Inclusion and exclusion strategies for conducting meta-analyses

Joshua K. Swift & Bruce E. Wampold

To cite this article: Joshua K. Swift & Bruce E. Wampold (2018) Inclusion and exclusion strategies for conducting meta-analyses, Psychotherapy Research, 28:3, 356-366, DOI: 10.1080/10503307.2017.1405169

To link to this article: https://doi.org/10.1080/10503307.2017.1405169

Published online: 23 Nov 2017.
CONSIDERATIONS OF HOW TO CONDUCT META-ANALYSES IN PSYCHOLOGICAL INTERVENTIONS

Inclusion and exclusion strategies for conducting meta-analyses

JOSHUA K. SWIFT¹ & BRUCE E. WAMPOLD²

¹Department of Psychology, Idaho State University, Pocatello, ID, USA & ²Modum Bad Psychiatric Center and Department of Counseling Psychology, University of Wisconsin-Madison, Madison, WI, USA

(Received 5 May 2017; revised 14 October 2017; accepted 16 October 2017)

Abstract
Some of the most important decisions that a researcher will make when conducting a meta-analysis are decisions about the search strategies and inclusion/exclusion criteria. Decisions regarding inclusion/exclusion criteria serve to define the scope of a meta-analysis and search strategy decisions can have a large impact on how well the results of a meta-analysis actually represent the existing body of literature. In this article, we provide descriptions and recommendations for performing searches and making inclusion/exclusion decisions. We also provide a review of 65 meta-analyses of psychotherapy versus pharmacotherapy in order to offer frequency data on the use of various search strategies and inclusion/exclusion decisions in the field.

Keywords: exclusion; inclusion; meta-analysis; search strategies; systematic review; psychotherapy research

Clinical or methodological significance of this article: This article provides a review of and recommendations for inclusion and exclusion strategies when conducting meta-analyses.
• The recommendations can provide guidance for future meta-analysts.
• This article also provides information for consumers that may be used to judge the quality of published meta-analyses.

Through conducting a meta-analysis one has the ability to concisely amalgamate the results in an area of study. Similar to primary research, while performing a meta-analysis researchers are able to use steps of the scientific method to answer research questions. These steps include developing testable questions and hypotheses, defining the target population and sampling frame, gathering data, and conducting statistical tests of the hypotheses. The results of a meta-analysis can provide information regarding average differences between conditions, event rates, degree of homogeneity in the findings within an area of study, and moderators and predictors of the effect size of interest (Cooper, Hedges, & Valentine, 2009). Given all of these strengths, it is perhaps no surprise that meta-analyses are becoming increasingly popular in psychotherapy research.

Although meta-analyses do have a number of strengths when conducted properly, several methodological issues, if not handled correctly, can lead to results that incompletely or inaccurately represent the research findings in a particular area of study. Decisions about inclusion strategies and exclusion criteria perhaps have the largest impact on the quality of the findings that are obtained through a meta-analysis. In this paper, we describe several considerations and recommendations regarding inclusion and exclusion decisions to help guide researchers in conducting meta-analyses that accurately and adequately address their research questions. The focus of this paper is on conducting meta-analyses in the area of psychotherapy research. Although most of the recommendations that we provide could be applied by any meta-analyst, others are more tailored toward psychotherapy researchers and consumers. Thus, throughout this paper, we provide several examples from recent meta-analyses that have been conducted in the
psychotherapy research field in order to illustrate the recommendations that are made. Figure 1 provides a general flow of the inclusion/exclusion strategies with key questions that should be asked at each stage.

**Developing a Research Question**

Inclusion and exclusion decisions actually begin with developing a well-defined research question. Similar to research questions in primary studies, research questions for meta-analyses should operationally define the constructs of interest and set parameters on the scope of the research that the meta-analysis is meant to cover. Thus, a well-defined research question will include specifics about inclusion and exclusion criteria which will then have an impact on the inclusion and exclusion strategies that are used (Wilson, 2009).

The most important determination of inclusion and exclusion criteria is that the studies selected are sufficient to test the hypotheses of the meta-analysis. For example, if the purpose of the meta-analysis is pragmatic, the inclusion/exclusion criteria must be focused on that purpose. For example, Lambert and Bergin (1994) suggested that patients who are provided a bona fide treatment will show similar outcomes no matter what that bona fide treatment is. Motivated by this assertion, which was based on a narrative review of the literature, Wampold et al. (1997) desired to meta-analytically test this conjecture. However, in order to do so they had to first develop criteria that differentiated bona fide treatments from non bona fide treatments and then only search for and include studies that directly compared two or more bona fide treatments to each other.

In developing a research question, researchers should remember to be focused, but not overly narrow. Although meta-analyses can be performed with data from a very small number of studies, it may be more appropriate to wait until a large enough body of literature can be accumulated to more definitively answer the research question. A meta-analysis that is conducted with data from only a handful of poorly conducted studies could actually portray incorrect results and discourage future more well-defined studies in a topic area. In trying to decide if there is an adequate body of research to conduct a meta-analysis, it is important to remember that power in meta-analyses depends on both the number of studies ($k$) and the sample sizes within each study ($N$). While it is often the case that a large $k$ will produce more power, including studies with very small sample sizes may introduce more heterogeneity, which in turn could reduce the power to detect effects (Hedges & Pigott, 2001). A sufficient number for $k$ to conduct a meta-analysis also depends on the type of model that will be used (random or fixed), the types of analyses that are planned (main effects, moderation analyses), the expected effect size, the level of significance, and expected variance. All of these values depend on the specific area of study and the questions that the meta-analyst wants to address. Thus, a universal minimum value for $k$ cannot be provided; however,
several sources are available that provide suggestions for calculating power for meta-analyses (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges & Pigott, 2001; Jackson & Turner, 2017).

Search Strategies for Study Inclusion

Once the research question is adequately defined, a meta-analyst is ready to search the literature for studies to include in the review. In general, we recommend that researchers use multiple search methods in order to identify as many of the existing studies as possible that are applicable to the specific research question. In addition, at the level of search for inclusion, we recommend that researchers remain broad and over-identify studies that can later be removed based on exclusion criteria. With the search strategies, researchers should primarily be concerned that they are not systematically missing any relevant literature, rather than be concerned with whether the studies fit the research question (that comes later with exclusion considerations).

Term Searches

Perhaps the most common search strategy used by meta-analysts is a term search. This strategy involves identifying terms that represent or that are often used in the particular area of study. Those terms are then entered into a scientific search engine to identify a list of potential articles that may fit the inclusion/exclusion criteria. Again, we recommend that meta-analysts remain broad at this stage of the search to identify as many potentially relevant studies as possible. Thus, the search words should not be overly narrow. When deciding on the terms to use, researchers may want to consider using medical subject headings (MeSH) terms; however, current MeSH listed terms may not be specific enough for many meta-analyses within the area of psychotherapy research. The effectiveness of using MeSH terms in a meta-analysis would also depend on researchers using MeSH terms when they publish their primary research. Instead, researchers can review the terms used in prior meta-analyses, identify terms that are commonly seen in highly cited papers in the field, and/or seek the input from “experts” in the area of study regarding terms that would fit the research question. In addition, researchers should use several synonyms for the constructs of interest. For example, if conducting a meta-analysis on premature termination in psychotherapy, researchers may want to include “attrition,” “dropout,” “premature discontinuation,” “premature termination,” and “unilateral termination.”

In addition, including a large number of search terms with the Boolean operator “AND” between them in most databases will severely limit the number of results that can be obtained. Instead, we recommend that researchers use the Boolean operator “OR” and multiple synonyms for the terms that are chosen. For example, a recent meta-analysis that compared psychotherapy and pharmacotherapy used the terms “psychotherapy OR therapy OR psychological treatment OR psychological intervention OR medication OR pharmacotherapy” (Swift, Greenberg, Tompkins, & Parkin, 2017). Using the “AND” specifier or fewer synonyms for the intervention types would have resulted in a more narrow list of articles and thus an increased likelihood of missing potentially relevant studies. Less experienced meta-analysts may want to recruit the assistance of a librarian to aid in the development of appropriate search terms, operators, and search restrictions.

When performing a term search, we also recommend the use of multiple databases. Previous research has found that searches conducted in different databases (e.g., PsycINFO, PubMed) yield discrepant proportions of relevant articles, and that the discrepancy depends in part on the topic of the search (Wu, Aylward, Roberts, & Evans, 2012). Although using multiple databases is usually recommended, there may be times when, based on the research question, a single database is sufficient. For example, Colman et al. (2016) only used PsycINFO for their search because they were only concerned with studies that included clinical and counseling psychology graduate students as participants. Given this research question, it is unlikely that a more medical database (e.g., PubMed) would have produced any results that were not already identified through PsycINFO. Whether using one or several databases, it is important that the databases that are chosen fit the research question. In addition, although historically only more traditional databases (e.g., PsycINFO, PubMed, EMBASE) have been used, databases such as Google Scholar may also be considered. In many cases Google Scholar will provide a wider array of results, including more unpublished work, professional presentations, and even classroom lectures. This has the advantage of providing a more inclusive search; however, it comes with the cost associated with needing to review additional irrelevant and frequently non-empirical material.

Although term searches are the most commonly used search strategy and will likely yield a high percentage of the relevant research, they are limited based on the coverage of the databases used and the limitations of the specific chosen terms. Meta-analysts may choose from several additional search
strategies to make sure they have adequately covered the existing literature.

**Root and Branch Searches**

A root and branch search typically occurs after the large pool of potential articles that were obtained from the term search has received an initial review. Through the initial review, a smaller subset of likely relevant articles can be identified. These articles represent the “trunk,” or the core set of studies. Each article included in a reference list of any of the articles that belong to the “trunk” represent the “roots.” Each article that cites any of the “trunk” articles represent the “branches.” Both “root” and “branch” articles should be reviewed to identify additional studies that were missed through the term search.

**Review of Previous Meta-analyses**

Another valuable strategy for identifying studies that meet inclusion criteria is to review previous meta-analyses that have been conducted on the same or related topics. Existing meta-analyses may be obtained through the initial term or root and branch searches, through a separate search for meta-analyses, or simply through knowledge of the research that has been conducted in a topic area. Once the related meta-analyses are identified, a meta-analyst can review each of the studies included in the previous meta-analyses to check to see if any meet the current review’s inclusion and exclusion criteria. Included studies from the previous meta-analyses can often be identified in the reference lists or by contacting corresponding authors. A review of related meta-analyses can often add substantially to the list of included studies. For example, in a meta-analysis of premature termination in psychotherapy, the term search produced 198 relevant studies, the meta-analysis search produced an additional 323 included studies (Swift & Greenberg, 2012).

**Talking to Experts**

Another approach for identifying studies is to talk to “experts” in the subject area. The definition of “experts” may range from well-known researchers who have written extensively on the topic to any author of an included study that was identified through a previous search strategy. Experts can be asked to share any work that they have published on the topic, any papers that they know about that have been published by others, and any unpublished studies that they might know about.

**Hand Search Journals**

A hand search of journals can provide another check for studies that may have been missed through term, root and branch, and meta-analysis searches. When using this strategy, a meta-analyst will want to first identify journals that are more likely to publish studies on the topic of interest. If a term search has already been conducted, the reviewer may be able to notice that the identified studies are primarily coming from a select group of journals. Or, a reviewer may specifically choose certain journals based on impact factors and/or reputation in the field. Once the journals are chosen, the reviewer simply checks every study published in the journal to see if it meets inclusion and exclusion criteria. Of course, some of the studies reviewed in the hand search will likely overlap with studies that were already identified through previous search strategies; however, it is likely that some new studies will also be found. In the previously mentioned meta-analysis on premature termination in psychotherapy, 148 of the 669 total relevant studies were found through a hand search of 8 journals. There are times when a hand search is the only choice. In 1997, Wampold et al. (1997) wished to meta-analyze all clinical trials that compared two different psychotherapies. Often the terms “comparison” or “comparative” were not used in the trials and other keywords were inadequate. As a result, they hand searched six journals that published clinical trials of psychotherapy over a 25-year span. This search strategy was repeated with the same six journals in a more recent meta-analysis (Marcus, O’Connell, Norris, & Sawaqdeh, 2014).

**Search of Databases**

Recently, several databases have been established as a way to assist reviewers in the search process. The goals of these databases are to allow a centralized location to gain open access “all” studies on a particular topic. In addition to being easier for reviewers to search and identify studies, these databases have the potential to allow for more transparency in what studies were included and uniformity in the search across meta-analyses. Some examples of these include Cuijpers and colleagues’ database for studies of treatments for depression (Cuijpers, van Straten, Warmerdam, & Andersson, 2008; evidence-based psychotherapies.org), databases for family therapy (Baldwin & Del Re, 2016; shinyserver.byu.edu/family_therapy) and the relationship between aspects of the therapeutic relationship and outcome (see Norcross & Lambert, 2018; Norcross & Wampold, 2018; Baldwin & Del Re, 2016;
shinyserver.byu.edu/alliance_outcome), and the Cochrane Collaboration’s register for controlled clinical trials (Dickersin et al., 2002; cochranelibrary.com/about/central-landing-page.html). Although these databases do have a number of advantages, they are likely not 100% comprehensive in the studies that they include, and so we recommend that meta-analysts consider them as one of several potential search strategies.

**Searches for Grey/Unpublished Literature**

In addition to searching the published literature, many meta-analysts will include a search of the “grey” or unpublished literature. Arguments for and against including grey literature are discussed in the exclusion section of this paper, but if one chooses to include unpublished research, there are several methods one might use to gain access to it. Some of these methods include searching databases that focus on dissertations and thesis projects (i.e., ProQuest Dissertations & Theses A&I), searching trial registries that include published and unpublished studies (e.g., clinicaltrials.gov, who.int/trial-search/), reviewing conference proceedings where relevant studies may have been presented, contacting authors who have previously published on the topic to see if they have any unpublished related work, and making an announcement via listservs specifically asking for unpublished work on the topic of interest.

Even if all of the above search strategies are applied, it is very likely that some studies will be missed—no search can capture everything. This is not necessarily a problem. In primary research, investigators rarely attempt to include an entire population in their study. Instead, they focus on gathering data from a representative sample. The same applies to meta-analyses. Although meta-analysts should attempt to be as comprehensive as possible, the most important questions are: (i) Was there a bias in the search strategies that led to systematically missing a group of studies? and (ii) Would the missed studies change the results if they had been included?

**Exclusion Decisions**

Logically, exclusion criteria are the inverse of inclusion criteria. If a study is included because if it has property X, then it could also be said that the study will be excluded if it does not contain property X. Those conducting meta-analyses always face the problem of whether to have few exclusion criteria (being broad) or many stringent inclusion criteria (be narrow). This has been an issue since the first application of meta-analysis when Smith and Glass (1977) went broad—they did not exclude treatments for those without formal diagnoses (e.g., college students) or studies with poor quality. And they were roundly criticized for those decisions (Andrews & Harvey, 1981; Eysenck, 1978; Landman & Dawes, 1982). Glass and colleagues rebuttal was that including a wide range of studies allows one to test whether a factor (e.g., study quality) moderates effects, an important analysis that is precluded by only including a narrow range (e.g., only studies with high quality) (see Glass & Kliegl, 1983; Glass & Smith, 1978).

As it turns out, selecting only high-quality studies of clinical populations did not change Smith’s and Glass’s conclusions (Wampold & Imel, 2015). Meta-analyses are one of the most open of research strategies because critics and defenders of a conclusion can reanalyze the effects to test various conjectures. For example, they may test the impact of including and excluding various studies, test whether a new variable (e.g., quality of research) moderates the result, and test whether the results have changed over time. As an example of the latter issue, in one meta-analysis Johnsen and Friborg (2015) found that the effects of cognitive-behavioral therapy (CBT) for depression have been decreasing over time.

**Exclusion Based on Study Design**

It is not uncommon for inclusion criteria to involve the type of design employed to answer a particular research question. For example, Borkovec (1990) argued that component studies, which either removed a component of a treatment (dismantling design) or added a component to a treatment (additive design), were the most valid designs for identifying the effect of therapeutic ingredients. Subsequently, two meta-analyses have been conducted that examined the effects of specific ingredients by using only studies with component designs (Ahn & Wampold, 2001; Bell, Marcus, & Goodlad, 2013). As another example, Kivlghan et al. (2015) conducted a meta-analysis of the enduring effects of psychodynamic treatments vis-à-vis other treatments and excluded studies that did not report follow-up results. A third example can be seen in Swift et al.’s (2017) meta-analysis where studies that were not designed as a head-to-head comparison of psychotherapy, pharmacotherapy, or their combination were excluded from the review. Of course, often the designs of interests to psychotherapy researchers assess the association between various constructs, which is often the case for studies of psychotherapy process. The working alliance is the most widely researched process variable and the most recent meta-analysis of the correlation between alliance
and outcome aggregated nearly 200 studies (Horvath, Del Re, Flückiger, & Symonds, 2011).

Researchers must make decisions about restricting their meta-analyses to only comparative trials or to include pre/post designs, to only use studies with follow-up data collection or to include studies that stop at post-treatment or mid-treatment assessments, to focus only on studies that used one particular outcome measure, and so on. There is no right answer when making these exclusion decisions; however, it is important to recognize that with more exclusion criteria, the scope of the results will be limited. Also, researchers who choose to be more inclusive, can also test study design characteristics as moderators of the effect size.

Exclusion Based on Definitions of Psychotherapy

Psychotherapy encompasses a wide variety of practices and definitions of exactly what is involved in therapy. Interestingly, meta-analyses typically do not reference a definition of psychotherapy, but often include various inclusion/exclusion criteria. For example, Thoma, McKay, Gerber, Milrod, and Kocsis (2012) investigated the effects of the quality of clinical trials of CBT for depression and included “group or individual CBT, but it had to be delivered in person by a trained therapist. Trials of bibliotherapy, teletherapy, computer therapy, or Internet therapy were excluded if they did not include a treatment arm with in-person psychotherapy” (p. 23). As is often the case, group and individual therapy were included and not differentiated in the analysis. Mayo-Wilson et al. (2014) in a network meta-analysis of treatments (psychotherapy and psychopharmacology) for social anxiety, included only “first-line” treatments, which was an ill-defined term, but for which consensus was reached. Another ill-defined criterion is “evidence-based treatment” because no generally accepted definition of this term exists (see Wampold & Imel, 2015). As mentioned previously, Wampold et al. (1997) excluded non-bona fide treatments, which were carefully defined as treatments that used a trained therapist in which the therapist developed a relationship with the patient and tailored the therapy to the patient, and that additional met two of the following four conditions: (i) a citation was made to an established approach to psychotherapy, (ii) a description of the therapy was contained in the article and the description contained a reference to psychological processes (e.g., operant conditioning), (iii) a manual for the treatment existed and was used to guide the administration of the psychotherapy, and (iv) the active ingredients of the treatment were identified and citations provided for those ingredients. In this way, various control conditions, such as supportive counseling, were excluded.

Exclusion Based on Classifying Treatments

Often meta-analysts wish to estimate the effectiveness of a particular treatment or to compare the effectiveness of two treatments. Such goals require a definition of a particular treatment as well as operations to decide if a given treatment is a member of the psychotherapy class. There have been three ways to accomplish this classification. The first method is to rely on the study authors’ designation: If a treatment is labeled as CBT or dynamic therapy in the report of the primary study, then it is classified as CBT or dynamic. There are obvious limitations to such a strategy—the treatment so labeled may not have any resemblance to the prototype of the treatment. An example of this is the emotion-focused therapy employed by Shear, Houck, Greeno, and Masters (2001), which bore no resemblance to emotion-focused therapy developed by Greenberg (2010).

The second method is to define the treatment and then develop criteria to operationalize that definition. The problem is that the definitions vary greatly among meta-analyses, as there are no generally accepted definitions of what constitutes a given treatment. What is CBT in one meta-analysis may be classified as non-CBT in another, for example, thus making conclusions about the effectiveness of CBT ambiguous (cf., Baardseth et al., 2013; Tolin, 2010, 2014, 2015; Wampold et al., 2017). A third way is to have experts in a particular treatment classify the treatment. For example, Baardseth et al. surveyed members of the Association of Behavioral and Cognitive Therapies and defined CBT based on consensus of the respondents (see also Kivlighan et al., 2015).

Exclusion Based on Control Treatments

Psychotherapy studies are replete with various types of control treatments. Smits and Hofmann (2009) meta-analyzed the effects of “control conditions” for CBT for adult anxiety disorders and located 19 trials that compared CBT to controls and then estimated the pretreatment to post-treatment effects of the controls. Unfortunately, they did not define what constituted a control treatment or the criteria that were used, but the names of the treatments included supportive counseling, non-directive therapy, discussion groups, relaxation, problem-solving therapy, educational supportive therapy, and digital audio visual integration device (see Table 1, p. 232). This is problematic because in other meta-analyses some of these
“control” conditions were classified as actual treatments (e.g., relaxation has been classified as a bona fide treatment in some meta-analyses of anxiety disorders; see Siev & Chambless, 2007; Siev, Huppert, & Chambless, 2009).

In fundamental ways, the effort to establish the specificity of treatment by comparing a treatment to control treatments without specific ingredients is flawed (see, e.g., Kirsch, 2005; Kirsch, Wampold, & Kelley, 2016; Wampold & Imel, 2015; Wampold, Frost, & Yulish, 2016 for recent discussions). One essential problem is that removing specific ingredients from psychotherapy yields something that no longer resembles psychotherapy, to the therapist delivering the treatment nor to the patient receiving the treatment. As an example, consider a “talking control” for CBT for depression in older people (Serfaty, Csipke, Haworth, Murad, & King, 2011). In this study, the control treatment was designed with special attention to purportedly providing all of the common factors, including therapist enthusiasm, sympathy, being non-judgmental, and encouraging the patient to talk about their history, family, and friends. However, at the same time, the therapist was instructed to stay with “neutral topics such as hobbies, news, holidays, etc.” and avoid setting an agenda, conceptualizing or explaining the patient’s distress, asking the patient about their view of the session, exploring belief systems, and “collaborating with client to solve problems [and] focusing on key problem areas” (p. 434, Table 1, emphasis added). For example, in the talking control, if the patient stated, “I am sure my children think I’m a burden and dread visiting me,” it is suggested the therapist say, “You have children? How many and how old are they?” (p. 434). Unfortunately, meta-analytic comparisons to these flawed controls are speciously used as evidence of specificity (e.g., Cuijpers et al., 2012; Honyashiki et al., 2014).

Exclusion Based on Publication Status

Another narrow/broad decision to make is whether to include the grey literature. From personal experience, we know that searching and coding the grey literature is much work. The question is not whether or not it should be done, but rather, in what circumstances is it necessary. Some arguments have been made against including unpublished literature in meta-analyses. Many of these arguments concern questions of study quality (Borenstein et al., 2009; Eysenck, 1978). Additionally, it may be very difficult to identify and obtain all of the existing unpublished literature even if several different search strategies are used. Thus, if a researcher chooses to use the unpublished literature that can be found, it may not accurately represent all of the unpublished research that is out there. Therefore, researchers may want to exclude grey literature in order to more concretely draw a line around the scope of their meta-analyses. In addition, researchers may have difficulty retrieving the grey literature on their own in order to check the researchers’ results. Thus, meta-analysts who do include grey literature should specifically state that individuals who desire to confirm their results can contact them to have access to the data and/or unpublished studies.

In contrast to the arguments made against including unpublished research, arguments have been made in favor of including grey literature based on the file drawer problem or publication bias (Rosenthal, 1979; Sutton, 2009). Publication bias is evident when publication decisions are based on statistical significance rather than methodological quality. Sometimes these decisions are made by journal editors and reviewers who believe that significant results have the potential for a larger impact on the field, and other times these decisions are made by researchers who never seek publication when their results are not significant. Either way, publication bias is a major issue for meta-analysts because if only published research is included, the results will likely be biased toward significance. That is, the grey literature may contain studies with null results that would reduce the size of the overall effect found in a meta-analysis. Fortunately, there are ways to detect publication bias based on the findings from the studies that are included (see Sutton for a more detailed description of these methods), so a search of the grey literature may not be needed or worth the effort. In addition, the presence of publication bias toward null findings is not guaranteed. For example, meta-analysts may be aware of unpublished studies, including their own or close colleagues, which have produced positive results that support their hypotheses (see Wampold et al., 2017) and unaware of unpublished studies with null results, producing a bias for significant results. Another example can be found in Horvath et al.’s (2011) meta-analysis of the alliance-outcome correlation, where effects from journal articles, unpublished theses, and chapters were all compared. They found no differences among the correlations from these three sources. This demonstrates a direct way to determine whether or not there is a publication bias, which in this case there was not.

Exclusion Strategies

Once exclusion decisions are made, meta-analysts have the difficult tasks of actually reviewing and
removing studies. Some of the more straight-forward exclusions can be done using advanced search options with the initial term search. Many databases, including PsycINFO, provide a number of options to limit the search. For instance, if one is doing a meta-analysis of treatments for anxiety in adults, one can select the Age group “Adulthood (18 Yrs & Older).” Similarly, “Treatment Outcome/Clinical Trials” can be selected for Methodology for psychotherapy researchers who would like to only synthesize the outcomes of clinical trials. Depending on the nature of the meta-analysis and the specific research questions, several other options may be relevant. Utilization of advanced search options can reduce significant amounts of time and effort that would otherwise be spent on screening studies retrieved through term searches. However, many exclusion decisions can only be made after a thorough review and coding of the full text articles.

**Multiple Searchers, Reviewers, and Coders**

In performing the initial search for studies to include in a meta-analysis and in making decisions about whether studies meet exclusion criteria, a researcher’s biases have the potential to influence the decisions that are made. However, these biases may be reduced by following a few simple steps. First, meta-analysts should set up the inclusion and exclusion criteria prior to actually reviewing the literature. By doing this, inclusion and exclusion criteria will be based on the research question, rather than on certain studies that the reviewer may want or not want to include in the review. Second, if at all possible, multiple independent reviewers should be used at each stage of the search and coding for inclusion. The initial review is often done in stages, beginning with a quick review of article abstracts and titles to quickly screen out irrelevant studies. This is followed by a more thorough, full text review to determine if a study meets all of the inclusion criteria. Prior to evaluating the articles found through the search strategies, the independent reviewers should be trained in the review process. This can be done with a small set of studies until proficiency is demonstrated. A codebook should be used to remind the reviewers of what they are looking for and to maintain consistency. Although we do recommend having two independent coders involved for each article at the full text review level, on occasion, the body of articles is too large to have two reviewers evaluate each study. In these situations, a subset of articles can be randomly selected for double review to check for consistency. As well, the reviewers should be blinded to the outcomes of the studies reviewed, for example by presenting to the coders only the methods sections, provided that is sufficient to make judgments.

Coder agreement or interrater reliability can be assessed through several methods (Orwin & Vevea, 2009). Calculating intraclass correlation coefficients (ICCs) is the most commonly used method for determining agreement when coding continuous variables. For categorical variables, the simplest method is to calculate the percentage of agreement \(\%\text{agreed} = \#\text{studies agreed}/\#\text{studies reviewed}\). However, given the number of studies that are often reviewed for inclusion and the fact that most of the studies will be excluded because they are clearly irrelevant, a percentage might give an artificially high impression of the agreement rates even if the reviewers disagreed on many of the pertinent studies. Cohen’s kappa coefficient can be used as an alternative method for testing interrater reliability and is often preferred because it accounts for chance agreement rates. Kappa is calculated through the following formula:

\[
\kappa = \frac{p_{\text{observed}} - p_{\text{expected}}}{1 - p_{\text{expected}}}
\]

The proportion of observed agreement \(p_{\text{observed}}\) is the simple percentage of agreement between the two coders. The probability of chance agreement \(p_{\text{expected}}\) is based on the base rate frequency with which both reviewers include or exclude studies.

\[
\begin{align*}
п_\text{expected} &= \left(\frac{\#\text{coder 1 says include}}{\text{total # reviewed}}\right) \\
&\times \left(\frac{\#\text{coder 2 says include}}{\text{total # reviewed}}\right) \\
&+ \left(\frac{\#\text{coder 1 says exclude}}{\text{total # reviewed}}\right) \times \left(\frac{\#\text{coder 2 says exclude}}{\text{total # reviewed}}\right).
\end{align*}
\]

Kappa values range from 0 to 1, with higher values representing greater agreement. Standards in the field suggest that for primary studies ICC values should be used to calculate agreement for continuous variables and kappa values for categorical variables; the same should be reported by those conducting meta-analyses.

Although interrater reliability estimates provide an estimate of the consistency of the reviewers, they do not provide instruction on how to handle disagreements. We would recommend that interrater reliability be calculated early on as an initial check of the reviewers. If the consistency is low, the reviewers may need re-training or new reviewers may be needed. If consistency is high, then the reviewers may continue with the remaining studies and disagreements can be handled on a case-by-case basis. When a small number of disagreements are present, the two coders may want to come together with a third researcher to collaboratively
Flow Diagrams and Reporting of Inclusion Strategies and Exclusion Decisions

Regardless of the inclusion and exclusion strategies and decisions that are chosen, it is critical that meta-analysts adequately report their criteria and the impact of the criteria in their manuscripts. The Meta-Analysis Reporting Standards which have been adopted by the American Psychological Association provides clear guidelines about the search, inclusion, and exclusion information that should be reported in all meta-analyses (American Psychological Association Publications and Communications Board, 2008). These standards allow the consumer to more adequately evaluate the scope and limitations of the meta-analytic results. One of the clearest ways to illustrate the impact of the search strategies and exclusion criteria is through a flow diagram (see Figure 2 for an example).

Frequency Data from Previous Meta-analyses

In order provide some data on what is frequently done in the field, we reviewed the search strategies and inclusion/exclusion criteria for 65 meta-analyses that compared psychotherapy to pharmacotherapy with adults. These meta-analyses were published from 2001 to 2016 and they included an average of 38.22 primary studies, ranging from two to 153. Starting with the search strategies, 49 of the 65 reviewed meta-analyses reported that a term search was performed. These 49 meta-analyses used a median of 3 search engines/databases, with 34.1% using 1–2 engines, 38.6% using 3 engines, and 27.3% using 4–8 engines. The most common databases that were searched include MEDLINE/PubMed (92%), PsycINFO (65%), EMBASE (51%), and Web of Science (22%). When term searches were performed, an average of 9.33 search terms were used, ranging from two to 34. Just over half of the reviews (54%) included a root search, but only 6% included both a root and branch search.

Figure 2. Example flow chart for meta-analyses.
References


