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

Dealing with effect size multiplicity in systematic reviews and meta-analyses

José A López-López[1]
Matthew J Page[1,2]
Mark W Lipsey[3]
Julian PT Higgins[1]

1. Department of Population Health Sciences, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK
2. School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
3. Peabody Research Institute & Dept. of Human & Organizational Development, Vanderbilt University, Nashville, TN, USA

Correspondence to: Dr. José Antonio López-López, ja.lopez-lopez@bristol.ac.uk. , Tel. +44 (0) 117 928 7343

Running head: Dealing with effect size multiplicity

Abstract

Systematic reviews often encounter primary studies that report multiple effect sizes based on data from the same participants. These have the potential to introduce statistical dependency into the meta-analytic data set. In this paper we provide a tutorial on dealing with effect size multiplicity within studies in the context of meta-analyses of intervention and association studies, recommending a three-step approach. The first step is to define the research question and consider the extent to which it mainly reflects interest in mean effect sizes (which we term a 'convergent' approach) or an interest in exploring heterogeneity (which we term a 'divergent' approach). A second step is to identify the types of multiplicities that appear in the initial database of effect sizes relevant to the research question, and we propose a categorization scheme to differentiate them. The third step is to select a strategy for dealing with each type of multiplicity. The researcher can choose between a 'reductionist' meta-analytic approach, which is characterized by inclusion of a single effect size per study, or an 'integrative' approach characterized by inclusion of multiple effect sizes per study. We present an overview of available analysis strategies for dealing with effect size multiplicity within studies, and provide recommendations intended to help researchers decide which strategy might be preferable in particular situations. Last, we offer caveats and cautions about addressing the challenges multiplicity poses for systematic reviews and meta-analyses.

Key-words: systematic review, meta-analysis, effect size, multiplicity, dependency, effect size.

Dealing with effect size multiplicity in systematic reviews and meta-analyses

Systematic reviews often encounter primary studies that report multiple effect sizes.[1, 2] For example, a study may have measured more than one outcome on each participant (e.g. depression and anxiety), may have used multiple measures of the same outcome (e.g. different depression scales), may have taken measurements at multiple time points (e.g. at 1 month and 6 months), and may report results for different subgroups with some degree of overlap (e.g. men, women, young adults and elderly adults). Such effect size multiplicity in primary studies may introduce statistical dependency into the meta-analytic data set.[3] Specifically, if the same participants contribute to two or more effect sizes, then those effect sizes cannot be regarded as statistically independent and should not be treated as if they are in the meta-analysis.[4] Ignoring statistical dependencies can lead to misleading results, as has been noted by various authors,[5, 6] and demonstrated in simulation studies.[7, 8] It is therefore important to deal with multiplicity in some appropriate way, and the decisions made can have a substantial impact on the results of the meta-analysis.[9]

Here we focus on studies providing effect sizes that represent an association between two constructs. This includes intervention studies linking treatment conditions with subsequent outcomes, correlational studies linking two variables, and observational studies linking exposures with subsequent outcomes. We focus on studies that supply multiple effect sizes that are based (at least partially) on the same study participants. Various situations in which multiple effect sizes of this sort may be reported (or can be computed) are described in Table 1, based on examples in the literature.[4, 10] We also describe in the table how the multiplicity can lead to statistical dependencies across effect sizes from the same study. Multiplicity is only an issue for effect sizes that are relevant to the purpose of the review, and multiple effect sizes may be found either within a single report of a study or across multiple reports of the same study.

TABLE 1 HERE

There are several different ways to handle multiple statistically dependent effect sizes of these various sorts, and no single way is appropriate for all situations. In this tutorial we provide a roadmap to the strategies and options available and offer guidance for making appropriate choices. We propose that multiplicity be addressed in a sequence of three steps: (1) specify the research question and recognize its implications for multiplicity, (2) identify the nature and amount of multiplicity in the initial database of effect sizes relevant to the research question, and (3) choose an appropriate procedure for handling each kind of multiplicity identified. We provide recommendations intended to help researchers decide which strategy might be preferable in their particular situation. Finally, we summarize the messages of the paper through a suggested decision algorithm, and conclude with some caveats and cautions.

We make the important remark at the outset that approaches for dealing with multiplicity should be pre-specified in the review protocol wherever possible. We also stress that processes are often iterative in practice, because researchers may need to revise assumptions made about which effect sizes should be combined, and how, in the light of the sources of multiplicity found in the primary studies.

1. Defining the research question and setting eligibility criteria

Decisions made in the early stages of a systematic review impact on the nature and extent of multiplicity in the collected data. A first step in any systematic review is to specify the research question and develop clear definitions of the constructs and relationships involved. These definitions

inform the inclusion and exclusion criteria for primary studies and determine which effect sizes are relevant to the purpose of the review. The scope of the research question and, correspondingly, of the eligible primary studies, will generally have considerable influence on the nature and number of multiple statistically dependent effect sizes the researcher will encounter.

In relation to the broad synthesis plans, two ends of a continuum can be distinguished, which we will label as *convergent* and *divergent* review frameworks. Convergent reviews usually take a confirmatory approach. They focus on a rather narrowly circumscribed research question, typically an association previously hypothesized in the field that involves constructs and associated effect sizes expected to be relatively homogenous. A review of this sort is implemented through inclusion criteria that constrain studies to be relatively similar to each other with regard to the relationship examined and the way those relationships are represented empirically. The aim of convergent reviews is mainly to estimate an overall mean effect size, or a few mean effect sizes. Limited variability around those means is expected and, consequentially, there are few relevant moderator relationships to explore. Convergent frameworks are frequent in the health sciences, where inclusion criteria are typically more restrictive and the relevant constructs and measures are likely to be more tightly defined than in other disciplines.[11] A prototypical example is a review of clinical trials of a particular treatment for patients with a particular condition that examines only the outcomes most important to patients.[12]

In contrast, *divergent* frameworks tend to take an exploratory approach. They address a broad research question with a primary focus on identifying sources of variation among effect sizes that are expected to show substantial statistical heterogeneity such that interpretation of an overall mean effect size might be challenging, or even meaningless.[13-15] Their inclusion criteria are typically broad and may involve a range of diverse constructs and relationships. A divergent framework may help generate new hypotheses through a wider examination of moderators than can be achieved in primary studies. This is particularly useful for reviews dealing with complex constructs such as multifaceted interventions, where many factors may affect the magnitude of the association between the two constructs of interest. Divergent frameworks are not unusual in the social and behavioural sciences. Prototypical examples are reviews of the differential strength of a range of predictors of an outcome of interest,[16] or the relative effectiveness of different treatment modalities within a broad family under different conditions.[17]

A comparison of convergent and divergent review frameworks is provided in Table 2. In practice, most systematic reviews will fall somewhere between the two extremes. Effect size multiplicity within studies will be more common among reviews on the divergent side. Furthermore, reviews within the convergent framework will often be characterized by inclusion of a smaller number of studies and effect sizes per study.

TABLE 2 HERE

As an example, consider the planning of a systematic review and meta-analysis to examine the effectiveness of cognitive-behavioural therapies (CBT) in adults with anxiety disorder. A highly convergent approach to such a review might target the specific comparison of individually administered CBT versus no CBT, focussing on its effects on depression as measured using the Beck's Depression Inventory (BDI) and requiring use of an intention-to-treat (ITT) analysis. At the other extreme, a highly divergent approach might instead allow any comparison of a variant of CBT against something different and any analysis of any outcome. In later sections we develop this example with reference to a published systematic review on the topic.[18] We would classify this as predominantly a divergent review because the authors' scope was broad and embraced studies examining participants with different psychiatric disorders treated with a wide range of psychological interventions and comparators. Furthermore, effect sizes for multiple outcome domains were

included (both anxiety and depression), and different analysis approaches were entertained (both ITT analysis and analysis of completers only).

2. Examining the effect sizes in the database

The specification of the research question and the associated study inclusion criteria provide the basis for identifying eligible studies and eligible effect sizes. At this point, the choice of effect sizes to extract becomes important. It may not be appropriate to extract some types of effect sizes that could create multiplicity, either because they are not relevant to the research question or because it is known that they are too infrequent to support analysis even if extracted. For example, effect sizes for specific subgroups of the total sample, or for specific outcomes, might not be extracted for such reasons.

Once the relevant effect sizes have been extracted, it is time to think about the extent to which these can be combined in a meta-analysis, a notion we refer to as *combinability*. A first consideration is the *statistical* combinability of effect sizes. Effect sizes using very different metrics (e.g. a difference in means and a ratio of means) cannot be combined directly with each other in a meta-analysis. Sometimes conversions can be made between effect metrics, such as among standardized mean differences, odds ratios and correlation coefficients, although these conversions rest on assumptions that need to be considered. Sometimes effect metrics can be regarded as similar enough to be combined (e.g. odds ratios, rate ratios and risk ratios when the event is rare; or regression coefficients from analyses that adjust for different sets of confounders). Decisions about what effect sizes are in principle combinable may reduce the amount of multiplicity in any individual synthesis. Note, however, that some meta-analytic techniques, such as multivariate meta-analysis, allow for different metrics to be combined legitimately in a single analysis.

A second consideration is around which effect sizes are sensible to combine, taking into account substantive rather than statistical issues. The research questions for the review are fundamental to decisions made in this regard, and issues of combinability for a convergent approach to the review are likely to be different from those for a divergent approach. Within any set of studies deemed combinable according to a particular research question, the sources of multiplicity listed in Table 1 should be considered.

Sources of multiplicity in a systematic review of CBT interventions in adults with different anxiety disorders are summarized in Table 3.[18] A highly convergent approach to the review (or to a particular meta-analysis within the review) might target the specific comparison of individual CBT versus placebo using ITT analyses of BDI scores. In this case there is no effect size multiplicity, since none of the studies in this example provides more than one result relevant to this question. However, there are only two studies that meet these narrow criteria, illustrating the limitations of asking such a specific question. For the most highly divergent approach, all effect sizes in the data set are relevant, so there is substantial multiplicity with 115 effect sizes arising from the 26 studies. Within each study, the multiple effect sizes are derived wholly or in part from the same individual participants and thus all are statistically dependent.

Note that the extent of multiplicity in this example varies by different aspects of the research question. For intervention comparisons there is minimal multiplicity: only one study (study 24) contributes more than one comparison. For outcomes, there is similar multiplicity in all studies in that most of them report effect sizes for anxiety and for depression. The extent of multiplicity is largest for outcome measures, since several studies used multiple scales to measure depression or anxiety. For the analysis approach, multiplicity arises in five studies presenting both ITT and completers analyses.

TABLE 3 HERE

Tables of study characteristics are useful to visualize the extent and structure of the multiplicities in the review database. It is a generally good practice to construct a table detailing each included study to report in every systematic review.[19] Selecting a format that displays sources of multiplicity, such as Table 3 above, may help identify the sources that are common enough to pose a methodological challenge. Alternative formats, such as tables or matrices that explicitly quantify the extent of one or more types of multiplicity in terms of number of studies and effect sizes, are also possible.

3. Deciding on an analysis strategy

After the characteristics of the effect sizes in the database have been thoroughly examined, the third step is to proceed with the meta-analysis. Although pre-specified analysis plans should be followed wherever possible, the frequency and nature of multiplicity can be difficult to predict, so the plans will often need to be refined. At one extreme, very few effect sizes may meet the predefined criteria and a broader question might then become the focus of the synthesis. More challenging problems of multiplicity arise at the other extreme, when there are many eligible effect sizes from each study. At this point it may become apparent that the most insight will be gained by combining multiple effect sizes per study in a single synthesis (therefore introducing a need to account for statistical dependencies), by performing multiple syntheses when only one was planned, or by undertaking extra statistical investigations of heterogeneity in effect sizes within (and possibly across) studies.

We distinguish between two general approaches to dealing with effect size multiplicity within studies in a meta-analysis, which we call the *reductionist* approach and the *integrative* approach. The *reductionist* approach seeks to reduce the data set to minimal representations of the relationships of interest before performing the meta-analysis. This typically leads to one effect size from each independent study being included in each meta-analysis so that dependencies do not need to be modelled. However, there may be some loss of precision in the results if this appreciably reduces the total number of effect sizes included in the analysis.

In contrast, the *integrative* approach includes multiple effect sizes from each study in a single meta-analysis. This may be either because the effect sizes within the same study are considered to be equivalent but informative, or because the effect sizes differ in ways that can potentially be accounted for by moderators. This approach may lead to gains in statistical precision in the meta-analysis, but will introduce statistical dependencies that need to be modelled or accounted for in the statistical analysis to avoid misleading results. A comparison of the reductionist and integrative approaches is presented in Table 4.

TABLE 4 HERE

The reductionist and integrative approaches are not mutually exclusive. The effect sizes in a given meta-analytic database may include instances of more than one of the types of multiplicity distinguished in Table 1. Depending on the nature of the multiplicity and the purposes of the meta-analysis, some of those might be handled with reductionist strategies while others are handled with integrative strategies.

Returning to the running example of CBT for anxiety disorders, consider first the multiplicity posed by the intervention comparisons being made, in which only one study introduced an issue by comparing three interventions. Simple approaches might be taken for this source of multiplicity, such as omitting the extra effect size. For outcomes, the availability of data for anxiety and for depression from most studies opens up the possibility of using multivariate meta-analysis to analyse the two outcomes jointly (see Section 5). For outcome measures, there is much more substantial

multiplicity. The actual measurement instruments for anxiety suggest that some of them might be comparable (e.g. different versions of the same scale, different anxiety inventories), whereas others could be measuring symptoms pertaining to different anxiety disorders (e.g. social phobia, obsessive-compulsive disorder, social anxiety, panic disorder). For this area of multiplicity, it would be important to take a careful approach, for example either by selecting specific scales or by accounting for statistical dependencies. For the multiplicity posed by ITT and completer analysis approaches, it is likely to be most reasonable to select the preferred approach over the less preferred approach, or to undertake separate analyses, rather than to account for the statistical dependencies between the two analyses.

In fact, the authors of the review performed separate meta-analyses for anxiety and depression, and separate analyses for ITT and completer approaches. They lumped all intervention comparisons into a single category (CBT vs. control), dropping one relevant comparison from study 24 and exploring the association between the specific intervention and comparator types and the effect sizes through moderator analyses at a later stage (an alternative analysis approach here is network meta-analysis, see Section 5). The authors averaged the effect sizes over multiple measurement instruments. We would characterize this as reductionist meta-analysis.

Reductionist approaches address multiplicity by resolving it within studies, thereby removing it from the meta-analysis. Clearly this limits the possibility of exploring sources of heterogeneity within studies. In contrast, integrative approaches allow within-study effect size differences to be explored but introduce statistical issues with regard to the dependent effect sizes. In the following two sections we review the main strategies within the reductionist and integrative approaches, respectively.

4. Strategies to obtain non-dependent effect sizes

4.1 Random selection. Some multiple effect sizes from the same study sample may be regarded as completely equivalent conceptually for the purpose of the meta-analysis. For example, there may be different ways of measuring an outcome construct that the meta-analyst views as equally valid and relevant to the research question. One straightforward way to deal with multiplicity of this sort is to select one effect size randomly from each set of such effect sizes. Effect sizes that are truly equivalent should yield essentially the same results from the statistical analysis irrespective of which is selected from the respective sets of multiple effect sizes (after accounting for the metric and providing the study is large enough to have sufficient precision). However, equivalence of effect sizes from the same study is a strong assumption that may not be realistic, and the meta-analysis results may differ in ways that matter depending on which effect sizes are chosen.[9] If there are many instances of multiple effect sizes handled this way, sensitivity analyses comparing the results with different random draws may be appropriate to test the assumption of equivalence. Also, if there are many such instances, a drawback of this approach is loss of information associated with the omitted effect sizes and the associated loss of statistical precision and statistical power stemming from the smaller number of effect sizes contributing to the analysis.

4.2 Averaging effect sizes. Another simple way to reduce multiple effect sizes within a study to a single independent effect size is to average them. Especially for systematic reviews aimed mainly at estimating the overall mean effect size, averaging multiple effect sizes within studies before averaging them across studies has a straightforward logic. Possibilities include computation of the arithmetic mean or the median values, although more sophisticated procedures are available[20, 21] and the choice may influence the results.[22] One advantage of the arithmetic mean is that its standard error can readily be derived using standard formulae, as discussed in Chapter 24 of

Borenstein and colleagues.[23] However, these formulae require the correlations between the variables whose effect sizes are being averaged (e.g. different outcome variables in an intervention study). Such information is often not reported in the primary studies[24] and typically cannot be computed directly without individual participant data. In practice the correlations would often need to be imputed.

Averaging effect sizes is a simple and arguably objective strategy that makes use of all the relevant effect sizes. Its limitation is that it ignores any within-study variability across the effect sizes combined in those averages. Also, as we have noted, the correct standard error of a within-study averaged effect will depend on the correlation structure of the summarized variables, which is rarely available. One alternative is to average the variances of the effect estimates and use that value as the variance of the averaged effect size. This approach does not convey the increased precision that comes with having more effect size estimates, although that might be considered a positive feature because it does not award greater influence to studies that provide multiple effect sizes relative to those that provide only one.

4.3 Selection using a decision rule. A third option to obtain non-dependent effect sizes is to select a single most relevant effect size from each source of within-study multiplicity according to a decision rule. This strategy attempts to maximize the comparability of the effect sizes across studies while still restricting the meta-analysis to one effect size per study sample. Decision rules may be especially appropriate when multiple effect sizes within studies are regarded as being loosely equivalent but not completely interchangeable, and when the effect size diversity within studies is not very similar to that between studies. With effect sizes involving measures of depression, for example, most studies might use the BDI, but some might also include a second depression measure. The selection rule in that case might be to choose the BDI in preference to other depression scales if it is available. Similarly, a decision rule might specify effect sizes based on ITT analyses in preference to completer analyses, or to prioritize adjusted regression coefficients in preference to unadjusted ones, or to choose the longest period of follow up.

When such decision rules are to be used, both Cochrane and the Campbell Collaboration consider it desirable that specification of the rule be done before data extraction. However, a recent survey of systematic reviews suggested that this recommendation is rarely followed in practice.[1] Pre-specification can be challenging when insufficient information is available in advance to predict the nature of the multiplicity, and therefore what kind of decisions will be needed. As stated in the *Cochrane Handbook for Systematic Reviews of Interventions*,[25] the rationale for specifying a decision rule before data extraction is to minimize bias, and other organizations also recommend that such decision rules be used.[26-29] If not prespecified, articulating decision rules as a *post hoc* strategy may still be appropriate if they are developed in a way that does not systematically select effect sizes based on their magnitude or direction.[26]

Table 5 provides some suggestions for decision rules that select a single effect size per study for each source of multiplicity. We recommend that similar criteria be stated at the protocol stage if a decision rule is deemed appropriate for the review. An additional task for the review team will be to determine the order in which the different rules across domains should be applied, and the resulting decision rule might then look like a hierarchical algorithm.

TABLE 5

An example of a particularly detailed algorithm is provided by a review on the unintended effects of three childhood vaccines.[30] The reviewers extracted 854 different effect sizes relating to 64 unique studies of association between each of the vaccines and all-cause mortality. Box 1 provides the

algorithm that was developed to reduce these effect sizes to one per study in order to overcome issues of multiplicity.

BOX 1 HERE

5. Modelling or accounting for dependency in the meta-analysis

In this section, we describe several meta-analytic strategies available to researchers when the meta-analytic data set includes dependency structures among effect sizes from the same participant samples. One option, not explicitly covered, is to ignore those dependencies and analyse the effect sizes as if they were all statistically independent. This approach will produce erroneous standard errors that will distort all the inferential statistics from the meta-analysis. While the consequences of this may be trivial if it involves only relatively few effect sizes in the data set,[5] we generally discourage this naïve approach.

5.1 Multivariate meta-analysis. A multivariate meta-analysis[31, 32] provides meta-analytic estimates for all effect sizes of interest, including those that are statistically dependent, within a single modelling framework, even if different effect size metrics are present in the data set. This is achieved by integrating a covariance matrix into the analysis that carries information about the extent to which any two effect sizes are correlated. Using multivariate techniques to integrate effect sizes on different outcomes measured on the same participants is particularly attractive, as these models provide separate estimates for each outcome that are more precise than those from separate meta-analyses of each outcome. For example, a multivariate meta-analysis may be used to produce summary effect sizes for mathematics test scores separately from language test scores while accounting for the correlation between these scores when measured on the same children. Moreover, fitting a single model to obtain overall effect size estimates across outcomes avoids the inflation of the Type I error rate that might arise if separate meta-analyses were conducted for each of these outcomes.[4, 10, 33, 34]

Multivariate meta-analysis has been recommended and shown to improve the performance of univariate meta-analysis in the presence of outcome reporting bias.[35] This is achieved by 'borrowing strength' across correlated outcomes so that the magnitude and direction of the missing effect sizes can be estimated even if some outcomes are only partially reported across studies. Another common application of multivariate meta-analysis is to examine multiple intervention comparisons against a common control group, an approach known as network meta-analysis.[36] A novel implementation of network meta-analysis allows mapping of effects from one outcome measure to another, providing an attractive alternative to the commonly used standardized differences in means as a dimensionless measure of effect.[37]

Although multivariate meta-analysis is an elegant and powerful approach to handling multiple dependent effect sizes, its implementation requires information about the correlations between the effect sizes estimated from the same study samples (in the form of covariance matrices). This correlational information is rarely reported in source studies, which substantially limits the applicability of these methods in practice. Some strategies have been proposed to address this problem, both within a frequentist[38] and a Bayesian[39] framework. Software tools to implement multivariate meta-analysis include the *mvmeta* function in Stata[40] and the *metafor* package in R.[41]

5.2 Multilevel models. Another meta-analytic strategy to handling multiplicity is to use multilevel models, also known as hierarchical or mixed models.[42, 43] A simple weighted average approach to meta-analysis can be conceived as a two-level hierarchical model as illustrated in Figure 1A. Here it is assumed that there is a single true effect size of interest from each study, and each study provides a

single estimate of that effect size based on data derived from its participant sample. The variability at the higher level, representing variation in the true effect sizes across studies, is commonly either modelled by assuming a distribution (as in random-effects meta-analysis) or assumed to be zero (as in a fixed-effect meta-analysis). The variability at the lower level, which reflects variation across participants within studies, is represented by within-study variances that are typically assumed to be measured without error. Effect size multiplicity within studies can be accounted for by introducing an intermediate level to the hierarchy, as displayed in Figure 1B for multiple measures of the same outcome. Additional complexity can be built in as needed to reflect the structure of the data, leading to models with more than three levels. An example is shown in Figure 1C, where a four-level model represents the combination of multiple time points and multiple measurement instruments within the studies.

FIGURE 1 HERE

The three-level model has been found to perform similarly to multivariate meta-analysis in terms of accuracy of parameter and standard error estimates,[7] and has the advantage of being applicable in situations in which estimates of the covariances among effect sizes are not available, which are very common in practice. Multilevel models can be implemented in R, using the *metaSEM*[44] or the *metafor*[41] packages, and in SAS (Proc Mixed; some code is provided in Van den Noortgate and colleagues[8]).

When synthesizing correlation matrices, an increasingly popular application of multilevel models is meta-analytic structural equation modelling (MASEM), bearing in mind the close link between multilevel modelling and structural equation modelling (SEM).[45] In addition to the statistical integration of the correlation matrices reported in the primary studies, a second stage in MASEM typically involves testing models within a SEM framework.[2, 46]. While Mplus and LISREL are software choices familiar to SEM analysts, a more generic software option tailored to meta-analysis is the *metaSEM* package in R[44], for which a tutorial has recently been published.[47]

5.3 Robust variance estimation. The primary problem created by multiple statistically dependent effect sizes is distortion of the sampling variance estimates on which the inferential statistics for the meta-analysis are based. One approach to this problem is the use of robust variance estimation techniques that, as the name indicates, are robust to distortions. This strategy, developed by Hedges and colleagues[48] and building on the work of Huber[49] and White,[50] has the advantage of simultaneously addressing all the different types of multiplicity that might appear in a meta-analytic data set with a single blanket procedure. It requires an estimate of the correlations across the effect sizes involved in instances of multiplicity, but the results are not very sensitive to a range of plausible estimates.

As with most statistical methods used in meta-analysis, robust variance estimation relies on large-sample approximations, and therefore estimates and statistical tests may yield inaccurate results when the number of studies and effect sizes is relatively small. To overcome this limitation, Tipton[51, 52] explored the performance of several small-sample corrections of the standard errors of the parameters in the meta-analytic model and of the degrees of freedom of the *t*-test on which statistical significance and confidence intervals were based. In her simulation study exploring the performance of these variants, some combinations of both adjustments were found to improve the performance of the statistical tests under a wide range of realistic scenarios. Robust variance estimation (with or without these small-sample corrections) can be implemented in R (using the *robumeta* or *clubSandwich* packages) or using macros developed for SPSS and Stata (for a tutorial, see Tanner-Smith and Tipton[53]).

5.4 Specific methods for multiplicity in exposures or interventions. Some specific techniques are available to address multiplicity of interventions and exposures. A dose-response analysis allows multiple levels of exposures to be modelled explicitly. Care is required to model statistical dependencies if comparisons are used against a common reference category. An advantage of this approach is that dose-response relationships can be modelled both within and across studies.[54] For multiple interventions, an important candidate is network meta-analysis, based on a connected network of within-study comparisons (the lines in a network diagram) from among a defined set of interventions (the nodes in a network diagram).

6. A suggested approach to addressing multiplicity

We provide a summary of the guidance offered in this paper in Figures 2 and 3. Figure 2 describes the three steps of a systematic review with which multiplicity can be addressed. The first step is to articulate the review question, specifying definitions of the constructs involved in the relationship(s) of interest, and clarifying the extent to which the review takes a convergent approach (in which narrow eligibility criteria are adopted to answer a highly focused question) or a divergent approach (in which wider eligibility criteria are used to bring diversity of effect sizes into the review).

The second step is to examine the collected effect sizes for their combinability. This step includes both statistical and non-statistical considerations. Effect sizes measured using different metrics should not be combined statistically (unless multivariate meta-analysis is used), whereas effect sizes using the same metric may be considered as statistically combinable. Sensible decisions about which effect sizes should be combined involve standard considerations of the similarity of studies, and avoiding meta-analysis is a legitimate option if the effect sizes are not regarded as similar enough to be synthesized quantitatively. For effect sizes deemed reasonable to combine in a meta-analysis, sources of multiplicity need to be identified, according to the sources listed in Table 1. Tables are helpful to visualize the extent and structure of the multiplicities found in the included studies with regard to each particular proposed meta-analysis.

The third step consists of choosing an analysis strategy to handle any multiplicity. Effect size multiplicity within studies may arise from multiple answers to the same research question or may be due to different questions being asked, such as effects on different comparisons or different outcome measures. A key decision is whether to perform separate meta-analyses to address the multiple research questions or whether to 'lump' the effect sizes into a single synthesis of a broader research question. Performing multiple separate meta-analyses might lead to inflated Type I error rates. Adjustments for multiple testing are available to address this problem, although they often yield over-conservative results and are not common in systematic reviews.[25] Within a 'lumped' synthesis, meta-regression approaches enable examination of one or more moderators, controlling Type I error.[55] They can be used to investigate between-study moderators (e.g. design characteristics) or within-study moderators (e.g. different outcome measures), or both.

FIGURE 2 HERE

Once a decision had been made to perform a meta-analysis (either a lumped analysis or one from a series of separate analyses), we propose the algorithm in Figure 3 to tackle each type of multiplicity. In the running CBT example (Table 3), there were four sources of multiplicity: intervention comparisons, outcomes (depression and anxiety), outcome measures (different scales) and analysis approaches (ITT vs. completers), and different strategies were used to address each of these.

A starting point is whether the effect sizes within studies are completely interchangeable (or equivalent) in the sense that each would answer the research question with similar relevance and

they are not expected to differ in relation to any moderators of interest. If this is the case, then averaging multiple effect sizes within studies is an especially straightforward way to handle multiplicity. However, if the extent of multiplicity is large, this strategy will discard potentially relevant information, resulting in loss of precision and statistical power in the model estimation, so that alternative strategies including all effect sizes and accounting for the statistical dependencies would be preferred. Among these, multilevel models and robust variance estimation are applicable in most situations, although the latter strategy (using small-sample corrections) will provide the best performance in terms of Type I error rate for the statistical tests with a small number of studies.[56]

If the effect sizes within each study are not considered to be equivalent, then we might ask whether they differ in ways that are of interest (i.e. have different values in relation to moderator variables) or whether they differ more as nuisance variation. Moderator analyses may be pursued in the former case, but in the latter case the synthesis will need to address the diversity in other ways. This is the situation in which a decision rule may be particularly useful, so that the most appropriate effect size is selected for a meta-analysis. If a hierarchy cannot be specified, or there is a desire to include as many effect sizes as possible in the meta-analysis, then we suggest several options in Figure 3.

FIGURE 3 HERE

7. Conclusion

Effect size multiplicity is a common challenge in systematic reviews because primary studies frequently report several effect sizes with at least partial overlap between samples of participants. We have offered recommendations to delimit, identify, and address such multiplicity in systematic reviews and meta-analyses. We provide an overview of different analysis strategies to deal with such multiplicity along with strengths and limitations of each alternative.

In general, approaches for dealing with multiplicity should be specified in advance in a review protocol, including strategies for dealing with populations of interest, key constructs of interest (such as exposures, interventions, outcomes and measures of outcomes), follow-up times, and analysis approaches. In practice, however, it is difficult to anticipate all issues of multiplicity that may arise, and transparent reporting of decisions made is critical.

Systematic reviews organized around convergent research questions will generally encounter few instances of effect size multiplicity. Furthermore, the types of multiplicity that may occur (Table 1) will often be predictable enough for some of them to be resolved in advance through the inclusion criteria. By contrast, divergent research questions will almost inevitably lead to systematic reviews that present many instances of multiplicity, often of quite diverse kinds. Nonetheless, many of the multiple dependent effect sizes may represent variation that is relevant to the purposes of the meta-analysis given its focus on sources of variation in the observed effects.[57] And, even if that were not the case, the number of multiple dependent effect sizes relevant to the research question will often be too large to be ignored and sufficiently large so that different choices about how to handle them could impact on the results of the meta-analysis. Careful planning for dealing with multiplicity and consideration of the implications of different choices are thus particularly important for meta-analyses of broad scope that address divergent research questions.

An examination of the database of effect sizes, as we suggest in Section 2, can lead to some immediate reductions in the extent of multiplicity. Some decisions made at this step may lead to data sets for separate meta-analyses. However, substantial multiplicity might still remain after this step and any chosen strategy entails caveats that need to be considered. Researchers might decide

to reduce the data set to one effect size per study sample by implementing one or more of the strategies described in Section 4, which results in loss of information and statistical power. Alternatively, analysis strategies in Section 5 allow combination of all effect sizes while accounting for dependency structures, but this might raise issues of disproportionate influence on the meta-analysis results from studies that contribute a large number of effect sizes relative to those that contribute fewer.

We acknowledge that practical systematic reviews and meta-analyses may entail more complex scenarios than those we have addressed, for example because available information (e.g. within-study correlation, number of studies) limits implementation of some strategies. A combination of different strategies, or even approaches, might sometimes be the best solution. We have focussed on reviews of association between two constructs, although most of the principles we outline apply to other types of studies. In spite of the challenges that it poses, effect size multiplicity provides exciting opportunities to explore different hypotheses in systematic reviews.

8. Acknowledgements

We are grateful to Trevor Day, from the Royal Literary Fund, who provided useful feedback on an early draft of the manuscript.

References

1. Page MJ, McKenzie JE, Chau M, Green SE, Forbes A. Methods to select results to include in meta-analyses deserve more consideration in systematic reviews. J Clin Epidemiol 2015;68(11):1282-91.
2. Wilson SJ, Polanin JR, Lipsey MW. Fitting meta-analytic structural equation models with complex datasets. Res Synth Methods 2016;7(2):121-39.
3. Stevens JR, Taylor AM. Hierarchical dependence in meta-analysis. J Educ Behav Stat 2009;34(1):46-73.
4. Littell JH, Corcoran J, Pillai V. Systematic reviews and meta-analysis. New York: Oxford University Press; 2008.
5. Hedges LV. Statistical considerations. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 37-47.
6. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21(4):589-624.
7. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. Behav Res Methods 2013;45(2):576-94.
8. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Meta-analysis of multiple outcomes: a multilevel approach. Behav Res Methods 2015;47(4):1274-94.
9. Tendal B, Nüesch E, Higgins JPT, Jüni P, Gøtzsche PC. Multiplicity of data in trial reports and the reliability of meta-analyses: empirical study. BMJ 2011;343:d4829.
10. Bender R, Bunce C, Clarke M, et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008;61(9):857-65.
11. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Med Res Methodol 2011;11(1):160.
12. Sadeghirad B, Siemieniuk RAC, Brignardello-Petersen R, et al. Corticosteroids for treatment of sore throat: systematic review and meta-analysis of randomised trials. BMJ 2017;358:j3887.
13. Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI. Understanding heterogeneity in meta-analysis: The role of meta-regression. Int J Clin Pract 2009;63(10):1426-34.

14. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009;172(1):137-59.
15. Viechtbauer W. Analysis of moderator effects in meta-analysis. In: Osborne J, editor. Best practices in quantitative methods. Thousand Oaks, CA: Sage; 2008. p. 471-87.
16. Wilson SJ, Lipsey MW, Derzon JH. The effects of school-based intervention programs on aggressive and disruptive behavior: A meta-analysis. J Consult Clin Psychol 2003;71(1):136-49.
17. Sánchez-Meca J, Rosa-Alcázar AI, Marín-Martínez F, Gómez-Conesa A. Psychological treatment of panic disorder with or without agoraphobia: a meta-analysis. Clin Psychol Rev 2010;30(1):37-50.
18. Hofmann SG, Smits JA. Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. J Clin Psychiatry 2008;69(4):621-32.
19. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. BMJ 2009;339:b2535.
20. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.
21. Rosenthal R, Rubin DB. Meta-analytic procedures for combining studies with multiple effect sizes. Psychol Bull 1986;99(3):400-6.
22. Marín-Martínez F, Sánchez-Meca J. Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. Span J Psychol 1999;2(1):32-8.
23. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester, UK: Wiley; 2009.
24. Hedges LV. Meta-analysis. In: Rao CR, Sinharay S, editors. Handbook of Statistics. 26. Amsterdam: Elsevier; 2007. p. 919-53.
25. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Chichester, UK: Wiley-Blackwell; 2011.
26. Agency for Healthcare Research and Quality. Methods guide for effectiveness and comparative effectiveness reviews. Agency for Healthcare Research and Quality; 2008.
27. Boers M, Kirwan JR, Wells G, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67(7):745-53.
28. Christensen R, Maxwell LJ, Jüni P, et al. Consensus on the need for a hierarchical list of patient-reported pain outcomes for metaanalyses of knee osteoarthritis trials: An OMERACT Objective. J Rheumatol 2015;42(10):1971-5.
29. Prinsen CA, Vohra S, Rose MR, et al. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. Trials 2014;15(1):247.
30. Higgins JPT, Soares-Weiser K, López-López JA, et al. Association of BCG, DTP, and measles containing vaccines with childhood mortality: systematic review. BMJ 2016;355:i5170.
31. Becker BJ. Multivariate meta-analysis. In: Tinsley HEA, Brown SD, editors. Handbook of applied multivariate statistics and mathematical modeling. San Diego, CA: Academic Press; 2000. p. 499-525.
32. Riley RD, Price MJ, Jackson D, et al. Multivariate meta-analysis using individual participant data. Res Synth Methods 2015;6(2):157-74.
33. Gleser LJ, Olkin I. Stochastically dependent effect sizes. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 357-76.
34. Imberger G, Vejlby AD, Hansen SB, Møller AM, Wetterslev J. Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews. PLoS One 2011;6(12):e28422.
35. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. Stat Med 2012;31(20):2179-95.
36. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods 2012;3(2):80-97.

37. Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. Res Synth Methods 2015;6(1):96-107.
38. Riley R, Thompson J, Abrams K. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics 2008;9:172-86.
39. Wei Y, Higgins JPT. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. Stat Med 2013;32(7):1191-205.
40. White IR. Multivariate random-effects meta-regression: updates to mvmeta. Stata J 2011;11(2):255.
41. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw 2010;36(3):1-48.
42. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. Res Synth Methods 2011;2(1):61-76.
43. Raudenbush SW, Bryk AS. Hierarchical linear models: Applications and data analysis methods. 2nd ed. Thousand Oaks, CA: Sage; 2002.
44. Cheung MWL. metaSEM: an R package for meta-analysis using structural equation modeling. Front Psychol 2015;5:1521.
45. Bauer DJ. Estimating multilevel linear models as structural equation models. J Educ Behav Stat 2003;28(2):135-67.
46. Cheung MWL, Chan W. Meta-analytic structural equation modeling: a two-stage approach. Psychol Methods 2005;10(1):40-64.
47. Cheung MWL. Fixed-and random-effects meta-analytic structural equation modeling: Examples and analyses in R. Behav Res Methods 2014;46(1):29-40.
48. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods 2010;1(1):39-65.
49. Huber P. The behavior of maximum-likelihood estimates under nonstandard conditions. In: LeCam LM, Neyman J, editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; 1967. p. 221-33.
50. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society 1980;48(4):817-38.
51. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods 2015;20(3):375-93.
52. Tipton E, Pustejovsky JE. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. J Educ Behav Stat 2015;40(6):604-34.
53. Tanner-Smith EE, Tipton E. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. Res Synth Methods 2014;5(1):13-30.
54. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol 1992;135(11):1301-9.
55. Polanin JR, Pigott TD. The use of meta-analytic statistical significance testing. Res Synth Methods 2015;6(1):63-73.
56. López-López JA, Van den Noortgate W, Tanner-Smith EE, Wilson SJ, Lipsey MW. Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. Res Synth Methods 2017;8(4):435-50.
57. Lipsey MW. Identifying interesting variables and analysis opportunities. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 147-58.
58. Mayo-Wilson E, Li T, Fusco N, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. J Clin Epidemiol 2017.

Tables

Table 1. Sources of effect size multiplicity in primary studies and how they might lead to statistical dependencies

Source	Examples	Potential for dependency
Populations/ groups	(1) Investigators report effect sizes based on different overlapping subsets of participants, such as men only and older adults only. (2) Investigators report effect sizes for the whole sample and also for subpopulations.	If separate effect sizes are reported based on (a) men only, (b) women only, (c) young adults, and (d) older adults, then all four effects should not be included in a meta-analysis assuming independent effect sizes because (a) and (b) include both younger and older adults, and (c) and (d) include both men and women.
Constructs to be compared		
1. Interventions or exposure groups	(1) An intervention study includes more than two arms, such as drug A, drug B, placebo, and thus allows for effect sizes from multiple comparisons (e.g. each drug versus drug placebo).	Researcher is interested in the effect of either drug, compared with placebo, but inclusion of drug A versus placebo and drug B versus placebo comparisons in the meta-analysis will double-count the placebo group.
2. Outcomes or dependent variables	More than one dependent variable is measured, such as panic, anxiety and depression, allowing examination of different outcomes.	Researcher is interested in the effect of an intervention on mental health symptoms generally, so all three domains are relevant to the meta-analysis, but may be measured on the same participants.
3. Variables correlated with each other	A correlational study looking at factors associated with educational achievement tested different variables (e.g. scores on several tests focused on different cognitive skills).	Researcher is interested in a comprehensive set of factors associated with educational achievement, and hence all explored associations are relevant to the meta-analysis, but may be measured using the same participants.
Ways of measuring the same construct	A construct is measured in more than one way, such as with different self-reported measures, or a combination of self-reported measures with external assessment measures (e.g. clinical interviews) and objective measures (e.g. biological samples). A special case here is part-whole relations among measures (e.g. the total score on a scale as well as the scores on the subscales that are included in that total score are reported).	All measures may be relevant to a particular meta-analysis, but they were measured on the same participants.

Time points	An outcome domain or association is measured at different follow-up times.	All time points may be of interest to a particular meta-analysis, but they include the same participants.
Analyses	(1) Investigators estimate the effect of an exposure factor using different effect size indices (e.g. risk ratio and risk difference), or report analyses both unadjusted and adjusted for confounders (2) Investigators estimate the magnitude of an association using total and partial correlations, or product-moment and rank correlation coefficients. (3) Investigators report results based on the intention-to-treat principle and also results including study completers only. (4) Different investigators report secondary analyses of data drawn from the same common database.	All analyses may be relevant to a particular meta-analysis, but there is partial or complete overlap of participants included in the different analyses.

Table 2. A comparison of convergent and divergent review frameworks

Convergent	Divergent
Characterized by • Confirmatory approach • Single primary research question • Aims to include studies and effect sizes addressing the same (or a very similar) question • Focus on overall effect *Implementation* • Potential exclusion of informative data • Useful if the main goal is to quantify the association between the two constructs • Dominant in health sciences	*Characterized by* • Exploratory approach • May involve multiple research questions • Aims to include diverse studies and effect sizes in the research area • Focus on variation *Implementation* • Often leads to large databases • Useful to generate new hypotheses • More common among social and behavioural sciences

Table 3. Table of study characteristics illustrating instances of multiplicity in a review of cognitive-behavioural therapies for anxiety disorders (modified from Hofman and colleagues[18])

Study	Intervention comparisons eligible for the systematic review	Outcomes reported (number of scales)	Analysis approach	Total number of effect sizes
1	CBT vs. Supportive counseling	Anxiety (1) Depression (1*)	Completer	2
2	CBT vs. Supportive counseling	Anxiety (1) Depression (1*)	Completer	2
3	CBT vs. Supportive counseling	Anxiety (1) Depression (1*)	Completer	2
4	CBT vs. Supportive counseling	Anxiety (1) Depression (1*)	Completer ITT	4
5	CBT vs. Nondirective therapy	Anxiety (5) Depression (2*)	Completer	7
6	CBT vs. Discussion group	Anxiety (4) Depression (2*)	Completer	6
7	ERP vs. Placebo pill	Anxiety (2)	Completer	2
8	BT vs. Systematic relaxation	Anxiety (2) Depression (1)	Completer	3
9	ERP vs. Anxiety management	Anxiety (3) Depression (1*)	Completer	4
10	CT vs. Placebo pill	Anxiety (3) Depression (1)	Completer ITT	8
11	CBT vs. Placebo pill	Anxiety (1)	Completer ITT	2
12	CT vs. Placebo pill	Anxiety (3) Depression (1)	Completer ITT	8
13	CBT vs. Nondirective supportive therapy	Anxiety (4)	Completer	4
14	CBT vs. Placebo pill	Depression (1)	ITT	1
15	CBT vs. Supportive counseling	Anxiety (5) Depression (1*)	Completer	6
16	CBT vs. Supportive counseling	Anxiety (3) Depression (1*)	Completer	4
17	PE vs. Supportive counseling	Anxiety (1) Depression (1*)	Completer	2
18	CBT vs. Relaxation	Anxiety (2) Depression (1*)	Completer	3
19	CBT vs. Problem solving therapy	Anxiety (1) Depression (1*)	Completer ITT	4
20	NET vs. Supportive counseling	Anxiety (2)	Completer	2
21	CT vs. Placebo pill	Anxiety (7) Depression (1*)	Completer	8
22	CBT vs. Supportive therapy	Anxiety (3) Depression (1*)	Completer	4
23	CCBT vs. Placebo pill	Anxiety (3)	Unclear	3
24	CBGT vs. Educational group supportive therapy CBGT vs. Placebo pill	Anxiety (8) Depression (1)	Completer	18

25	CBGT vs. Educational group supportive therapy	Anxiety (4) Depression (1*)	Completer	5
26	BT vs. Psychological Placebo	Anxiety (1)	Completer	1

BT = Behaviour therapy; CBGT = Cognitive-behavioural group therapy; CBT = Cognitive-behavioural therapy; CCBT = Comprehensive cognitive-behavioural therapy; CT = Cognitive Therapy; ERP = Exposure and response prevention; ITT = Intention-to-treat analysis; NET = Narrative exposure therapy; PE = Prolonged exposure therapy. * = depression scales include the Beck Depression Inventory.

Table 4. Meta-analytic approaches to multiplicity

Reductionist	Integrative
Characterized by • Inclusion of one effect size per study in a meta-analysis *Implementation* • Estimation accuracy and statistical power may be diminished • Effect sizes included in each meta-analysis are statistically independent	*Characterized by* • Inclusion of multiple effect sizes from all or some studies in a meta-analysis *Implementation* • May lead to large meta-analytic data sets • Gain in statistical power • Likely to introduce statistical dependencies

Table 5. Considerations for selecting a single effect size per study

Source of Multiplicity	**Suggestion**
Populations/ groups	Select the effect size calculated based on the broadest population or the largest sample size. Note that it is often possible to use the data reported for subgroups to estimate the effect size for an aggregate that reassembles the full sample (e.g. combining the male and female subgroup data appropriately to get the effect size for the full sample even though the full sample results may not have been reported). Note that separate meta-analyses can be conducted for subgroups if those are reported widely enough across studies, (e.g. male and female subsamples).
Constructs to be compared	
1. Interventions or exposure groups	Select the intervention(s) or exposure(s) most relevant (e.g. interventions recommended for practice, most used, most fully-developed, or most implementable; or contrast using the largest sample sizes). Note that it is often possible to collapse data across intervention or exposure groups, if the interventions are sufficiently similar, to estimate a single effect size for a study.
2. Outcomes or dependent variables	Select the outcome most relevant (e.g. most familiar to readers, most frequently measured in practice). A practical strategy is to select the outcome or dependent variable for which the most studies contribute a result; the impact of reporting biases may need to be considered if such a strategy is used. It may be preferable to perform separate meta-analyses for each outcome or dependent variable if possible.
3. Variables being correlated with each other	Select the variables most relevant on a substantive basis (e.g. based on theoretical reasons, not numerical results).
Ways of measuring the same construct	Select operationalization with evidence of highest validity and reliability (e.g. objective measure); if different measures are believed to be equally valid, consider the most commonly used.
Time points	Select the time point most relevant or perhaps those most frequently reported; the decision will depend on the topic and nature of the review, noting that effect sizes often decrease over time.
Analyses	Select the effect size computed using the methodological approach considered to be most appropriate in the context of the review. For instance, select intention-to-treat-analyses if the intention is to estimate the effect of being assigned to an intervention rather than starting and adhering to the intervention. In the same vein, select partial correlation (or adjusted regression) coefficient if the aim is to estimate the strength of the association between two variables once a third variable is controlled for.
Data sources (e.g. different, journal articles, conference abstracts, unpublished sources)	Select the effect size from the most reliable or most recent source. For example, a Clinical Study Report from a randomized trial is typically less prone to reporting biases than a journal article about the same trial, and a journal article is usually more reliable than a conference abstract.[58]

Box

Box 1. Example of an algorithm to select one result from each study in a review of the effects of Bacillus Calmette–Guérin (BCG) vaccine, diphtheria-tetanus-pertussis (DTP) vaccine and measles-containing vaccine (MCV), to avoid multiplicity in the meta-analysis[30]

1. Select comparison with vaccination sequence according to the World Health Organization (WHO) recommendations (e.g., BCG, then DTP, then MCV). We depict 'DTP after BCG' as 'BCP<DTP'.
2. Select estimates from randomized comparisons in preference to observational comparisons.
3. Select estimates adjusted for age and other vaccines in preference to those not adjusted for these.
4. Estimates of primary interest
   - BCG
     A. BCG at birth vs. no BCG in preference to
     B. BCG vs. no BCG
   - DTP
     A. BCG<DTP (any number of doses) vs. BCG in preference to
     B. BCG<DTP (1 or 1-2 doses) vs. BCG in preference to
     C. BCG<DTP (2 or more doses) vs. BCG in preference to
     D. DTP (any number of doses) vs. no DTP in preference to
     E. DTP (1 or 1-2 doses) vs. no DTP in preference to
     F. DTP (2 or more doses) vs. no DTP
   - MCV
     A. BCG<DTP<MCV vs. BCG<DTP in preference to
     B. BCG<MCV vs. BCG in preference to
     C. DTP<MCV vs. DTP in preference to
     D. MCV vs. no MCV
5. Select comparison with least co-administration of other vaccines, particularly when vs. unvaccinated children.
6. Select comparison involving children from the same area in preference to a different area.
7. Select estimate obtained using landmark rather than retrospective approach to analysis.
8. Select estimate obtained from general population children rather than subgroups (e.g. hospitalized children).
9. Select comparison including the most comprehensive adjustment for potential confounders.
10. Select result for the shortest period of follow-up.
11. Select result with the largest sample size.
12. Select comparison with vaccination strategies according to the WHO recommendations (e.g. BCG at birth, MCV vaccine at 9 months) in preference to alternative strategies.
13. Select estimate using the methodological approach claimed to be superior or more correct
14. Select result from more recent article.

Figures

Figure 1. Multi-level (or hierarchical) structures of meta-analytic data sets. The data at the lowest level (y) are effect size estimates.

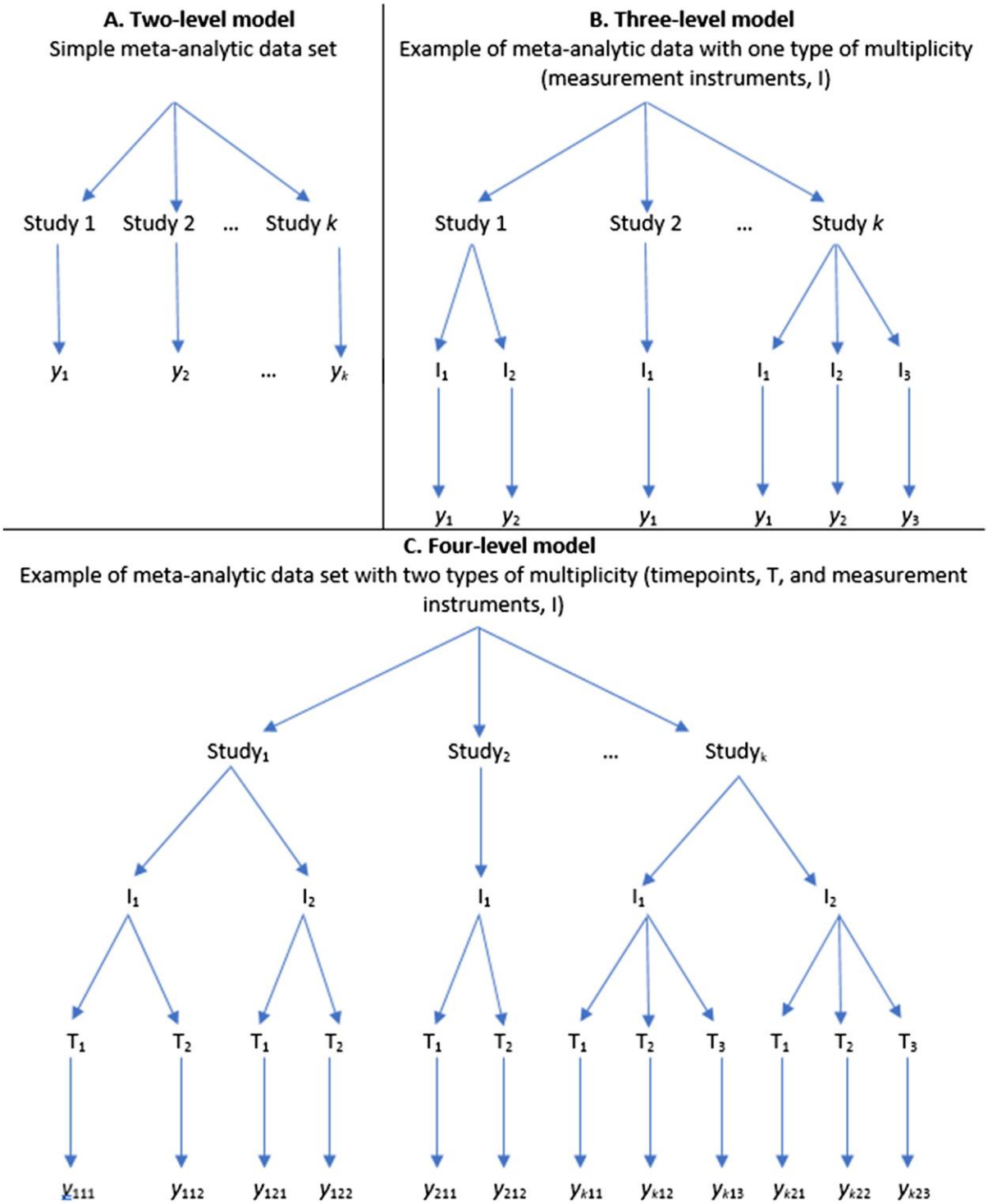

Figure 2. Summary of steps of a systematic review in which multiplicity might need to be addressed.

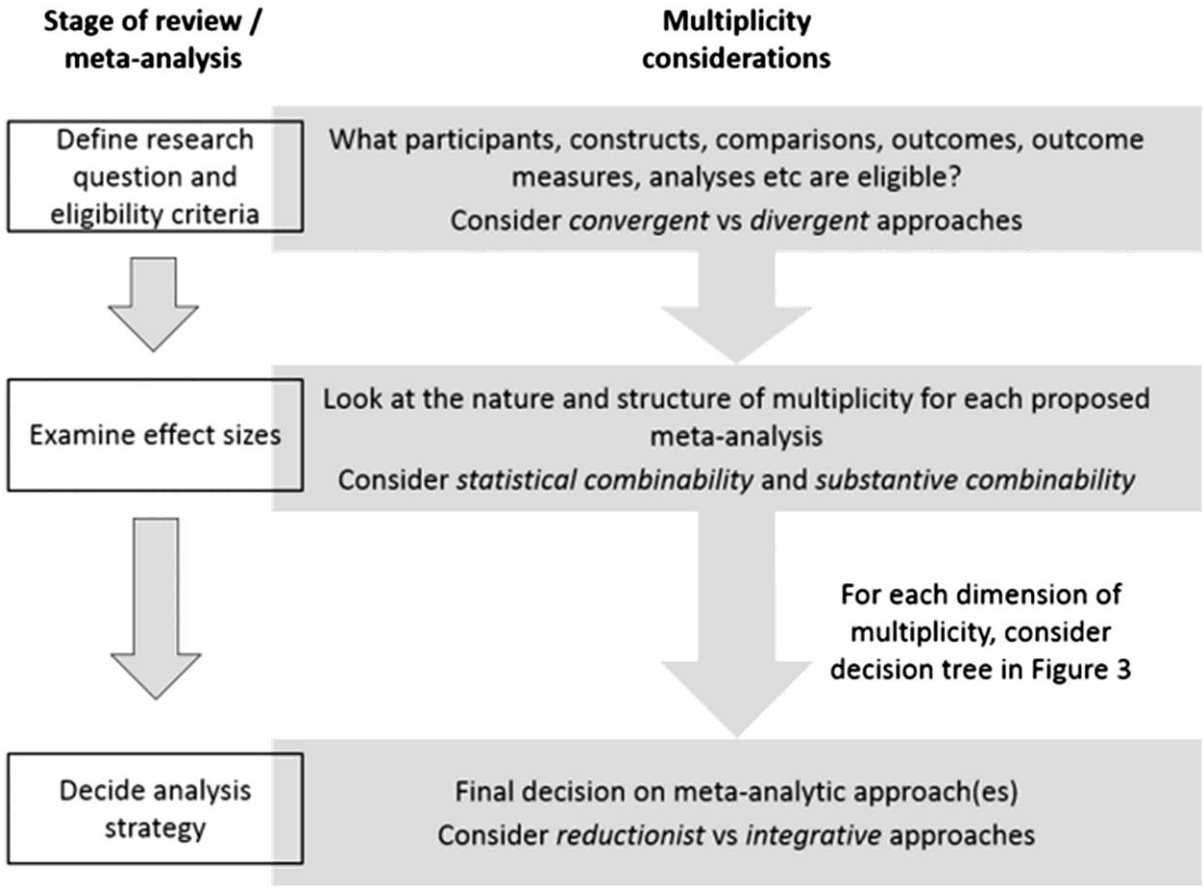

Figure 3. Suggested decision tree for dealing with multiplicity in any particular dimension.