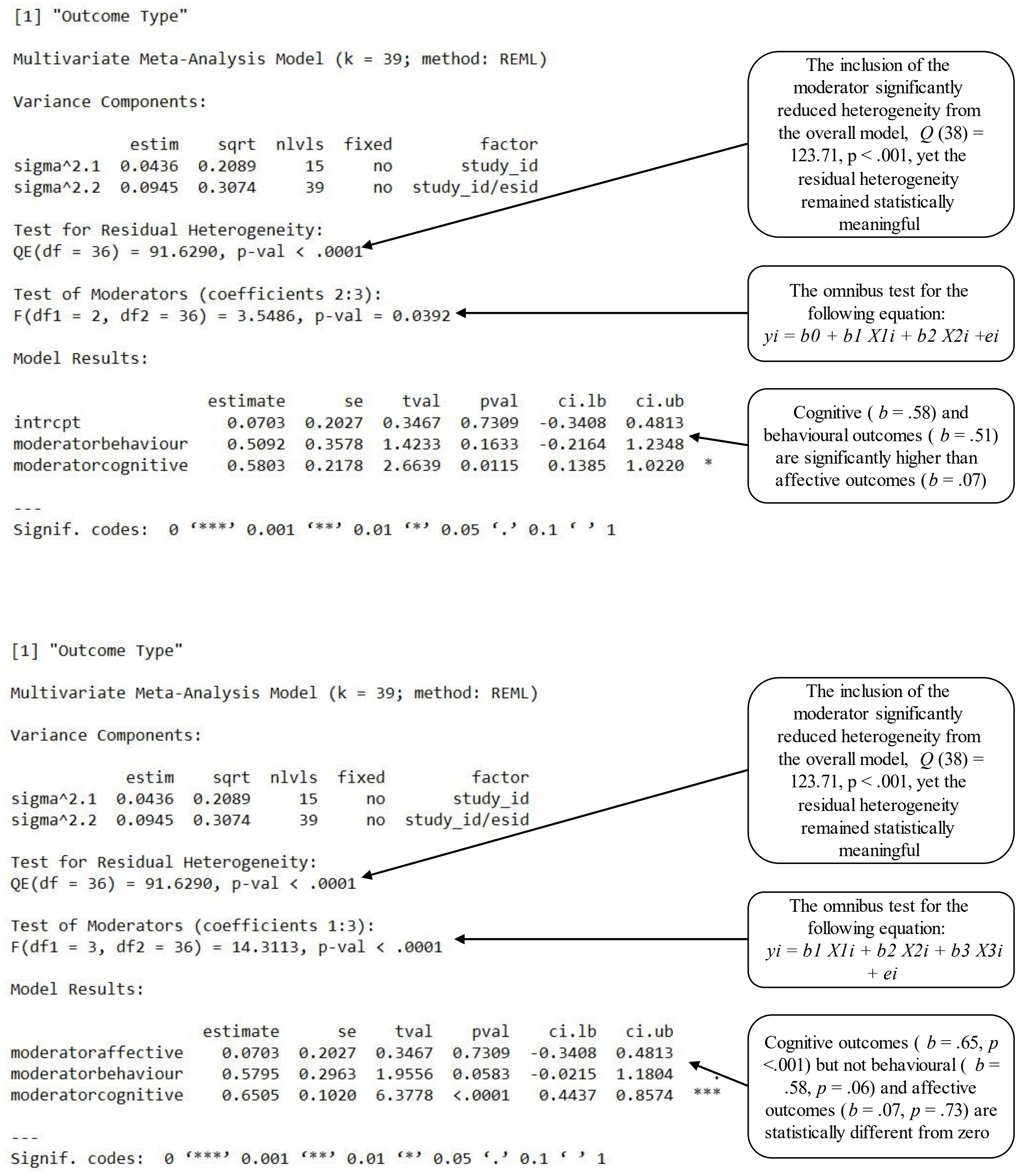
Dotted vertical line represents the random-effects model estimate (*g* = 0.55)

Significance contours at .05 and .01 levels are noted by dark shaded areas

Discrete colour-coded power regions computed via a two-tailed test with significance at .05

α = 0.05, δ = 0.5 | med_{power} = 38.4%, $d_{33\%}$ = 0.46, $d_{66\%}$ = 0.71 | E = 15.53, O = 16, p_{TES} = 0.877, R−Index = 35.7%

Figure 3. Funnel (top) and power-enhanced funnel plots (bottom) for individual effect sizes in the meta-analysis.



```
[1] "Outcome Type"

Multivariate Meta-Analysis Model (k = 39; method: REML)

Variance Components:

            estim    sqrt  nlvls  fixed         factor
sigma^2.1  0.0436  0.2089     15     no       study_id
sigma^2.2  0.0945  0.3074     39     no  study_id/esid

Test for Residual Heterogeneity:
QE(df = 36) = 91.6290, p-val < .0001

Test of Moderators (coefficients 2:3):
F(df1 = 2, df2 = 36) = 3.5486, p-val = 0.0392

Model Results:

                    estimate      se    tval    pval    ci.lb   ci.ub
intrcpt               0.0703  0.2027  0.3467  0.7309  -0.3408  0.4813
moderatorbehaviour    0.5092  0.3578  1.4233  0.1633  -0.2164  1.2348
moderatorcognitive    0.5803  0.2178  2.6639  0.0115   0.1385  1.0220  *

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The inclusion of the moderator significantly reduced heterogeneity from the overall model, *Q* (38) = 123.71, p < .001, yet the residual heterogeneity remained statistically meaningful

The omnibus test for the following equation: *yi = b0 + b1 X1i + b2 X2i +ei*

Cognitive (*b* = .58) and behavioural outcomes (*b* = .51) are significantly higher than affective outcomes (*b* = .07)

```
[1] "Outcome Type"

Multivariate Meta-Analysis Model (k = 39; method: REML)

Variance Components:

            estim    sqrt  nlvls  fixed         factor
sigma^2.1  0.0436  0.2089     15     no       study_id
sigma^2.2  0.0945  0.3074     39     no  study_id/esid

Test for Residual Heterogeneity:
QE(df = 36) = 91.6290, p-val < .0001

Test of Moderators (coefficients 1:3):
F(df1 = 3, df2 = 36) = 14.3113, p-val < .0001

Model Results:

                    estimate      se    tval    pval    ci.lb   ci.ub
moderatoraffective    0.0703  0.2027  0.3467  0.7309  -0.3408  0.4813
moderatorbehaviour    0.5795  0.2963  1.9556  0.0583  -0.0215  1.1804    .
moderatorcognitive    0.6505  0.1020  6.3778  <.0001   0.4437  0.8574  ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The inclusion of the moderator significantly reduced heterogeneity from the overall model, *Q* (38) = 123.71, p < .001, yet the residual heterogeneity remained statistically meaningful

The omnibus test for the following equation: *yi = b1 X1i + b2 X2i + b3 X3i + ei*

Cognitive outcomes (*b* = .65, *p* <.001) but not behavioural (*b* = .58, *p* = .06) and affective outcomes (*b* = .07, *p* = .73) are statistically different from zero

Figure 4. Excerpt of the output of the moderator analyses of the overall pooled effect (Note: top panel includes the intercept in the statistical model, whereas the bottom panel excludes the intercept from the statistical model).