27 Meta-analysis

Judith A. Hall and David Miller

Meta-analysis has come a long way since the word was introduced to psychology and other fields in the 1970s. While some analytic procedures were developed and used in psychology earlier (e.g., methods for combining probabilities across studies: Mosteller & Bush, 1954; Rosenthal, 1966), meta-analysis became widely known in the field when Smith & Glass (1977) reviewed a large number of psychotherapy outcome studies using meta-analytic methods. The controversy that ensued was fierce, but the validity of meta-analysis was soon recognized. Fortunately, meta-analysis is no longer called "mega-silliness" (Eysenck, 1978). Many thousands of meta-analyses have been done in psychology, public health, medicine, education, and many other scientific fields.

As the field of meta-analysis has grown, it has, however, also undergone internal methodological revolutions and debates. By and large, these debates have produced useful discourse and progress. On the other hand, some debates have not reached consensus, nor should they necessarily because choices must depend on the literature under review, the questions the researcher wishes to answer, and the inferences one wishes to draw. Meta-analysis is not a monolithic or orthodox set of procedures. It is an *approach*, a philosophical commitment to the idea that often we can improve understanding of a research topic by doing a thorough, quantitative, integrative analysis of the relevant empirical studies. It is also evolving, as new norms and statistical procedures emerge.

Resources on understanding and conducting meta-analysis are abundant. These include textbooks and handbooks (Borenstein et al., 2009; Cleophas & Zwinderman, 2017; Cooper, 2016; Cooper et al., 2019; Lipsey & Wilson, 2001; Schmid et al., 2021), chapters giving overviews or focusing on specific themes (Del Re & Flückiger, 2016; Johnson & Eagly, 2014; Rosenthal, 1994; Valentine, 2009, 2012; White, 2009), guidelines for recommended best practices (Johnson, 2021), and countless articles on methodology. New methods are proposed, and old ones debated, in journals such as *Research Synthesis Methods*, *Psychological Methods*, and the *Psychological Bulletin*.

The abundance of resources and the sophistication of available methods can, however, be overwhelming. A glance at meta-analyses published across the decades shows a sharp increase in complexity, which can be daunting even to seasoned meta-analytic researchers. This chapter therefore aims to demystify some of that complexity, offering conceptual explanations instead of mathematical formulas, along with introducing some points of disagreement within the meta-analytic community. We aim to help readers who have not conducted a meta-analysis to get started, as well as to help those who simply want to be intelligent consumers of published meta-analyses.

27.1 Definitional Issues and Aims of Meta-analysis

The term *meta-analysis* refers to the quantitative synthesis of results across multiple studies. Early efforts at quantitative reviews used "vote counting," which usually took the form of tabulating the number of statistically significant and non-significant studies in one's literature (Bushman & Wang, 2009). Because a thumbs-up–thumbs-down

approach is insensitive and because the conclusions can be seriously biased by variations in the sample sizes in the studies being tabulated, this approach is relatively uncommon. In contrast to narrative and vote-counting reviews, meta-analyses mainly extract and analyze *effect sizes* that capture the direction and magnitude of findings. Many indices of effect size are available (Cohen, 1989; Rosenthal, 1994), and their application depends on the nature of the data to be summarized and the reviewer's own preferences. The value of effect sizes is that they are a standardized metric (common ones being standardized mean differences such as Hedges' *g* and Cohen's *d*, and the Pearson correlation) that permit comparison of outcome magnitudes across studies whose original measurement metrics and instruments were not necessarily identical. Based on effect sizes, meta-analysts can use statistical methods to summarize overall effects, compare differences across studies, and diagnose biases in the literature (Borenstein et al., 2009).

Typically, once the main research questions are decided on, a meta-analyst proceeds to do a *systematic review* following a set of norms for searching the target literature and screening studies for their relevance. This process aims to be "methodical, comprehensive, transparent, and replicable" (Siddaway et al., 2019, 751), with features such as careful explanation of the inclusion criteria, databases searched, search terms, and numbers of studies that were included and excluded (Moher et al., 2009). This chapter later describes how to conduct these steps with high quality and to document decisions using reporting guidelines such as the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) checklists and flow diagrams (Page et al., 2022).

Meta-analysis and systematic review are distinct concepts, as one can happen without the other. One can do a systematic review without a meta-analysis of results, such as when describing methods or conceptual usage in the evidence base. Some review journals such as the *Psychological Bulletin* expect review authors to follow systematic search and documentation guidelines, even if the review is qualitative in nature and does not include a quantitative meta-analysis (Johnson, 2021). One can also do a meta-analysis without a systematic review, such as when a researcher conducts several studies and performs a "mini" meta-analysis to summarize them (Goh et al., 2016). Systematic searching is not necessary in that case because the researcher already has all the relevant studies in hand (e.g., Freudenberg et al., 2020; Razpurker-Apfeld & Shamoa-Nir, 2021). However, when aiming to quantitatively synthesize results beyond the authors' own work, a high-quality systematic search and its documentation should usually precede a meta-analysis.

27.2 When Is Meta-analysis Appropriate?

Not all research literatures and not all research questions are suitable for meta-analysis. The studies of interest might use qualitative methods, which offer opportunities for using synthesis methods other than meta-analysis (Thomas & Harden, 2008). Even when the studies are quantitative, narrative review is sometimes the best choice because the studies are judged to be truly incommensurable in method or because the researcher's focus is too broad for a quantitative treatment. Narrative reviews can have high theoretical and evidentiary value, meaning there is no intrinsic competition between quantitative and nonquantitative approaches.

A question often asked is, "How many studies do you need to do a meta-analysis?" Though journals such as the *Psychological Bulletin* tend to publish meta-analyses with dozens of studies, meta-analytic techniques can be usefully applied to even a few studies. Pooling evidence across two or more studies can offer greater precision in estimating the effect of interest than any

individual study considered on its own (Valentine et al., 2010). However, tests of differences in effects across studies should be interpreted cautiously if based on a small number of studies.

To conduct a meta-analysis, the researcher must have a *focused* research question. This requirement generally means that the effect of interest in a study can be estimated with a single degree of freedom, such as comparisons of two means, planned contrasts, or trends tested within a set of means, interaction effects with a single degree of freedom (such as in a two-by-two analysis of variance), or linear correlations (Rosenthal, 1994). A defining feature of these comparisons is that one can put a sign (+, –, or, if justified, 0) on the derived effects. Is one mean bigger than the other? Is the correlation positive or negative? In some cases, however, one can also meta-analyze simple means (point estimates) if the measurement scale is consistent across studies (e.g., Donnelly & Twenge, 2017; Konrath et al., 2011; Miller et al., 2018).

Studies that produce only unfocused results such as multi-*df* (omnibus) *F*-ratios are generally not amenable to standard meta-analytic practices. For that reason, squared indices should be used with caution as they may obscure whether a result is focused or omnibus (Rosenthal, 1994). Squared indices also lack a sign, which is another serious problem for meta-analysis. However, the meta-analyst can sometimes use omnibus results to produce the focused comparisons of interest (Rosenthal et al., 2000). As an illustration, suppose the original author reports the omnibus *F* for a one-way analysis of variance (ANOVA) of five experimental conditions. If appropriate information is provided or can be obtained, the meta-analyst could calculate their own contrast, such as a linear or quadratic trend across conditions or a contrast between the control condition and one of the other conditions. Calculating new effects not originally reported is one of the meta-analyst's tools; meta-analysts become versatile at converting the data as they are reported into the quantities necessary for their analysis.

27.3 Affordances of Meta-analysis

Meta-analysis is usually a significant advance over a corresponding narrative review. These advantages include being able to analyze many studies, draw precise conclusions, and detect subtle differences and patterns that a narrative reviewer could never hope to detect or confirm. Four core functions of systematic reviews and meta-analyses include (a) mapping, (b) summarizing, (c) comparing, and (d) assessing bias.

27.3.1 Mapping

Systematic reviews allow one to carefully describe what has (and has not) been done in a field, including the questions that have been asked and the methods researchers have used. As noted previously, meta-analytic projects typically begin with a systematic review, allowing for this opportunity. Even without extracting effect sizes, these mapping efforts can inform researchers where they should direct future research efforts or offer critiques of methodology or conceptual usage.

27.3.2 Summarizing

The summarizing feature of meta-analysis gives the answer to your overall research question, which is typically (though not always) a main-effect type of question such as, "Does this kind of intervention work? or "Are narcissism and aggression correlated?" Calculating the central tendency or overall outcome of one's focused question is usually the first substantive result one seeks. For example, if the meta-analysis asks how variables *X* and *Y* are correlated, the central tendency could be the average of the *X*–*Y* correlations across all of the studies in the database. Summarizing generally includes a significance test to see whether the summarized effect deviates from zero or "no effect" and also includes analyses of variation among effects (i.e., heterogeneity).

27.3.3 Comparing

The "comparing" function is about detecting *moderators* – ways in which the basic effect varies with features of study methodology or population characteristics. In the example of narcissism and aggression, the moderator could be gender or age, or different ways in which the key variables are operationalized. Moderator analyses generally cannot reveal a causal connection between the moderator and the outcomes, because studies are not randomly assigned to their respective moderator categories; a moderator analysis can only show that there is a correlation between the moderator and the outcome, as in finding that narcissism and aggression are more strongly related among younger than older samples.

27.3.4 Assessing Bias

Systematic review and meta-analysis tools allow for assessing several types of bias, including with regard to the search and selection of studies for the review, the methodological quality of studies, and the representativeness of the studies available to be retrieved (usually called publication bias or selective-reporting bias). Meta-analysts should carefully consider these biases throughout their project design and interpretation of results.

27.4 A Brief History

Errors of judgment, understanding, or calculation can occur with any method but these can multiply as methods become more sophisticated and opaque. As Rosenthal (1995, 183) said, "In 20 years of reviewing meta-analytic literature syntheses, I have never seen a meta-analysis that was 'too simple,' but I have often seen meta-analyses that were very fancy and very much in error." Things have become much fancier since then, prompting Borenstein (2019) to write a book on how to avoid mistakes in meta-analysis. As a general point, simplicity is not necessarily a fault and, along the same lines, meta-analyses performed in an earlier decade do not necessarily deserve to be viewed with suspicion based on methodology alone.

As we said, early critics questioned the legitimacy of meta-analysis as a method. One early complaint was that it is overreaching to use meta-analytic significance testing to generalize across studies. People wondered, for example, whether it was justifiable to conclude that an overall effect is significantly different from zero even though some, or maybe even all, of the individual studies failed to reach significance. This question is no longer debated; researchers now agree that a key advantage of meta-analysis is being able to combine evidence across multiple studies.

Another complaint was that meta-analysis "combines apples and oranges." This complaint captures the fear that disparate methodologies are lumped together in a meaningless way when calculating the overall effect. Typically, the meta-analyst uses the same conceptual heading ("fruit") to group studies that differ to some extent conceptually or methodologically ("apples" and "oranges"). Perhaps "aggression" was sometimes physical and sometimes verbal, or sometimes measured by self-report and sometimes by teacher ratings. Perhaps "narcissism" was measured with several different self-report scales in the literature. The meta-analyst must define and justify the criteria they use for applying a given conceptual label to diverse-seeming variables; sometimes there could be legitimate debate because different meta-analysts might make different, but equally defensible, decisions. In any case, if there are sufficient studies, the meta-analyst should perform moderator analyses to see whether different operational definitions of key variables matter.

Another complaint is called "garbage in, garbage out" to describe the perils of trying to draw a valid conclusion based on a collection of flawed studies. This concern can be valid if the meta-analyst does not address the possible biasing impact of weakly designed studies. Some meta-analysts therefore use

coding schemes or rubrics to measure the methodological quality of studies, which can then be used as a moderator variable to investigate a possible "garbage" factor. (But even studies potentially rejected as "garbage" should be inspected closely in case they yield previously unconsidered, substantive insights.) Other meta-analysts craft their inclusion criteria so that only high-quality studies are included. Another variant of the "garbage-in–garbage-out" complaint is concern with selective-reporting bias – the possibility that studies with small or nonsignificant or counterintuitive effects have not been published. Relatedly, the quality of one's database is jeopardized if a fair and exhaustive search has not been made. Selective reporting by the original authors, selective publication, and incomplete searching could all bias the meta-analyst's conclusions.

Along with a steady increase in statistical analysis options, meta-analysis has also experienced notable advances over the years in understanding (and dealing with) the implications of studies' different sample sizes. Although meta-analysis generally emphasizes effect sizes rather than the statistical significance of individual studies, the field is not entirely free of concerns with studies' sample sizes. The role of sample size in significance testing is, of course, obvious and is one reason why meta-analysts grew impatient early on with simple tallies of whether individual studies were significant or not. But sample size has more far-reaching effects than that in the context of meta-analysis. The result of a study – let us say, the correlation between narcissism and aggression – has its own variance, which is to say its own degree of uncertainty. Other things being equal, one would have more confidence in the result of the larger study because, being based on many research participants, it is a better estimate of the population's "true" value. Most meta-analysts nowadays weight the effects of bigger studies (using various functions of sample size) more than those of smaller studies.

Starting in the late 1970s, researchers sometimes used a few weighted analyses, mainly the combined-*p* statistic called the Stouffer test (Mosteller & Bush, 1954; Rosenthal & Rubin, 1979), but they also looked at their results in an unweighted (random-effects) way, counting each study's effect size equally and using ordinary statistical tools for analysis. An early example is Hall's (1978) meta-analysis of gender differences in decoding nonverbal cues, which analyzed unweighted effect sizes using ordinary statistics (a one-sample *t*-test for testing the mean effect against zero and Pearson correlations for looking at moderators). In the early years, researchers often mingled the unweighted and weighted approaches without comment.

Then came the introduction of a fuller collection of weighted models (Hedges & Vevea, 1998), within which a central distinction is between a *fixed*- and a *random*-effects model. Fixed-effects models assume that there is one true effect underlying all studies in the database, which means that all variation between studies is due solely to sampling error. A random-effects model assumes that there is variance due to factors other than sampling error, with attendant greater generalization to new studies. Fixed-effects analyses harnessed study size in calculating confidence intervals and *p*-values, yielding much higher statistical power because "*N*" was effectively participants within studies, not studies. There emerged a quickly growing dogmatism about using the fixed-effects model and most meta-analyses were done that way for over twenty years (as documented by Schmidt, Oh, and Hayes, 2009). Some mixing, and sometimes comparing, of weighted and unweighted approaches still occurred (e.g., DiMatteo, 2004; Hall et al., 2005). Most authors did not discuss the inferential implications of using the different types of model, but surely enjoyed the much greater statistical power afforded by the fixed-effects approach compared to the various random-effects models.

Of course, studies differ in many ways besides sample size – indeed one of the key advantages of meta-analysis is its capacity to compare studies that are not identical in design. Heterogeneity among effect sizes beyond what would be expected based on sampling error is the rule, not the exception, in published meta-analyses (Kenny & Judd, 2019), raising serious questions about the appropriateness of a fixed-effects analysis in many instances. Furthermore, a fixed approach in meta-analysis (just as when applying a fixed-effects model in one's ANOVA of a single study) means that generalization must be to the very same studies but with new participants, not to new studies that might vary in other respects. And so the pendulum swung dramatically away from the fixed approach, as statistical models were developed to incorporate variance due to studies (meaning it is not assumed that there is "one true" effect) as well as to sample size. The prevailing habit as of this writing is a random-effects model along those lines, which we call the weighted random-effects model. This model often has less statistical power than the fixed approach, reflecting its more ambitious inferential goals (i.e., generalization to new and not identical studies).

What happened to the simple, unweighted (random-effects) approach? There are features of this approach that merit a new look and comparison against other models (Hall & Rosenthal, 2018; Shuster et al., 2012). Not weighting by sample size avoids the risk of accidentally confounding sample size with other methodological features (such as study design or population characteristics), a problem of biased interpretation flagged many times by commentators (e.g., Borenstein, 2019). Such confounding can make weighting a threat to validity because one is actually giving more weight to certain kinds of study, not simply those with larger sample size. An unweighted approach also allows for simpler analytic methods, such as the regular descriptive and inferential statistics that most people with graduate psychology education know how to do already. The unweighted approach remains in use (Schlegel et al., 2017; Schlegel et al., 2020; Zuckerman et al., 2013; Zuckerman et al., 2016), with some meta-analysts wisely presenting both weighted and unweighted approaches (Dickens & Robins, 2022; Tucker-Drob et al., 2019).

The shift to random-effects models has also prompted broader methodological attention to the question, How can we explore and understand why effects vary across studies? This goal contrasts with the question often at the center of methodological debates about weighting, which is, How should we estimate the mean effect and its significance? This broadening of analytic goals is important because, just as in primary research, the mean is only one parameter for describing an outcome's distribution and relationships with other theoretically meaningful constructs (for a broader review of this methodological shift, see Tipton et al., 2019). For instance, in one meta-analysis on children's gender science stereotypes, the authors focused on understanding how children's stereotypes varied across age and historical time (Miller et al., 2018). The mean effect size was only of secondary interest, though it was computed and interpreted. Similarly, a researcher on narcissism and aggression might care little about the overall relation between the two variables across studies of all types, but instead focus attention on specific patterns of relations that have practical or theoretical importance.

27.5 Choices and Challenges

Meta-analysis is not a cookbook enterprise. One of the first surprises facing a new meta-analyst is the plethora of choices that need to be made. The meta-analyst is allowed to make executive decisions (and must make many), and must painstakingly justify and document them. The meta-analyst should carefully think about the guiding questions up front and, where possible, make theory-driven predictions and

preregister them before conducting the review (Lakens et al., 2016). However, other meta-analyses may have more exploratory, descriptive goals – to simply "tell it how it is" – in which case the questions investigated may evolve as the work unfolds. Such exploratory approaches are valuable, but the meta-analyst should be explicit about the exploratory nature of the project in writing up the results.

When making these many decisions, the "correct" choice is often not clear because there is no objectively "correct" choice. This point applies to one's scheme for coding studies and to the analyses one does. All meta-analytic results are predicated on the choices the meta-analyst has made. Sometimes one can do things several different ways and compare them. One must, of course, understand the implications of whatever one does.

Here are some of the common areas where choices have to be made:

- What is the guiding research question?
- What are the limits of your inquiry – for example, do you want only adults, or only typically functioning individuals, or only studies using behavioral observation?
- What methods and databases will you use for the literature search?
- What variables do you want to code to describe the studies' samples and methods?
- Although all coded study characteristics can be analyzed as moderators, which ones have special importance given your theoretical interests?
- How should the different categories of a coded study characteristic be defined operationally?
- What index of effect size will you use?
- How do you handle missing effect sizes?
- How do you handle multiple effect sizes within one study?
- How do you find or calculate the effect size when a study offers you more than one way, or does not give you the effect you need directly?
- How should you check on coder reliability?
- What statistical software will you use?
- What kinds of model will you run?
- How will you decide if there has been selective-reporting bias?

Each such choice requires careful thought about your goals and about criticisms that you may need to fend off. Furthermore, many early choices will inevitably be revisited as you become fully familiar with the literature. You might change the inclusion and exclusion criteria, you might add or drop variables describing study characteristics, and you might revise how you code such variables. Even in preregistered meta-analyses, your choices might evolve, as it can be difficult to fully anticipate all complexities up front; such changes are acceptable as long as they are transparently noted, such as whether they were made before or after analyzing the data.

Conducting a meta-analysis generally takes a long time, is tiring, and brings many frustrations, some of which are implied in the list of decisions above and the backtracking that is often required. A meta-analyst is taxed by the mental fatigue of documenting every step in the search and by the work of coding study characteristics and extracting effect sizes, both of which require high cognitive effort and cannot usually be routinized or outsourced because each study presents its own challenges. Learning new statistical methods and software is challenging as well.

The rewards, of course, are great, as a meta-analysis almost always pushes a field forward and receives generous attention. Also, doing a meta-analysis puts one in a new and much more intimate, discerning, and critical relationship with the literature. One is also reminded that no one study, no matter how brilliant it is or how impactful its results are, can ever stand alone. If any activity has the potential to settle questions about replicability (see Fiedler & Ermark, Chapter 3 in this volume), it is meta-analysis.

27.6 Steps in Doing Meta-analysis

A prototypical meta-analysis project has four phases: finding the literature, quantifying (coding) characteristics of the chosen studies, extracting effect sizes that describe the direction and the magnitude of findings, and, then finally, analyzing and reporting on the extracted data. Guides are available that touch on specific topics in detail. However, in our experience, textbooks cannot cover all the nuances and challenges, meaning that a new meta-analyst is advised to have an experienced person available for consultation. Below we discuss a select set of issues that have come up repeatedly in our own practice and teaching.

27.6.1 State a Directional Research Question

As we said at the outset, a starting premise is having a directional research question or hypothesis such as "higher scores on a scale of narcissistic personality will be associated with higher aggression." For every finding in one's database, the outcome of this directional hypothesis needs to be reflected in the sign that the meta-analyst assigns to the given effect. Although it might seem intuitive that the signs for this hypothesis are + for a positive correlation and − for a negative correlation, these are not the only options for choosing a signage scheme. The meta-analyst is entitled to choose a signage scheme, of which there are three types: (1) + (or –) if the effect is consistent with the meta-analyst's specific prediction, – (or +) if not; (2) + (or –) if the effect direction matches the prevailing direction of effects in the database, – (or +) if not; and (3) arbitrary, as in + (or –) for a gender difference favoring women, or – (or +) for a gender difference favoring men. The choice of signage scheme depends on the meta-analyst's theory and preference, and on the studies at hand. The crucial considerations are consistency and accuracy in applying signs to effects. This step can be surprisingly confusing and error-prone, considering that the variables in the various studies may have inconsistent polarities. For instance, a meta-analysis of intervention effects might include both prosocial and antisocial behaviors as eligible outcomes, requiring flipping the effect size sign to be consistent with one's signage scheme.

27.6.2 Define Inclusion/Exclusion Criteria

A very lucky meta-analyst will decide the inclusion criteria at the outset and never have to revise them. But, nearly always, what kinds of study to include evolves as one becomes aware of what kinds of study are in the literature. In preregistered meta-analyses, the authors might preregister the inclusion criteria after piloting them on a small subset of articles, still knowing that elaboration or modification might be needed as the work unfolds; any deviations from the initial plan should be transparently noted. Many inclusion/exclusion criteria will be unique to one's literature, but some generally relevant ones pertain to type of publication source, year, type of population (e.g., minimum acceptable sample size, gender, age, psychiatric status, location), and study design (e.g., experimental versus correlational, pre-post versus post-only, self-report versus behavior).

27.6.3 Search Exhaustively

Over time, expectations have become stricter for the search process and its reporting. Occasionally a meta-analyst can justify using only published studies, but this practice is now rare. Finding unpublished work can be arduous. Fortunately, systematic methods, search engines, and advice are available (Cooper et al., 2019; Polanin et al., 2019). One can also make direct inquiries to authors and post invitations on listservs or other similar outlets.

Some topics are more difficult to search via keywords than others. This scenario can happen when the phenomenon of interest does not have

standard nomenclature attached to it or when the phenomenon of interest is likely to be of minor interest to the original authors and therefore is not highlighted in titles or abstracts. The latter situation applies, for example, to the study of gender differences, which are often reported as an aside or in footnotes, or not reported at all.

27.6.4 Screen Studies for Eligibility

After gathering relevant citations, the next step is screening for eligibility. This step can be arduous, sometimes involving reviewing thousands of citations. Screening typically happens in two stages: (a) abstract screening to eliminate obviously irrelevant studies based on the abstract alone and (b) full-text screening to make final eligibility decisions. Methodologists commonly recommend developing a written protocol to guide decisions at each screening stage, conducting training sessions on the protocol, and meeting regularly as a screening team (Polanin et al., 2019). Deciding whether a given study qualifies for inclusion is often not straightforward and, in the early stages, discussion often produces refinements of the inclusion/exclusion criteria (which can then require backtracking).

Embedding dual screening (i.e., two humans screen the same citation) throughout the process can help assess and resolve differences in interpretation, reducing the chances of erroneous judgments (Waffenschmidt et al., 2019). Though dual screening is resource-intensive, not every record needs to be dual-screened to realize these benefits (e.g., a reviewer could randomly assign 30 percent of citations to be dual-screened). New machine learning tools like Abstrackr (a web-based tool for abstract screening) can also help by prioritizing dual screening for the records that are more likely to be relevant, while reverting back to single (human) screening for less relevant records (Rathbone et al., 2015). We further elaborate on this point when discussing screener and coder reliability.

27.6.5 Make a Codebook for Study Characteristics

A codebook describes the characteristics that coders should extract from eligible studies; this protocol invariably goes through numerous revisions as one discovers what can and cannot be coded from studies. With each revision, one has to backtrack to see whether previously coded studies need recoding; one approach to addressing this issue is to first pilot-test the coding scheme on a small number of studies and then aim to minimize any subsequent changes for the full set of studies.

The variables to be coded fall into two overlapping categories: those that describe the database (populations, designs, measures, and so on) and the subset of those that are intended to be examined as theoretically meaningful moderators, based on a priori reasoning. Sometimes meta-analysts find themselves overwhelmed with moderator analyses; narrowing the focus and returning to the guiding research questions can help in such cases. One can significantly conserve time and resources, for instance, by intentionally *not* coding study characteristics less central to the review's research questions and theoretical goals, resisting the urge to code every aspect that might seem "interesting."

One of the meta-analyst's many executive decisions is to define the phenomena of interest. The chosen definition may lead the meta-analyst to contradict what an original author said about their instrument or intervention. For example, consider a meta-analyst doing a review that focuses on self-report measures of "empathy." Based on reviewing the survey items, an instrument called "compassion" might fit the meta-analyst's operational definition of "empathy" and therefore be included, while an instrument that is called "empathy" might not fit the meta-analyst's definition and therefore be excluded. In other words, the meta-analyst should look at what was done in a study, not at what it was called.

Junior personnel can be engaged for coding straightforward characteristics but reliability checks and careful supervision are still required; finding the information needed to code each item in the codebook can be confusing and error-prone even for experienced meta-analytic coders. Research papers can be surprisingly unsystematic on where, how, and whether they describe the features one wants to code. Journals often have different style requirements, and individual scholars also vary in their preferences. A perennial question is the extent to which the coder can assume a given answer when the original author has not explicitly stated it (e.g., can one assume that a sample is predominantly White if it was selected in Norway?).

Coded moderators can be considered low-inference or high-inference. Low-inference means the information is clearly stated in the study, such as the year of publication or the proportion of women in the sample. High-inference means the coder makes a judgment about what a given study feature likely means to the study's original participants; in this way, altogether new information is added to the database. High-inference coding can be difficult to validate because the coders are typically asked to imagine themselves in the original participants' situation, in order to gain insight into their psychological experience. The meta-analyst therefore relies first on establishing strong inter-rater reliability (sometimes using substantially large groups of raters for the purpose) and second on whether the new moderator elucidates the meta-analytic findings. Eagly & Steffen (1986) used high-inference coding to gain insight into when and why men in their meta-analysis aggressed more than women. A large group of raters read brief descriptions of the context in which the particular kind of aggression occurred and then rated how dangerous it would be for them, themselves, to aggress in that way, in that situation. The gender difference in those ratings significantly predicted the aggression gender difference in the literature, yielding insight into how dangerousness likely factored into the literature's findings. Early on, high-inference coding was viewed with suspicion, but now it is understood to be a valuable source of new information.

27.6.6 Quality Assessment

The quality of studies is of concern for evidence syntheses, in terms of how the reviewer might either exclude low-quality studies or compare low- versus high-quality studies. Some standardized rubrics exist for evaluating study quality and some are developed by the meta-analyst for their particular literature (e.g., Higgins et al., 2019; Wells et al., 2000). Measurement can consist of a checklist of desirable features or a set of evaluative rating scales, with the content including (for example) whether the reliability or validity of instruments was reported (see Revelle & Garner, Chapter 20 in this volume), whether the study had a comparison group, whether there was random assignment, or whether hypothesis blinding was adequately achieved (Valentine, 2009).

27.6.7 Screener and Coder Reliability

Ensuring consistent decisions across team members (i.e., reliability) is important for both the selection and coding of studies. Reliability of coding the selected studies can be based on coding the entire database or a sufficient sample of studies using two or more coders. This process of dual coding should begin *early* and continue *throughout* the review process, not be saved for the end, so that differences in interpretation can be proactively addressed, especially if only a subset of studies is dual-coded. The standard methods for assessing intercoder reliability in primary research are appropriate (e.g., kappa for categorical variables, Pearson correlation or intraclass correlation for continuous variables; see Thorson & West, Chapter 16 in this volume). Disagreements on categorical variables are

generally adjudicated by consensus. When resolving discrepancies on the assessment of effect sizes, it is helpful if the coders have kept good notes on their process, to help reconstruct how they arrived at their respective values.

27.6.8 Extract Effect Sizes

The idea that study results can be put into a standardized metric underpins meta-analysis. What might have seemed incommensurable – the many different ways original authors calculated and expressed their study results – could now be examined systematically, using consistent statistical methods. Hence meta-analysts learn to convert one statistic into another, as needed, using surprisingly simple formulas (e.g., Lakens, 2013; Rosenthal, 1994). To give examples, a *t*-test (which is not an effect size, only a significance test) can be converted to the point-biserial (i.e., Pearson) correlation with a simple formula. The same is true for a two-by-two chi-squared test. Or, given a *p*-value and *N*, one can derive an effect size exactly or estimate it. These conversions are just examples. The point is that the meta-analyst develops versatility in finding the relevant analysis in the original study and then converting it, as needed, to the desired outcome metric. Extracting effect sizes typically requires more experience than very junior assistants can bring to the task; often, indeed, senior authors do this work. Needless to say, reliability checking is very important.

In psychology, two "families" of effect sizes are most commonly used: the "*r*" (correlation) family and the "*d*" (standardized mean difference) family. Both of these describe the relation between two variables in a standardized (unit-free) metric that permits their aggregation (mean, median, and so on) and other statistical manipulations. These two types of effect size are themselves connected by easy mathematical relations, and often one needs to go from one to the other. Within each of these families there are subtypes that we need not describe here because there are many excellent sources available (e.g., Borenstein et al., 2009). Also, other indices of effect exist, for example for comparing two independent correlations, or differences between proportions (Cohen, 1989). Other complexities include whether to use unadjusted or covariate-adjusted statistics (Taylor et al., 2022) and whether to account for clustering (e.g., students nested within schools) in the primary study samples (for formulas, see What Works Clearinghouse (WWC), 2022, Appendix E).

One ubiquitous question is what to do when an effect size is not retrievable despite all efforts (including writing to the authors). Is that study cast aside? Or does one call the effect size "zero" on the assumption it was not significant and was therefore probably negligible in magnitude? Methods for imputation of missing effect sizes are available (Albajes-Eizagirre et al., 2019). Analyzing and reporting the meta-analytic results in all of these ways is informative. Leaving out unknown effects may overestimate the overall effect size, while calling all missing effects "zero" likely underestimates it.

An additional challenge occurs when some of the analyses in the original studies were between-participant and others were within-participant (as in repeated measures, pre-post designs, or matched *t*-tests; Lakens, 2013). Lipsey & Wilson (1993) discussed this issue in their review of meta-analyses on the impact of interventions in psychology. Variability is typically much smaller within participants than between participants. Hence effect sizes computed using within-participant standard deviations will generally be larger than those based on between-participant standard deviations. Combining both types of effect in the same analysis can introduce an artifactual inflation of heterogeneity and distort the meaning of the pooled results. Lipsey and Wilson made the decision to report the two kinds of effect separately and, as expected, the within effects were larger than the between effects. One could

also code within versus between as a moderator and see, empirically, if it matters, or use the between-participant standard deviations as the standardizing term in both types of design (one could also use the pre-post correlation, if available, to make these conversions: WWC (2022), Equation E.19).

27.6.9 Choose Your Analyses

Analytic choices are myriad, constrained to some extent by one's software choices. We will discuss choice of model first, expanding our earlier introduction of the fixed versus random options. Analysis choices, of course, are not entirely either–or, as one can usually do things in multiple ways and compare.

27.6.10 Model Choice

Unweighted models assign equal weight to each effect size, avoiding the potential biasing effects of sample size being confounded with effect size moderators, though at the cost of reduced statistical power compared to weighted models. The weighted random-effects model, which weights studies as a function of the between-studies variation and within-study uncertainty, is the most common at present. The fixed-effects model assumes that there is "one true" population effect size and that all studies are sampled from that same population, with variation in effect sizes due only to uncertainties deriving from sample size. Compared to the weighted random-effects model, this approach yields more extreme differences in weights for smaller versus larger studies. As noted earlier in this chapter, the fixed approach is out of favor because most meta-analyses show high heterogeneity in effect sizes and usually reveal moderator effects, meaning that the assumption of "one true" population effect cannot be supported. Inherently, none of these models is more correct than any other; one's choice has to depend on the database as well as on one's goals.

27.6.11 Corrections for Statistical Artifacts

Depending on the kinds of variable in the database, it may be possible to correct effect sizes for artifacts, the most common being the reliability of measured variables and restriction of range (Schmidt, Le, and Oh, 2009). For example, if the meta-analysis is about the correlation between narcissism and aggression, one could correct every effect size for how reliable each scale is so that one's aggregate effect size is what one would get if both variables had perfect reliability (see Revelle & Garner, Chapter 20 in this volume). Not all meta-analytic contexts are suitable for this kind of correction, and not all meta-analysts are on board with generating meta-analytic results that are ideal rather than the ones actually obtained in the literature. Insight into statistical artifacts can also be gained by including scale reliability as a moderator and seeing what impact it has (Schlegel et al., 2017).

27.6.12 Multiple Effect Sizes within Studies

Traditional meta-analysis models typically assume that the effects are independent of each other. However, honoring this assumption is a challenge when a single study contributes multiple effect sizes, as is often the case. Sometimes an executive review decision solves the problem; for example, if some intervention studies measure the outcome at several follow-up timepoints while others measure the outcome only once, the meta-analyst might decide, a priori, to include only the timepoint that is most temporally comparable across studies. However, in many cases, the meta-analyst will be interested in multiple effect sizes per study. For example, the study might examine narcissism in relation to both verbal and physical aggression, producing two effect sizes for the same overarching question. The analyst must account for this multiplicity in some

way because otherwise the independence assumption is violated and the meta-analyst may find "significant" results when no true effect exists (i.e., yielding high false-positive rates; López-López et al., 2017).

Several approaches exist for addressing this issue (also often called *effect size dependencies*). For instance, the analyst could segregate the effects, conducting one meta-analysis for studies measuring verbal aggression and another for studies measuring physical aggression. Though some studies would appear in both meta-analyses, each meta-analysis would have only one effect size per study. However, this approach breaks down if there are other forms of dependency. For instance, the study could report correlations for verbal and physical aggression, each measured with two different scales, yielding two effect sizes for each type of aggression, making four effects in total. Or a study might report male and female subsamples, again yielding multiple effect sizes that may introduce dependencies. Separate meta-analyses of verbal and physical aggression would no longer have one effect size per study in these cases. The analyst might also be interested in investigating whether correlations with narcissism are stronger for verbal versus physical aggression, which is not possible when conducting separate meta-analyses.

Recent approaches for addressing these issues have therefore focused on creating flexible analysis models that explicitly account for multiple effect sizes per study (Tipton et al., 2019). These flexible analytic approaches are especially attractive when the analyst wants to understand the heterogeneity of effects both within and across studies (as when comparing correlations for verbal versus physical aggression). One such approach is called *robust variance estimation* (Pustejovsky & Tipton, 2022), which provides appropriate meta-analytic standard errors and significance tests, even in the presence of multiple effect sizes per study. The approach builds on prior flexible tools from multilevel and multivariate meta-analysis (e.g., Cheung, 2014) while offering practical advantages in terms of what information users need in order to use it. For instance, earlier multivariate approaches required the user to exactly specify the correlations between outcomes within a study to yield appropriate inferences, but studies often do not report such correlations. Robust variance estimation relaxes such requirements by generating inferences that are still appropriate even if the exact correlations between outcomes are unknown. Multiple software tutorials exist for implementing robust variance estimation in SPSS, Stata, and R (Tanner-Smith & Tipton, 2014; Tanner-Smith et al., 2016), though the most recent methodological developments exist in R, especially for handling small-study adjustments (see Pustejovsky & Tipton, 2022).

Simpler approaches for handling multiple effect sizes per study can sometimes be appropriate depending on the meta-analytic goal and research question (for a review, see López-López et al., 2018). A commonly used approach is to average effect sizes within studies to generate one effect size per study (Borenstein et al., 2009). Averaging effect sizes is inappropriate if one is interested in heterogeneity within studies but can be sensible if the meta-analysis focuses on study-level moderators (e.g., publication year) or if there are no meaningful conceptual differences between the several effects from one study.

27.6.13 Choosing Software

Early on (and still, when some unweighted models are used), meta-analysts used no special software beyond their standard packages such as SPSS. Special software that includes algorithms for weighting are available, both commercially (e.g., Biostat's Comprehensive Meta-Analysis) and for free public use (e.g., the *metafor* package in R: Viechtbauer, 2010). Additional macro routines can also be downloaded for specific purposes. Meta-analysts may use more than one

software package to suit their purposes. Hence multiple options are possible, though the latest developments tend to become available first in the R software, as it supports an active community of open-source developers (Polanin et al., 2017). Multiple resources exist for learning meta-analysis approaches in R (see, e.g., www.meta-analysis-research-institute.com/learning-information-center). Regardless of the software package used, meta-analysts should carefully document their code and make it publicly available upon publication to ensure that others can reproduce the findings (see Crandall, Giner-Sorolla, & Biernat, Chapter 2 in this volume). We say more about this in the section 27.8, Presentation.

27.7 Analyses

27.7.1 Study Characteristics

The meta-analyst should describe study characteristics in terms of frequencies or percentages of studies occurring within every coded category. Often, insufficient attention is given to examining correlations among potential moderators; these correlations can be substantial and, if left unaccounted for, can seriously bias the interpretation of moderator effects (e.g., see discussion in Del Re & Flückiger, 2016). One such correlation that is rarely reported is that between study sample size and study characteristics. This confound is potentially biasing because of the ubiquitous use of sample size in weighting formulas. Meta-analysts must ascertain, and appropriately control for, correlations between study size and study characteristics.

27.7.2 Central Tendency

Typically, one's first substantive question is, What is the overall effect size? The mean effect size can be weighted or unweighted, reported with and without outliers, corrected for statistical artifacts or not, and reported with unknown effect sizes excluded or included as "zero" or imputed. The median effect size is also informative and should be reported. As said earlier, tests of significance for the central tendency estimate are routinely done.

27.7.3 Heterogeneity

The standard deviation of the observed effect size distribution is one simple metric for describing variation in effect sizes. However, this metric conflates two sources of variability: (a) differences in true underlying effects across studies and (b) sampling error for each study. Consider neuroscience research, for instance, which typically has small sample sizes (Button et al., 2013). Each study's estimate therefore might be quite noisy. The simple standard deviation across studies could mislead the meta-analyst by suggesting that effects are quite variable across studies when actually this variability might be due to chance alone.

An alternative metric that addresses this interpretation issue is the *between-studies heterogeneity*, sometimes also referred to by the Greek letter tau (τ). This metric aims to estimate the standard deviation of the true underlying effects across studies, rather than of the noisy effect size estimates. This parameter has theoretical meaning as it describes how much the psychological effects substantively vary, independent of the sample sizes used to estimate those effects (Borenstein et al., 2017). An estimated heterogeneity of 0 means that the observed variance in effect size estimates can be attributed to chance alone. This heterogeneity parameter can be estimated with or without weights; common meta-analysis software applications routinely provide this parameter in model output or provide the variance τ^2, which can be converted to τ by taking the square root (Viechtbauer, 2010).

27.7.4 Moderators

As we have said, one of the most informative aspects of a meta-analysis is understanding how specific study characteristics can explain variability in effect sizes. Are effect sizes larger for verbal or physical aggression? Do self-reports differ from direct observations? These analyses are analogous to tests of moderation in primary research, which are typically addressed by looking at interaction effects. In investigating moderators, the meta-analyst must remember the earlier point about confounds; for instance, studies that use verbal or physical aggression might differ in other ways than just the type of aggression measured. For example, studies of verbal aggression might have a higher proportion of female participants than studies of physical aggression. Meta-analysts need to maintain the same vigilance for confounders that they would have in their primary research.

The analyst can account for these confounds through regression-based approaches that include multiple predictors in the same model, which is often called *meta-regression* (Tipton et al., 2019). (Such regressions can be weighted or unweighted.) Controlling for other variables can help, though other confounders may still be unaccounted for. The analyst should consider always controlling for methods confounders (e.g., study design, effect size computation details) that are not central to the substantive research questions but could contaminate the results of interest. Testing one moderator at a time does not account for these considerations regarding confounding. Simple tests of moderation can be a useful starting point but, for robustness, should be followed up with approaches such as regression that help address confounding. More advanced topics for investigating moderators include ways to build multiple-predictor models (Cinar et al., 2021) or use machine learning to automatically model nonlinearities and interactions between predictors (Van Lissa, 2020).

Moderator analyses are often underpowered, especially for small numbers of studies (Hedges & Pigott, 2004). This concern is particularly salient when conducting a large number of moderator tests; the analyst could find a stray significant moderator and herald that result as a main finding, but that result might just be due to chance. Following Tipton et al.'s (2019) recommendations, analysts should therefore distinguish confirmatory and exploratory analyses, limiting the confirmatory analyses to just a small number of moderators theorized to be most important a priori.

27.7.5 Selective Reporting Bias

Worry about the representativeness of the studies in one's database has long plagued meta-analysts (Marks-Anglin & Chen, 2020; Siegel et al., 2021), a concern also called *publication bias*. Results that are very small or nonsignificant, or that run counter to accepted wisdom, may remain, unpublished, in the researchers' proverbial "file drawers." Researchers might lack interest or be disappointed in their nonsignificant findings, perceive that journals only publish significant results, or perceive that others would lack interest, leading to selective reporting even before manuscript writing. Journal editors and peer reviewers might then also exert pressure to present a "clean" story (e.g., omitting nonsignificant findings) or reject studies with nonsignificant findings. These types of omission can occur at both the publication level (entire studies not being published) and the results level (studies being published, but not all outcomes or analyses from them). If the published literature only contains statistically significant results, then the meta-analyst may form biased conclusions.

Concerns over selective reporting have intensified in recent years, especially in the field of social psychology, given empirical evidence on the lack of reproducibility of effects in the published psychological literature (Open Science

Collaboration, 2015). As one example, Kvarven et al. (2019) compared the results of fifteen meta-analyses (potentially subject to selective reporting, as all meta-analyses typically are) to large-scale, multi-laboratory preregistered replications on the same psychological topic (not subject to selective reporting). Average effect sizes were systematically larger in the meta-analyses than in the preregistered replications on the same topic. Though multiple explanations are possible, selective reporting in the meta-analyzed primary literature may partly explain the discrepancy. A larger-scale review of meta-analyses estimated the severity of selective reporting bias across several fields (Mathur & VanderWeele, 2021). For top psychology journals, statistically significant results in the expected direction were estimated to be 1.54 times as likely to be reported relative to nonsignificant results or those in the unexpected direction. This estimated severity of bias was larger for psychology than for other fields such as medicine or other social science fields. These biases can contaminate conclusions in meta-analyses as well as those when reading individual studies.

Several methods exist for investigating selective-reporting bias in meta-analysis. These meta-analytic tools provide at least one advantage compared to reading individual studies, for which reporting bias may exist, but tools do not readily exist to investigate such biases when reading an individual study. Recent advances have emphasized moving beyond simply *detecting* selective-reporting bias (e.g., testing for evidence of bias) toward *adjusting* for bias (e.g., re-estimating average effect sizes adjusted for bias). This shift is an important one, given that significance tests for detecting selective-reporting bias depend heavily on the number of included studies. With few studies, a significance test for bias could be nonsignificant, even if selective reporting is rampant. With many studies, a significance test could be significant, even if selective reporting barely changes the meta-analytic results. Methods to adjust for bias instead focus the meta-analyst's attention on the question, How important is selective reporting to the meta-analytic conclusions I want to draw?

Before describing specific methods, we emphasize that these bias adjustment methods are not a panacea. They should not be viewed as providing definitive estimates of "true" corrected effects, as much of the methodological literature has highlighted the limitations of these methods, especially when effect heterogeneity is large (Carter et al., 2019). Hence they are not replacements for improving reporting practices in the primary literature (e.g., via preregistration) or for conducting a thorough search for unpublished literature. These bias adjustment methods, however, are appropriate as *sensitivity analyses*, helping the meta-analyst and readers understand how sensitive the results are to the chosen set of assumptions and modeling strategies. Multiple adjustment options are often defensible in a given context, and the meta-analyst should consider the range of plausible adjusted estimates to better understand the robustness of results.

Approaches for investigating selective-reporting bias largely fall into four families of methods: (a) comparison of published and unpublished studies, (b) methods based on small-study effects, (c) methods based on *p*-values, and (d) sensitivity analyses without estimating bias. These four basic categories have existed for decades, with the available methods within each category significantly evolving over time.

27.7.5.1 Comparison of Published and Unpublished Studies

The simplest method for investigating selective-reporting bias is to compare effect sizes from published versus unpublished studies. Though meta-analysts apply differing definitions of "published" and "unpublished," published studies generally include peer-reviewed journal articles or edited book chapters, whereas unpublished studies generally include master's theses,

dissertations, unpublished manuscripts, and investigators' unpublished data sent on request. Analysts should keep three caveats in mind, however. The first is that published versus unpublished studies can differ in other respects, such as differences in the methods used. This problem can be addressed by coding for important aspects of the study methodology and controlling for them in meta-regression models. Evidence of bias is strengthened if the published/unpublished difference remains even after controlling for other potential confounds. The second caveat is that the retrieved unpublished results may not resemble the universe of all unpublished results, which often they may not. The third caveat is that, within a retrieved unpublished work such as a doctoral dissertation, there could still be selective reporting at the results level. Failing to find a published/unpublished difference should therefore be interpreted cautiously for these reasons (Siegel et al., 2021).

27.7.5.2 Methods Based on Small-Study Effects

A large family of methodological approaches rests on the assumption that smaller studies might show larger effects in the published literature due to selective reporting (however, as noted below, these small-study effects can occur for other reasons too). The rationale is that small studies with small observed effects fail to reach statistical significance and are therefore not published. In contrast, large studies with small effects are still published because they have statistically significant results. Early methods exploiting this idea focused on *detecting* bias, such as through "funnel plots" of effect sizes graphed against some function of the study size to determine whether small studies with small effects were selectively missing (Sterne et al., 2005).

More recent advances have focused on *adjusting* for bias by leveraging the small-study effects assumption. For instance, the precision-effect test (PET) and precision-effect estimate with standard error (PEESE) methods involve extrapolating to a theoretical population of studies with infinite sample size (Stanley & Doucouliagos, 2014). These approaches do so by including the effect size standard error or variance as predictors in a meta-regression model (see Pustejovsky & Rodgers, 2019, for further improvements of this approach). The key assumption is that studies with infinite sample size do not suffer from selective-reporting bias because they always yield statistical significance (if a true effect exists). This meta-regression approach has the advantages of being easy to apply (i.e., uses the same modeling approach as a moderator analysis) and of performing better than some other analogous bias adjustment options based on the small-study assumption (e.g., the trim-and-fill procedure, which performs poorly in methodological research, despite being widely used in practice; see Carter et al., 2019).

A key limitation of these methods (e.g., including for both funnel plots and more recent regression-based approaches) is assuming that the only explanation for small-study effects is selective-reporting bias. In contrast, smaller and larger studies could differ in other ways, as we explained earlier, such as the methods or participant populations they use. Similar to published/unpublished differences, analysts can partially address this issue by coding for those other potential characteristics and controlling for them via meta-regression. Nevertheless, this point reiterates that these bias adjustments should be viewed as a sensitivity analysis, not as definitive estimates of true corrected effects.

27.7.5.3 Methods Based on P-Values

Other methods, called *selection models*, more explicitly model that reporting decisions often depend on the *p*-value (McShane et al., 2016). For instance, finding $p = 0.049$ in the expected direction may be more publishable than

$p = 0.051$, which may be more publishable than $p = 0.131$ or findings in the unexpected direction. Selection models explicitly capture this dependence on the *p*-value and adjust meta-analytic mean estimates for bias, showing favorable methodological performance across many contexts relevant to meta-analysis in psychology (Carter et al., 2019). Selection models have existed for decades (e.g., Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995) and have become more widely used in recent years due to new software products allowing for easier use, such as the *weightr* R package (Coburn & Vevea, 2019) or the new *selmodel* function in the metafor R package (Bartoš et al., 2022). For instance, an interactive website allows users to upload their data and apply this approach without requiring the analyst to have a coding background (https://vevealab.shinyapps.io/WeightFunctionModel).

More recently proposed methods based on the *p*-value such as *p*-curve analysis (Simonsohn et al., 2014) and *p*-uniform (Van Aert et al., 2016) are closely related in principle to selection models. However, *p*-curve analysis has fundamentally distinct inferential goals, which has led to some controversy in applying the method. Unlike the methods previously described, *p*-curve analysis discards findings of $p > 0.05$ and explicitly does not aim to make inferences about them (Simonsohn et al., 2017). Detailing the full reasons and implications for this distinction is complex and beyond the scope of this chapter, but we emphasize three key points here. First, this distinction matters when true effects vary from study to study (i.e., there is effect heterogeneity), which characterizes most psychological research, as previously discussed. Second, *p*-curve analysis can be substantially biased if one's inferential goal is to estimate population means for all studies conducted on a topic (Carter et al., 2019). Third, before applying the method, readers should consult further literature to better understand these inferential goals, including through critiques of *p*-curve analysis (e.g., McShane et al., 2016) and responses to those critiques (e.g., Simonsohn et al., 2018).

27.7.5.4 Sensitivity Analyses without Estimating Bias

Other methods start with question, How bad does selective reporting need to be to change the meta-analytic conclusions? Rather than empirically estimating bias or correcting for it, these methods instead focus on the sensitivity of meta-analytic results if selective reporting exists. One such approach is to estimate the number of unpublished "file-drawer" studies with results that are, on average, zero that would have to exist to render the obtained meta-analytic average non-significant (Rosenthal, 1979). This number, called either the *file-drawer N* or the *failsafe N*, is commonly used and offers simplicity, but methodologists have noted that it suffers from limitations such as the arbitrariness of defining a "large" file-drawer *N* and questionable assumptions about the average effect size of unretrieved studies (Becker, 2005; Siegel et al., 2021). More recent innovations have improved on these limitations, shifting attention away from considering an arbitrary number of unpublished studies and instead toward directly considering how severe selective reporting needs to be to change the meta-analytic conclusions (Mathur & VanderWeele, 2020).

27.8 Presentation

The reporting of meta-analyses should aim to be *transparent* and *interpretable*. Being transparent means comprehensively noting key decisions, publishing data files and analysis scripts, and overall ensuring that others can reproduce the process based on the reported materials. Being interpretable means interpreting effect sizes with sensible benchmarks, describing heterogeneity, and using graphs to display results where appropriate.

The current reporting of meta-analyses in psychology often falls short of these standards (Polanin et al., 2020). As a result, readers can struggle to understand what was done, what was learned, or how to reproduce the findings, threatening the credibility and usability of the meta-analysis (Lakens et al., 2016). Fortunately, however, meta-analysis reporting practices have improved over time (Polanin et al., 2020), and several helpful guides and resources exist to aid meta-analysts in writing transparent and interpretable manuscripts.

27.8.1 Comprehensively Note Key Decisions

External reporting guidelines play a critical role in improving transparent reporting by helping authors comprehensively note key decisions. One widely used set of such guidelines are the PRISMA guidelines noted earlier (www.prisma-statement.org), which include a two-page checklist for reporting methods details and figure templates for reporting the flow of studies during eligibility screening. A PRISMA elaboration and explanation document also provides many examples that follow these guidelines (Page et al., 2021). Authors can include a completed checklist with page numbers in the supplemental materials, helping peer reviewers and readers find key information. Another helpful resource is the Meta-Analysis Reporting Guidelines (MARS), though the MARS guidelines are generally less extensive than the PRISMA guidelines (American Psychological Association, 2020).

Figure 27.1 shows an example PRISMA flowchart from the second author's ongoing meta-analysis on children's gender stereotypes about academic abilities (see https://osf.io/8ktnj for the preregistration). Moving from top to bottom, the figure reports the number of citations identified through three search methods (18,490 in total), the number excluded at abstract screening, the number excluded due to full-text unavailability, the number excluded at full-text screening (with specific exclusion reasons recorded and counted), and then finally the number of reports and studies included (the same study sample sometimes appeared across multiple reports).

27.8.2 Publish Data Files and Analysis Scripts

Meta-analysis authors should publish the extracted data and analysis scripts on a repository such as the Open Science Framework website or the Inter-university Consortium for Political and Social Research website. Doing so can help others better understand the analytic procedures and decisions, catch errors, reproduce the findings, interrogate the robustness of the findings, and extend the findings in new analyses, similar to the benefits of open science in primary research (Lakens et al., 2016). Though meta-analyses in psychology commonly report the extracted effect size data, they rarely provide the code used to analyze the data (Polanin et al., 2020). This current state of reporting is a critical limitation as both the data and analysis code are vital to reproducing the meta-analytic findings. Sometimes study authors may provide the meta-analyst with additional unpublished data and may be concerned about sharing the individual-level data set more broadly; however, the meta-analyst can mitigate this concern by publishing the effect size summary statistics necessary to reproduce meta-analytic results without publishing the individual-level data sets by study.

27.8.3 Use Effect Size Benchmarks

Beyond making the methods transparent, authors should help readers interpret the meta-analytic findings in broader context. One such approach is to compare average effect sizes to those found in related research contexts. For instance, Lovakov & Agadullina (2021) compiled the mean effects sizes from 134 meta-analyses in

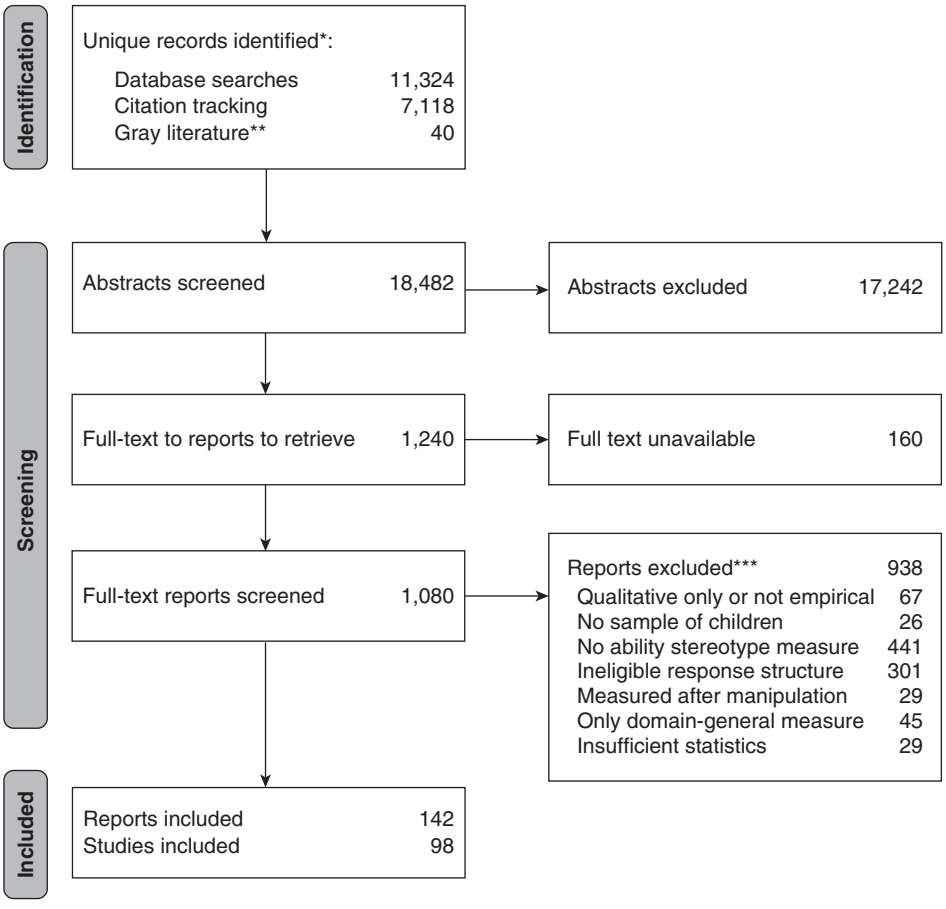

Figure 27.1 Flow of citations through the literature search and screening

* These counts already remove exact duplicates of reports. Each row is the number of unique new records identified compared to the previous rows (e.g., number of new records identified by backward citations that were not identified by keyword searches or forward citations).

** In practice, the number reviewed for grey literature searches was higher due to sources like conference programs where tracking the exact number of records reviewed was intractable.

social psychology, providing empirical guidelines for effect size interpretation. Overall, median effect sizes were 0.24 for correlation coefficients and 0.36 for standardized mean differences, though there was also considerable variation across research topics. In practice, a broad benchmark such as this should be supplemented by benchmarks that are specific to the research area in which one's meta-analysis is situated. For example, the meta-analytic gender difference in smiling was compared by Hall (2006) to the obtained effect size for other social psychological gender differences and to that found for other correlates of smiling.

27.8.4 Describe Heterogeneity

Interpreting average effects is only one goal. Understanding how effects vary across studies can be just as theoretically important, if not

more so. Authors can help readers understand the overall magnitude of heterogeneity through prediction intervals (for details, see Borenstein et al., 2017). Prediction intervals can answer questions such as, What is the middle 95 percent of true underlying effects across studies? Then, to describe specific sources of heterogeneity, authors should return to their guiding research questions to examine moderators, as we have discussed already.

27.8.5 Use Graphics

Graphics can be a powerful way to display the central tendency, variability, or relationships in meta-analytic data. Kossmeier et al. (2020) provided a comprehensive overview of more than 200 graph types developed or used for displaying data in systematic reviews and meta-analyses. Although the sheer number of options can be overwhelming, the options fell into several commonly reoccurring categories. Meta-analysis authors could scan the vignettes presented in Kossmeier et al. (2020) for inspiration and select among the options that seem most relevant for their research question. Figure 27.2 shows one way to visualize the effect sizes from each study and their uncertainty using a rainforest plot, which aims to improve on more traditional displays called forest plots (Schild & Voracek, 2015). The graph represents uncertainty by stretching the distribution for each data point. Data points with longer, lighter shading (such as study 6) indicate less precise estimates (i.e., smaller samples) relative to data points with shorter, darker shading (such as study 13).

27.9 Conclusions

Meta-analysis has proven invaluable in summarizing literatures, settling theoretical and methodological disputes, inspiring new research directions, and helping to correct for replication and other challenges facing social and personality psychology. We have emphasized how important meta-analysis is in establishing lasting and valid conclusions about research literatures, an especially important goal in an era marked by extreme proliferation of research and skepticism about what we have learned from it.

In closing, we wish to say that the benefits to the meta-analysts themselves are great as well. Aside from the high visibility typically attained by publishing a meta-analysis, the meta-analyst acquires habits of mind that can forever change how they read not just meta-analyses but also primary research studies in general. Those who have gone through the process attest to developing a stronger focus on an author's methods and data, with less reliance on authors' summary statements. For us, at least, doing meta-analysis bestows both heightened skepticism about the evidentiary value of individual studies and authors' claims, and optimism for the future of our field. We wish to thank the many methodologists who have developed these priceless tools, and we hope that new investigators will take on the crucial task of performing meta-analyses in their subject areas.

References

Albajes-Eizagirre, A., Solanes, A., and Radua, J. (2019). Meta-analysis of non-statistically significant unreported effects. *Statistical Methods in Medical Research*, 28(12), 3741–3754.

American Psychological Association. (2020). Meta-analysis reporting standards (MARS), https://apastyle.apa.org/jars/quant-table-9.pdf.

Bartoš, F., Maier, M., Quintana, D. S., and Wagenmakers, E. J. (2022). Adjusting for publication bias in JASP and R: Selection models, PET-PEESE, and robust Bayesian meta-analysis. *Advances in Methods and Practices in Psychological Science*, 5(3), 25152459221109259.

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, and M. Borenstein (eds.) *Publication Bias in*

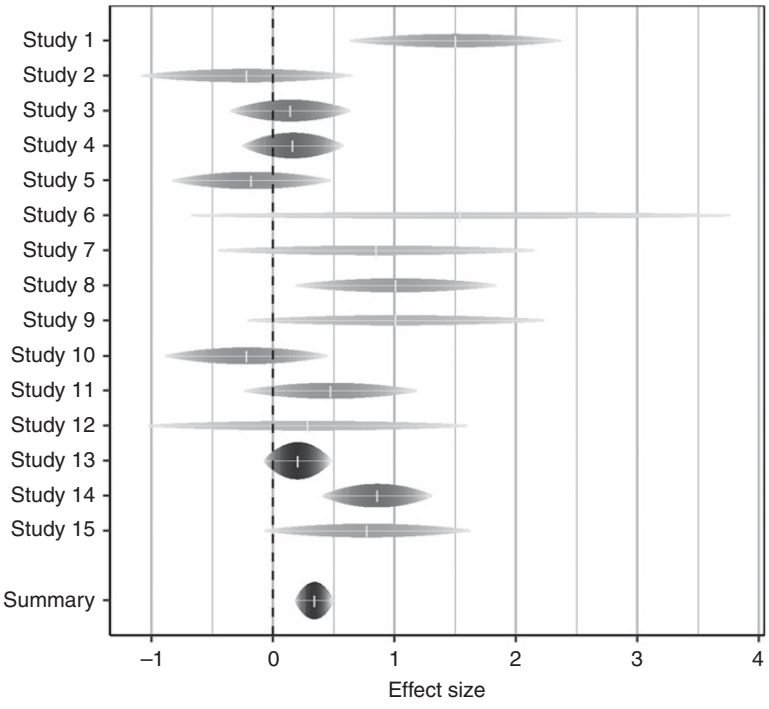

Figure 27.2 Example rainforest plot. The lengths of the ovals ("raindrops") represent the 95 percent confidence intervals for each study's effect size estimate or the overall meta-analytic mean (bottom raindrop). The shadings within the raindrops indicate the probability density within those confidence intervals, and the inner white vertical ticks represent the point estimates. Data points with longer, lighter shading (such as study 6) indicate less precise estimates (i.e., smaller samples) relative to data points with shorter, darker shading (such as study 13). BMC is the original publisher of this figure, which was Figure 12.7 in the supplemental materials from Kossmeier et al. (2020). Reproduction of figures or tables from any article is permitted free of charge and without formal written permission from the publisher or the copyright holder, provided that the figure or table is original, that BMC is duly identified as the original publisher, and that proper attribution of authorship and the correct citation details are given as acknowledgment. If you have any questions about reproduction of figures or tables please contact us

Meta-analysis: Prevention, Assessment and Adjustments. John Wiley and Sons, Ltd.

Borenstein, M. (2019). *Common Mistakes in Meta-analysis and How to Avoid Them*. Biostat, Inc.

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

Borenstein, M., Higgins, J. P. T., Hedges, L. V., and Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1), 5–18.

Bushman, B. J., and Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges, and J. C. Valentine (eds.) *The Handbook of Research Synthesis and Meta-analysis*. Russell Sage Foundation.

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., and 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.

Cheung, M. W. L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavioral Research Methods*, 46(1), 29–40.

Cinar, O., Umbanhowar, J., Hoeksema, J. D., and Viechtbauer, W. (2021). Using information-theoretic approaches for model selection in meta-analysis. *Research Synthesis Methods*, 12(4), 537–556.

Cleophas, T. J. M., and Zwinderman, A. H. (2017). *Modern Meta-analysis: Review and Update of Methodologies*. Springer.

Coburn, K. M., and Vevea, J. L. (2019). Weightr: Estimating weight-function models for publication bias. R package version 2.0.2, https://CRAN.R-project.org/package=weightr.

Cohen, J. (1989). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates.

Cooper, H., Hedges, L. V., and Valentine, J. C. (eds.). (2019). *The Handbook of Research Synthesis and Meta-analysis*. Russell Sage Foundation.

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

Del Re, A. C., and Flückiger, C. (2016). Meta-analysis. In J. C. Norcross, G. R. VandenBos, D. K. Freedheim, and B. O. Olatunji (eds.) *APA Handbook of Clinical Psychology: Theory and Research*. American Psychological Association.

Dickens, L. R., and Robins, R. W. (2022). Pride: A meta-analytic project. *Emotion*, 22(5), 1071–1087.

DiMatteo, M. R. (2004). Social support and patient adherence to medical treatment: A meta-analysis. *Health Psychology*, 23, 207–218.

Donnelly, K., and Twenge, J. M. (2017). Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*, 76(9), 556–565.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, and M. Borenstein (eds.) *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. John Wiley and Sons.

Eagly, A. H., and Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin*, 100(3), 309–330.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, 33(5), 517.

Freudenberg, M., Albohn, D. N., Kleck, R. E., Adams, R. B., Jr., and Hess, U. (2020). Emotional stereotypes on trial: Implicit emotion associations for young and old adults. *Emotion*, 20(7), 1244–1254.

Goh, J. X., Hall, J. A., and Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, 10, 535–549.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 85, 845–857.

Hall, J. A. (2006). How big are nonverbal sex differences? The case of smiling and nonverbal sensitivity. In K. Dindia and D. J. Canary (eds.) *Sex Differences and Similarities in Communication*. Lawrence Erlbaum Associates Publishers.

Hall, J. A., Coats, E. J., and Smith LeBeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131, 898–924.

Hall, J. A., and Rosenthal, R. (2018). Choosing between random effects models in meta-analysis: Units of analysis and the generalizability of obtained results. *Social and Personality Psychology Compass*, 12(10), article e12414.

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, and J. C. Valentine (eds.) *The Handbook of Research Synthesis and Meta-analysis*, 2nd ed. Russell Sage Foundation.

Hedges, L. V., and Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426–445.

Hedges, L. V., and Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486–504.

Higgins, J. P., Savović, J., Page, M. J., Elbers, R. G., and Sterne, J. A. (2019). Assessing risk of bias in a randomized trial. In *Cochrane Handbook for Systematic Reviews of Interventions*, 205–228, at https://training.cochrane.org/handbook.

Iyengar, S., and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3(1), 109–135.

Johnson, B. T. (2021). Toward a more transparent, rigorous, and generative psychology. *Psychological Bulletin*, 147(1), 1–15.

Johnson, B. T., and Eagly, A. H. (2014). Meta-analysis of research in social and personality psychology. In H. T. Reis and C. M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

Kenny, D. A., and 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.

Konrath, S. H., O'Brien, E. H., and Hsing, C. (2011). Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review*, 15(2), 180–198.

Kossmeier, M., Tran, U. S., and Voracek, M. (2020). Charting the landscape of graphical displays for meta-analysis and systematic reviews: A comprehensive review, taxonomy, and feature analysis. *BMC Medical Research Methodology*, 20(1), 1–24.

Kvarven, A., Strømland, E., and Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behavior*, 4, 423–434.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(26), 863, DOI:10.3389/fpsyg.2013.00863.

Lakens, D., Hilgard, J., and Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4(1), 1–10.

Lipsey, M. W., and Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181–1209.

Lipsey, M. W., and Wilson, D. B. (2001). *Practical Meta-analysis*. Sage.

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

López-López, J. A., van den Noortgate, W., Tanner-Smith, E. E., Wilson, S. J., and 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.

Lovakov, A., and Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3), 485–504.

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730–749.

Marks-Anglin, A., and Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, 11(6), 725–742.

Mathur, M. B., and VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5), 1091–1119.

Mathur, M. B., and VanderWeele, T. J. (2021). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, 12(2), 176–191.

Miller, D. I., Nolla, K. M., Eagly, A. H., and Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of five decades of U.S. Draw-A-Scientist studies. *Child Development*, 89(6), 1943–1955.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and the PRISMA Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, 6(7), e1000097.

Mosteller, F. M., and Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (ed.) *Handbook of Social Psychology*, vol. 1, *Theory and Method*. Addison-Wesley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, n160.

Page, M. J., Moher, D., and McKenzie, J. E. (2022). Introduction to PRISMA 2020 and implications for research synthesis methodologists. *Research Synthesis Methods*, 13(2), 156–163.

Polanin, J. R., Hennessy, E. A., and Tanner-Smith, E. E. (2017). A review of meta-analysis packages in R. *Journal of Educational and Behavioral Statistics*, 42(2), 206–242.

Polanin, J. R., Hennessy, E. A., and Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, 15(4), 1026–1041.

Polanin, J. R., Pigott, T. D., Espelage, D. L., and Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, 10(3), 330–342.

Pustejovsky, J. E., and Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71.

Pustejovsky, J. E., and Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23, 425–438.

Rathbone, J., Hoffmann, T., and Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews*, 4(1), 1–7.

Razpurker-Apfeld, I., and Shamoa-Nir, L. (2021). Is an outgroup welcome with open arms? Approach and avoidance motor activations and outgroup prejudice. *Journal of Experimental Psychology: Applied*, 27(2), 417–429.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper and L. V. Hedges (eds.) *The Handbook of Research Synthesis*. Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118, 183–192.

Rosenthal, R., Rosnow, R. L., and Rubin, D. B. (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge University Press.

Rosenthal, R., and Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, 86, 1165–1168.

Schild, A. H., and Voracek, M. (2015). Finding your way out of the forest without a trail of bread crumbs: development and evaluation of two novel displays of forest plots. *Research Synthesis Methods*, 6(1), 74–86.

Schlegel, K., Boone, R. T., and Hall, J. A. (2017). Individual differences in interpersonal accuracy: A multi-level meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior*, 41, 103–137.

Schlegel, K., Palese, T., Mast, M. S., Rammsayer, T. H., Hall, J. A., and Murphy, N. A. (2020). A meta-analysis of the relationship between emotion recognition ability and intelligence. *Cognition and Emotion*, 34(2), 329–351.

Schmid, C. H., Stijnen, T., and White, I. R. (eds.) (2021). *Handbook of Meta-analysis*. Routledge/ Taylor and Francis Group.

Schmidt, F. L., Le, H., and Oh, I. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, and J. C. Valentine (eds.) *The Handbook of Research Synthesis and Meta-analysis*, 2nd ed. Russell Sage Foundation.

Schmidt, F. L., Oh, I.-S., and Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical

comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, 62(1), 97–128.

Shuster, J. J, Guo, J. D., and Skyler, J. S. (2012). Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods*, 3, 30–50.

Siddaway, A. P., Wood, A. M., and Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, 70, 747–770.

Siegel, M., Eder, J. S. N., Wicherts, J. M., and Pietschnig, J. (2021). Times are changing, bias isn't: A meta-meta-analysis on publication bias detection practices, prevalence rates, and predictors in industrial/organizational psychology. *Journal of Applied Psychology*, 107(11), 2013–2039.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547.

Simonsohn, U., Simmons, J., and Nelson, L. (2017, June 15). Why *p*-curve excludes ps > .05 (blog post), https://datacolada.org/61.

Simonsohn, U., Simmons, J., and Nelson, L. (2018, January 8). P-curve handles heterogeneity just fine (blog post), https://datacolada.org/67.

Smith, M. L., and Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32(9), 752–760.

Stanley, T. D., and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78.

Sterne, J. A. C., Becker, B. J., and Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, and M. Borenstein (eds.) *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. John Wiley & Sons.

Tanner-Smith, E. E., and 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.

Tanner-Smith, E. E., Tipton, E., and 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, 85–112.

Taylor, J. A., Pigott, T., and Williams, R. (2022). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher*, 51(1), 72–80.

Thomas, J., and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 45, at https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-8-45.

Tipton, E., Pustejovsky, J. E., and Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, 10(2), 161–179.

Tucker-Drob, E. M., Brandmaier, A. M., and Lindenberger, U. (2019). Coupled cognitive changes in adulthood: A meta-analysis. *Psychological Bulletin*, 145(3), 273–301.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, and J. C. Valentine (eds.) *The Handbook of Research Synthesis and Meta-analysis*, 2nd ed. Russell Sage Foundation.

Valentine, J. C. (2012). Meta-analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (eds.). *APA Handbook of Research Methods in Psychology*, vol. 3. American Psychological Association.

Valentine, J. C., Pigott, T. D., and Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247.

van Aert, R. C. M., Wicherts, J. M., and van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p*-values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, 11(5), 713–729.

van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using MetaForest. In R. Van De Schoot and M. Miočević (eds.) *Small*

sample size solutions (open access): A guide for applied researchers and practitioners. CRC Press, www.crcpress.com/Small-Sample-SizeSolutions-Open-Access-A-Guide-for-Applied-Researchers/Schoot-Miocevic/p/book/9780367222222.

Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419–435.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., and Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology*, 19(1), 1–9.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., and Tugwell, P. (2000). The Newcastle–Ottawa Scale (NOS) for assessing the quality of non-randomized studies in meta-analysis, www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

What Works Clearinghouse. (2022). *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, https://ies.ed.gov/ncee/wwc/handbooks.

White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. V. Hedges, and J. C. Valentine (eds.) *The Handbook of Research Synthesis and Meta-analysis*, 2nd ed. Russell Sage Foundation.

Zuckerman, M., Li, C., and Hall, J. A. (2016). When men and women differ in self-esteem and when they don't: A meta-analysis. *Journal of Research in Personality*, 64, 34–51.

Zuckerman, M., Silberman, J., and Hall, J. A. (2013). The relation between intelligence and religiosity: A meta-analysis and some proposed explanations. *Personality and Social Psychology Review*, 17, 325–354.