Annual Review of Psychology

How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses

Andy P. Siddaway,[1] Alex M. Wood,[2] and Larry V. Hedges[3]

[1]Behavioural Science Centre, Stirling Management School, University of Stirling, Stirling FK9 4LA, United Kingdom; email: andy.siddaway@stir.ac.uk

[2]Department of Psychological and Behavioural Science, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

[3]Department of Statistics, Northwestern University, Evanston, Illinois 60208, USA; email: l-hedges@northwestern.edu

Annu. Rev. Psychol. 2019. 70:9.1–9.24

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-010418-102803

Keywords

guide, meta-analysis, meta-synthesis, narrative, systematic review, theory, evidence

Abstract

Systematic reviews are characterized by a methodical and replicable methodology and presentation. They involve a comprehensive search to locate all relevant published and unpublished work on a subject; a systematic integration of search results; and a critique of the extent, nature, and quality of evidence in relation to a particular research question. The best reviews synthesize studies to draw broad theoretical conclusions about what a literature means, linking theory to evidence and evidence to theory. This guide describes how to plan, conduct, organize, and present a systematic review of quantitative (meta-analysis) or qualitative (narrative review, meta-synthesis) information. We outline core standards and principles and describe

commonly encountered problems. Although this guide targets psychological scientists, its high level of abstraction makes it potentially relevant to any subject area or discipline. We argue that systematic reviews are a key methodology for clarifying whether and how research findings replicate and for explaining possible inconsistencies, and we call for researchers to conduct systematic reviews to help elucidate whether there is a replication crisis.

Contents

INTRODUCTION

The task of reading and making sense of the literature on a particular topic was relatively easily achieved in the early decades of psychological science for the obvious reason that literatures were small. Many decades on, the research landscape is vastly different, and this task can now be complicated, time consuming, and stressful. Scientific literatures are now enormous or expanding exponentially, and knowledge is produced and shared rapidly across the world through the Internet. New theories, constructs, and literatures are constantly emerging, and many literatures are being integrated with the goal of locating and understanding a smaller number of core constructs, processes, and mechanisms.

If this situation were not complex enough, research exploring the same question often produces varying or even conflicting findings that could potentially be due to a range of factors. When findings conflict or do not replicate, it can be unclear why this might be the case, what the overall picture is, what important questions remain unanswered, or which results are most reliable and should be used as the basis for practice and policy decisions.

Literature reviews, particularly systematic reviews, are the proposed solution to this complexity. High-quality literature reviews bring together, synthesize, and critique one or more literatures to provide an overall impression of the extent, nature, and quality of evidence in relation to a particular research question, highlighting gaps between what we know and what we need to know. Literature reviews potentially provide a means of making sense of vast quantities of scientific information and are often highly cited and influential. They sit at the top of hierarchies of evidence because they have great potential for informing practice and public policy. The best review articles comment on, evaluate, extend, or develop theory (Baumeister 2013), linking theory to evidence and evidence to theory. Literature reviews also form a key methodology for clarifying whether and how important research findings replicate.

However, literature reviews are certainly not a panacea. They do not automatically contain high-quality, reliable evidence; they are simply a means of synthesizing whatever evidence is available (for a discussion of threats to validity in research synthesis, see Matt & Cook 1994). As discussed below, there are important differences between writing an empirical paper and writing a literature review. We believe that conducting a high-quality, publishable literature review is a highly sophisticated task, and one that has the potential to become stressful or overwhelming. It therefore seems unfortunate that most scientists do not receive training in how to write literature reviews (Baumeister 2013).

This article focuses on systematic reviews (sometimes referred to as research syntheses or research reviews), a particular type of literature review that is characterized—as the name suggests—by a methodical, replicable, and transparent approach. We present a comprehensive guide to conducting and reporting systematic reviews of quantitative or qualitative information for scholars of all levels. Many excellent textbooks and articles have discussed how to conduct and/or present systematic reviews (e.g., Cooper 2016, Cooper et al. 2009, Higgins & Green 2011) and literature reviews more generally (e.g., Baumeister 2013, Baumeister & Leary 1997, Lipsey & Wilson 2001). For a complete picture, we recommend that these sources be read alongside this article. However, textbooks require a substantial time investment for the typical user, and it can be difficult to select among the wide range of resources available. Additionally, many of the existing texts do not methodically detail both the conducting and reporting of systematic reviews, or they have potential relevance for both quantitative and qualitative information. This article condenses all that we have learned and read about systematic reviews into a concise guide in the hope of providing readers with an easily accessible and comprehensive resource.

OVERVIEW

This guide focuses on how to plan, conduct, organize, and present the results of a systematic review, covering all aspects of the review except for the results section. We specifically avoid discussing how to use the data that make up the product of a systematic review (the results section) because there are different, specialized customs and methods for doing this with qualitative or quantitative information. Readers are referred to core texts in each specialty for this information (as discussed below).

We begin with a discussion of the advantages that literature reviews, particularly systematic reviews, confer, including the potential for systematic reviews to contribute to the debate on the

so-called replication crisis. We describe different types of literature reviews to help readers select a tool that fits their purposes and the research context. The bulk of this article details the key stages of conducting a systematic review. We discuss core standards and principles that need to be adhered to and real-world problems that a prospective systematic reviewer is likely to encounter, as well as the means of avoiding or overcoming such problems. In the final section, we describe core principles and standards for presenting systematic reviews to ensure that what is written is both practically useful (in terms of research impact) and appropriate for submission to field-leading journals.

WHY CONDUCT A LITERATURE REVIEW?

To our mind, there are two main reasons for conducting some form of literature review. The primary reason is the desire to synthesize a body of evidence on a topic in order to achieve robust and broad conclusions and implications (Baumeister 2013). Each individual study involves a researcher or a team collecting a sample using particular methods and measures. Because individual studies cannot ever be definitive, bringing together the results of many different individual studies, synthesizing and evaluating them, and uncovering consistency far extend what any single study can ever achieve (Baumeister & Leary 1997, Cumming 2014). Thus, by their nature, review articles have a power and value that no single study can match (Baumeister & Leary 1997, Cumming 2014).

The whole is far greater than the sum of its parts, and high-quality literature reviews involve bringing together and integrating a body of studies in order to (*a*) draw robust conclusions about big questions, principles, and issues, and (*b*) explain how and why existing studies fit together and what that means for theory and future research. The nature and scope of a literature is not always apparent in the absence of a review, and conducting a literature review can thus serve as an extremely useful exploratory exercise. This means that a literature review might be driven by theory (seeking to examine how closely the existing literature supports an existing or proposed theory), or new ideas and conceptualizations might emerge from the process of reviewing and integrating the existing literature.

The second reason for conducting some form of literature review is that doing so is a requirement. A literature review is usually expected in some form or another at most levels of academic study to demonstrate a student's knowledge of a research topic, and it is often expected by funding bodies to demonstrate the need for a proposed research grant. One trap commonly encountered by novice literature reviewers is to simply summarize everything they have come across or can bring to mind on a particular topic, with little critical evaluation or integration. Rather than being a comprehensive, critical, and coherent synthesis, such reviews present a collection of unconnected information and offer little that is novel or that shows evidence of reflection or critical thinking. One of the aims of this article is to help novice literature reviewers sidestep this and other common pitfalls that can arise when conducting and reporting a literature review so that good habits can be established early.

DISPELLING TWO COMMON MISUNDERSTANDINGS ABOUT LITERATURE REVIEWS

Literature Reviews Versus Reviewing Literature

Before discussing systematic reviews and the different types of literature review, it may be instructive to dispel two common misunderstanding about literature reviews. The first is that conducting a literature review is the same as the task of reviewing literature, which occurs when writing the introductory sections of all quantitative and qualitative journal articles (including review articles). Reviewing literature involves selectively discussing the literature on a particular topic to make

the argument that a new study will make a new and/or important contribution to knowledge. In contrast, literature reviews make up a distinct research design and type of article in their own right. Rather than selectively reviewing relevant literature to make a flowing rationale for a study's existence, they provide a comprehensive synthesis of the available evidence to allow the researcher to draw broad and robust conclusions.

Vote Counting

It is also worth strongly cautioning against vote counting (Light & Smith 1971), which might at first glance seem a useful means of summarizing quantitative research findings. Vote counting involves assigning one of three outcomes (positive, negative, or no relationship) to each study in a review based on that study's statistical significance. The basic idea is that a research hypothesis is deemed to be supported if a large proportion of studies on a topic find a statistically significant effect (Bushman 1994, Hedges & Olkin 1980).

Although vote counting has an appealing simplicity, it is deeply flawed. It does not take into account sample size, which affects statistical power and the precision with which a sample is representative of the population of interest, and it does not provide an estimate of the size of an effect (Bushman 1994, Hedges & Olkin 1980). It also performs increasingly poorly as the number of studies increases (see Hedges & Olkin 1980).

WHAT IS A SYSTEMATIC REVIEW?

A systematic review is a special type of literature review that confers added advantages. It is "a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review" (Cochrane Collab. 2003). Systematic reviews are characterized by being methodical, comprehensive, transparent, and replicable. They involve a systematic search process to locate all relevant published and unpublished work that addresses one or more research questions, as well as a systematic presentation and synthesis of the characteristics and findings of the results of that search. The systematic methodology and presentation aim to minimize subjectivity and bias. The best and most useful systematic reviews use the literature reviewed to develop a new theory or evaluate an existing theory and/or have clear implications for policy or practice.

The criteria for inclusion and exclusion in the systematic review are explicitly stated and consistently implemented such that the decision to include or exclude particular studies is clear to readers and another researcher using the same criteria would likely make the same judgements. This explicit approach allows readers of the review to assess the author's assumptions, procedures, and conclusions and enables other researchers to update and extend the review at a later time.

To best achieve the purposes of a systematic review, we like Baumeister's (2013) advice to adopt the mindset of a judge and jury rather than a lawyer. A judge and jury skeptically evaluate the evidence to render the fairest judgment possible, whereas a lawyer's approach is to make the best case for one side of the argument. Returning to the differences between a literature review and the task of reviewing literature, the introduction section of a quantitative or qualitative article is usually written using the lawyer's approach.

The nature of systematic reviews means that they are potentially able to achieve the following outcomes (Baumeister 2013, Baumeister & Leary 1997, Bem 1995, Cooper 2003):

- draw robust and broad conclusions by producing an unbiased summary of what the cumulative evidence says on a particular topic;

- critique and synthesize one or more literatures by identifying relations, contradictions, gaps, and inconsistencies and exploring the reasons for these;
- develop and evaluate a new theory or evaluate an existing theory or theories to explain how and why individual studies fit together;
- provide implications for practice and policy; and
- outline important directions for future research (e.g., highlighting where evidence is lacking or of poor quality).

WHY CONDUCT A SYSTEMATIC REVIEW RATHER THAN ANOTHER TYPE OF LITERATURE REVIEW?

Systematic reviews are becoming increasingly popular, but they are not the default literature review strategy. We recommend that a systematic review be conducted whenever possible for several reasons. First, their very nature means that they tend to be of higher quality, more comprehensive, and less biased than other types of literature review, which makes them more likely to be published and to have an impact. If done well, a systematic review is a novel and important substantive contribution to knowledge in its own right.

Second, the high quality and transparency of systematic reviews means that they are a relatively safe bet with academic markers and journal peer reviewers. Indeed, many of a marker's or reviewer's comments or concerns can be assuaged if a researcher has conducted a systematic review in line with best practice (e.g., by answering the following questions: How and why were particular studies included or excluded? What is the extent, nature, and consistency of the literature? Is the review coherent and clear? Do its conclusions seem like they can be trusted because they directly relate to the available evidence?).

Another reason for conducting a systematic review may be the most emotionally salient: It is simply far less stressful and far more manageable to conduct a systematic review than to conduct some other type of literature review. This is because the systematic structure and methodology that is apparent throughout the systematic review process imposes a discipline and a focus that make the task of conducting and presenting the review tangible and digestible. Conducting a systematic review involves breaking a potentially massive task down into sections and subsections and enables progress to be monitored concretely. These things are good for the soul when focusing on the same piece of work for months or years! In the rare instances in which a systematic review is not suitable (discussed below), many facets of the systematic review approach can still be utilized with commensurate benefits in terms of quality and rigor.

So far, we have argued that conducting a systematic review entails several major benefits for the researchers conducting the review, for the literature being synthesized, and for clinicians and policy makers. We also believe that systematic reviews offer broader benefits still—for science itself. This guide is written at a time when some scholars are concerned that some areas of science may be experiencing a replication crisis (see Nelson et al. 2018, Shrout & Rodgers 2018). We see systematic reviews as a critical means of clarifying whether important research findings meaningfully replicate. They are therefore likely to become an increasingly central pillar of psychological science.

What is the so-called replication crisis? The replication crisis literature has drawn attention to the fact that some key scientific findings do not replicate (e.g., Open Sci. Collab. 2015). This is a critical issue because reproducibility—the extent to which consistent results are observed when individual studies are repeated—is one of the defining features of science. However, as with everything in psychology, there are no simple or absolute answers, and increasing discussions of the nuances and facets of reproducibility have accompanied discussions of the alleged crisis.

The issue of reproducibility is, in fact, complex. A failure to replicate a finding does not conclusively indicate that an original finding was false because there are myriad possible reasons for findings not replicating, including insufficient power; researcher degrees of freedom (discretion in collecting and analyzing data); publication bias; questionable research practices (e.g., rounding down *p*-values, falsifying data); problems with the design, implementation, or analysis of the original or replication study; failure to recognize and document the circumstances and social context in which research took place; changes in the population over time; and other known and unknown factors (Braver et al. 2014, Cesario 2014, Cumming 2014, Earp & Trafimow 2015, Etz & Vandekerckhove 2016, John et al. 2012, Klein et al. 2012, Maxwell et al. 2015, Open Sci. Collab. 2015, Patil et al. 2016, Stroebe & Strack 2014). Several authors have also highlighted that there are no established or agreed-upon criteria for deciding that a finding has replicated or what replication means (e.g., Valentine et al. 2011). Taken together, these issues indicate that there should be less emphasis on individual studies—even landmark studies—and instead that the emphasis should be on a consensus of findings across studies and methods that have matured to the stage where there is a clear and consistent overall picture. This is especially the case when recommendations are being made for policy and clinical practice. [For a sobering illustration of the consequences of not basing clinical practice on the consensus of the evidence, see Chalmers (2007), Lau et al. (1992).]

It follows that, rather than conducting additional costly replication projects to examine the alleged replication crisis, the practice of systematic reviewing could be more widely employed. Systematic reviews offer the most robust means of clarifying the extent, nature, and quality of the evidence on a particular topic. They can therefore contribute to the issue of replicability in important ways, potentially fostering scientific rigor and maintaining a robust reputation for psychological science.

First and most obviously, the very nature of systematic reviews means that they themselves (i.e., their results) are reproducible. As discussed, systematic reviews aim to be comprehensive, methodical, explicit, transparent, and as unbiased as possible in the questions they explore and how they explore them. Inclusion criteria are explicitly described and consistently implemented, meaning that if another researcher conducted exactly the same search, they should end up with exactly the same results. Of course, different researchers might present the results of a search slightly differently or make slightly different decisions about how to use those results; however, systematic reviews are certainly less biased and more reproducible than other types of literature review.

Systematic reviews can additionally contribute to the issue of replicability through concerted efforts to include and critique all potentially relevant published and unpublished work (we discuss the inclusion of unpublished work in the section titled Publication Bias). Again, this is not a foolproof strategy (see Ferguson & Brannick 2012), but it is a possible advantage usually conferred by systematic reviews over other types of literature review and certainly over individual studies.

High-quality literature reviews of any type should be less affected by measurement error, publication bias, and several other biases that impair replicability in individual studies precisely because literature reviews occur at a higher level of abstraction than individual studies. Taking a bird's-eye, critical view means that literature reviewers are less likely to capitalize on statistical chance or fall victim to the simplistic and fallacious thinking that often comes as a consequence of null hypothesis significance testing (NHST) (see Cumming 2014).[1]

[1]Meta-analyses often have high power, but not always. Type II errors can still be a problem (Cohn & Becker 2003, Hedges & Pigott 2001). A 5% significance test in meta-analysis has a 5% chance of a Type I error, just like in primary research.

Additionally, in contrast to authors of individual studies who operate within the publish (novel findings) or perish zeitgeist, systematic reviewers can allow themselves to be led by the available evidence, whatever that looks like. Systematic reviewers are constrained by what other researchers have already done, and this context offers the advantage of making the reviewer somewhat immune to the pressure to publish statistically significant findings that authors of individual studies may perceive (this is publication bias and is discussed below). A literature review can make a useful contribution to the field by concluding that the existing data are inadequate to answer some question or that the current literature challenges a favored theory or line of research, if that is what the current evidence indicates (Baumeister & Leary 1997).

DECIDING WHEN TO DO A QUANTITATIVE OR A QUALITATIVE RESEARCH SYNTHESIS

One of the two starting points for conducting a literature review is to decide what type of review is most appropriate (the other starting point is establishing whether a review is needed). At the broadest level, there are two classes of review article, one involving quantitative information (quantities) and one involving qualitative information (qualities or types). Whether a qualitative or a quantitative approach is most appropriate will depend on the nature and state of the existing literature, the research questions, and theoretical and empirical issues. Just as particular statistical tests are selected to address particular research questions, the type of literature review conducted must logically fit the aims of the review.

Meta-Analyses

When the reviewer wishes to bring together many studies that have empirically tested the same hypothesis, a quantitative review is called for. This is a meta-analysis, for which there are many excellent textbooks (e.g., Borenstein et al. 2009, Cooper 2016). Meta-analysis is concerned with estimation (e.g., To what extent do job search interventions facilitate job search success?).

Meta-analysis would be appropriate when a collection of studies

- report quantitative results (data) rather than qualitative findings or theory;
- examine the same or similar constructs/relationships;
- are derived from similar research designs;
- report the simple relationships between two variables (bivariate relationships, zero-order correlations, single-degree-of-freedom contrasts), rather than relationships that have been adjusted for the effect of additional variables (e.g., partial or multivariate effects);[2] and
- have results that can be configured as standardized effect sizes (Borenstein et al. 2009).

There are many possible effect size indexes, and which one is the most appropriate depends on the nature of the data and the research design used by the included studies. When studies compare mean scores from treatment groups on continuous outcome variables, effect sizes based on the standardized mean difference (Cohen's *d* or Hedges's *g*) may be appropriate (see, e.g., Borenstein et al. 2009). When studies examine the relation between two continuous variables, effect sizes based on the correlation coefficient may be appropriate. When studies compare two treatment groups using dichotomous outcome variables, effect sizes may be based on the difference in proportions, the ratio of proportions (a risk ratio), or a more complex comparison called the odds ratio (see,

[2] Meta-analytic methods are constantly evolving, and methods to synthesize individual answers or use partial or multivariate effects are emerging. Interested readers are referred to the meta-analysis literature.

e.g., Fleiss & Berlin 2009). Sometimes these effect size measures are transformed as part of the analysis (for example, ratios of proportions are almost always log-transformed before statistical analysis), but summaries are usually more understandable when presented in the original metric (e.g., ratios of proportions rather than log ratios).

Meta-analysis usually summarizes effect sizes by a measure of central tendency (often a weighted mean) and a representation of its uncertainty, such as the standard error of that weighted mean or a confidence interval. It is also conventional to provide some measure of the consistency or heterogeneity of study results because effect sizes can be influenced by a potentially large number of characteristics that vary among studies. The *Q* statistic provides the standard significance test for between-study heterogeneity, and the T^2 statistic quantifies the amount of true heterogeneity between studies (Borenstein et al. 2009, 2017).

Two sources of variability might cause heterogeneity among the studies included in a meta-analysis. One is variability due to sampling error (within-study variability). This variability is always present in a meta-analysis because every study uses a different sample. The other source of variability is between-studies variability, which can appear when there are true differences among the population effect sizes estimated by individual studies. If there is statistically significant between-studies heterogeneity, moderator variables can be examined to explain it (e.g., participants, measures, treatment conditions, study design).

Graphical displays such as forest plots can be an economical means of displaying study effect sizes and their uncertainties (in the form of confidence intervals) so that the distribution of estimates can be evaluated. For example, forest plots make it easy to determine whether one or more studies appear to substantially disagree with the bulk of the other studies. Meta-analysis is able to test and account for publication bias and to make a particularly valuable contribution to the replication crisis debate because instead of simply noting whether each replication attempt did or did not reach statistical significance or replicate, the data from all the studies on a topic can be combined to estimate the effect in the population.[3]

Narrative Reviews and Meta-Syntheses

To date, systematic reviews have generally tended to be associated with meta-analysis. However, reviews of qualitative information can also be conducted and reported using the same replicable, rigorous, and transparent methodology and presentation. There are two types of qualitative research synthesis: narrative reviews and meta-syntheses.

A narrative review would be appropriate when a literature review is desired in relation to a collection of quantitative studies that have used diverse methodologies, or which have examined different theoretical conceptualizations, constructs, and/or relationships (Baumeister 2013). Narrative reviews synthesize the results of individual quantitative studies with no reference to the statistical significance of the findings. They are a particularly useful means of linking together studies on different topics for reinterpretation or interconnection in order to develop or evaluate a new theory (each piece of evidence reviewed draws its value from how it helps build or evaluate the overarching theory; see Baumeister & Leary 1997). For example, Baumeister & Leary (1995) synthesized a wide range of separate literatures to elaborate the theory that a need to belong is a pervasive and powerful human motivations. Narrative reviews can also be used to provide a historical account of the development of theory and research on a topic (although the contribution

[3] Bayesian statistics might even be used to compare the weight of evidence for competing (null and alternative) hypotheses (Braver et al. 2014, Etz & Vandekerckhove 2016).

to knowledge will be relatively minor; see Baumeister & Leary 1997). Readers who are interested in further discussion of narrative reviews and how to conduct them are referred to Baumeister & Leary (1997).

A meta-synthesis [also referred to as meta-ethnography (Noblit & Hare 1988) and qualitative meta-analysis (Schreiber et al. 1997)] would be appropriate when a review aims to integrate qualitative research. The aim of a meta-synthesis is to synthesize qualitative studies on a topic in order to locate key themes, concepts, or theories that provide novel or more powerful explanations for the phenomenon under review (Thorne et al. 2004). Readers who are interested in further discussion of meta-syntheses and how to conduct them are referred to Noblit & Hare (1988), Paterson et al. (2001), and Thorne et al. (2004).

KEY STAGES IN CONDUCTING A SYSTEMATIC REVIEW

Below, we discuss key considerations for conducting a systematic review and address directly to the reader directly in the hope of increasing clarity and usefulness.

Scoping

Several key issues need to be considered as a first step in conducting a systematic review.

Formulate one or more research questions. What do you want to know, and about what topics? Do you have a clear idea of the type of research findings that will be relevant to addressing your research questions? Who will be your audience? Why will a review be useful? Clear, specific, and answerable research questions are the starting point for a clear and comprehensive review.

Consider the breadth of the review. Next, give some consideration to the breadth of your research questions. Examining a narrow research question obviously makes a reviewer's task simpler, faster, and easier, but it also limits the breadth of the conclusions that can be drawn. A review's breadth will depend on the nature of the literature, the reviewer's aims, time constraints, and pragmatics. If the reviewer is an undergraduate or master's student, then, because of time constraints and skill, they would probably need to focus their research questions quite narrowly or select literatures that will yield relatively few results to make their task achievable. PhD students, depending on their skill level, ambition, and the possibility of collaborating with other PhD students, can potentially focus their research questions more broadly. It might be that a researcher has a grant that will fund one or more postgraduate research assistants to work for several years on a particular review; such a situation would potentially allow for a very broad or large literature review to be conducted.

Clarify whether a review has already been done in this area. Search thoroughly to discover whether a systematic review of your research questions has already been done or is registered as an ongoing review. This search will begin the process of familiarizing you with the literature, save you weeks or months of wasted work if a systematic review already exists and does not need updating, or help provide a rationale for an updated systematic review.

Become familiar with the literature. We advise that this searching for existing systematic reviews should be supplemented by some reading of the literature to get a fair idea of what the literature looks like. This will give you a general sense of the scope of the review, potential patterns

in the literature, and the types of research questions that could potentially be examined to make a novel, significant contribution to scientific knowledge.

Updating an existing systematic review. If a review has already been done in the area you are interested in, all is not necessarily lost. Some rationales for conducting an updated rather than an original systematic review might be (*a*) It has been 10 years since the last systematic review on this topic and the literature has rapidly expanded since then, meaning that new important studies and developments need to be accounted for; (*b*) The last review of this topic area was methodologically flawed in various ways that you intend to address with your review; (*c*) The last review focused on X but you think it is worth focusing on Y for particular important theoretical or empirical reasons. For example, a review was conducted 15 years ago on the relationship between trait self-control and a wide range of behaviors. The review was not systematic, and the way that self-control is conceptualized and measured has dramatically changed over the past 15 years. These conditions would make a new systematic review very appropriate in this area.

Planning

Having established a clear need for a systematic review, the next step is to carefully plan the review.

Formulate unambiguous search terms that operationalize your research questions. Break research questions down into individual concepts to create search terms. Search terms are needed to conduct a search that successfully locates all potentially relevant work. For less experienced researchers, reading the existing literature and consulting collaborators and supervisors can help translate research questions into clear and relevant search terms.

Consider different terminology. It is always worth thinking of alternative terms and concepts that may have potentially addressed the same question, as it is common for a range of terms to be used to describe the same phenomenon or research area (as you will learn from your scoping efforts). This is particularly the case when conducting a narrative review. Consider

- synonyms (e.g., "recycle," "refuse," "salvage," "recover");
- singular versus plural forms, verbal forms, adjectives (e.g., "recycling," "recycled");
- different spellings (e.g., "color," "colour");
- broader versus narrower terms (e.g., "Britain," "England," "Scotland," "Wales"); and
- classification terms used by databases to sort their contents into categories listed by headings and subheadings, if relevant to your search.

There is a balance between sensitivity (finding as many articles as possible that may be relevant) and specificity (making sure those articles are indeed relevant). We recommend that, at this stage, your search terms err on the side of sensitivity so that you do not miss anything. Although this will yield more results, and most studies will not be relevant, a large pool of studies can be whittled down relatively swiftly and you will not miss important studies.

Formulate preliminary inclusion and exclusion criteria. Based on your emerging knowledge of the literature, formulate a list of inclusion and exclusion criteria that will allow you to specifically address your research questions and clearly define the boundaries of the review. The inclusion and exclusion criteria used will depend on the topic of the systematic review, as well as on theoretical, empirical, and methodological issues specific to the literature. Best practice involves formulating inclusion and exclusion criteria purely based on your research questions (before even encountering

a literature, so that they are unaffected by what studies are out there) and applying these consistently throughout the review process. Studies that are eligible for inclusion will meet the inclusion criteria and not meet the exclusion criteria.

Justify inclusion and exclusion criteria. It goes without saying that your reasons for including particular studies need to be based on theoretically and/or empirically defensible grounds, rather than, for instance, disagreeing with a particular author's conclusions or theory. Readers of the review will interpret the results and conclusions within the context of your inclusion and exclusion criteria, with implications for the review's generalizability and relevance.

Common inclusion and exclusion criteria concern

- research questions (topic, scope),
- definition or conceptualization (terms and concepts are often defined differently, depending on theoretical or empirical considerations; e.g., "stress management" may be defined or conceptualized differently by different authors),
- measures or key variables (what is measured and how; e.g., whether measures need to meet particular psychometric criteria to be included),
- research design (e.g., observational studies, experimental studies, quantitative studies, qualitative studies),
- participants (e.g., adults, children, individuals with a learning disability, business leaders),
- time frame (e.g., since the start of the literature, since the last systematic review), and
- data (e.g., to be included in a meta-analysis, studies need to report an effect size on the relationship of interest or else provide sufficient information that could be used to compute an effect size).

Revisit and reflect on inclusion and exclusion criteria. As stated, best practice is to formulate inclusion and exclusion criteria before you begin sifting the literature. However, research is rarely a straightforward process and often involves going back to the drawing board. Furthermore, if the prospective reviewer is not already an expert on the subject area, it may prove difficult or impossible to formulate the final inclusion and exclusion criteria before beginning the review, particularly if the review is being conducted on diverse topics and literatures. Indeed, this situation is probably the norm, because the majority of literature reviews conducted are undertaken by undergraduates, master's students, and PhD students.

Because it may take some time before one becomes familiar enough with a literature to make a clear plan for a systematic review, how does one formulate inclusion and exclusion criteria a priori? We advise lots of critical and careful thinking when formulating which inclusion and exclusion criteria to adopt (weighed against the nature of the literature, pragmatic considerations, etc., as discussed above). Some flexibility and reflection when planning, scoping, and formulating inclusion and exclusion criteria may be necessary. However, inconsistently applying those criteria to a body of individual studies or when presenting the results of a literature sifting process is not acceptable.

The initial process of undertaking a large-scale, complex systematic review (until the focus is completely clear), therefore, often involves cycling between coming up with potentially appropriate inclusion and exclusion criteria and seeing whether systematically applying these during literature searching satisfies the research questions and pragmatic issues. For example, we have found it useful to start with a set of studies that we already know are relevant (e.g., from a prior review of the topic or because they seem obviously relevant and includable). These studies can then be used to test one's search strategy by checking that the search at least yields the studies that

are known to be relevant. This process is useful for identifying why any omitted studies were not located by a search and allows the researcher to improve the search terms.

Borderline cases. When you begin sifting the literature, you may encounter studies that are near misses or borderline cases for which either inclusion or exclusion could be argued (e.g., studies that breached a particular age range with part of the sample). The inclusion of these studies requires careful consideration, recourse to theory and/or evidence, and probably discussion and shared decision making between colleagues. If you are conducting a meta-analysis, you can test whether including borderline cases matters to the results.

Of course, if borderline cases that emerge as a result of literature sifting do highlight a potential conceptual or empirical flaw with your current inclusion and exclusion criteria, then those criteria will need to be revised. You will then need to carefully repeat the entire literature searching and sifting process to ensure that all potentially relevant studies are included and all potentially irrelevant studies are excluded. This again points to the importance of thorough planning and scoping before beginning literature sifting.

Create clear record-keeping systems and keep consistent and meticulous records by working systematically. One last step is strongly recommended before comprehensively searching and sifting the literature. We recommend creating one or more record-keeping systems to record what you do and why (i.e., your decision making) at different stages of the systematic review. This may seem an unnecessary effort, but if the literature is large, it is literally impossible to remember exactly what you did, when, and why for thousands of different decisions over months or years. If you need to repeat or check anything, this record will save you a lot of time. You can use the information you record to write the method section of your article, to compute an inter-rater agreement statistic (see the section titled Inter-Rater Reliability), and to respond formally or informally to peer reviews or queries. We suggest that you do the following:

- Make a record of the details of the searches you do and their results.
- Make a list of the number of studies excluded at the screening stage (i.e., based on their title and/or abstract).
- Make a table to record individual studies that were excluded at the potential eligibility stage (based on reading the full text), along with reasons for excluding each study based on your inclusion and exclusion criteria. Common reasons for exclusion are publication type (e.g., nonempirical article), study design (e.g., unsuitable data), measure (e.g., unvalidated measures), and participants (e.g., too old or young). This step is particularly important because it justifies the exclusion of studies that some readers might have expected to be included.
- Make a table that briefly describes the efforts made to find and retrieve unpublished work.
- Make a table to describe in detail the characteristics of studies included in the review. This step is described further below.

Adhere to recommended reporting standards. There are a number of guidelines outlining how to report systematic reviews and meta-analyses, including many that are methodology specific (e.g., CONSORT guidelines for reporting randomized controlled trials; see Moher et al. 2010). Some of the main guidelines and checklists are PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses; Moher et al. 2009), MARS (Meta-Analysis Reporting Standards; APA Publ. Commun. Board Work. Group J. Artic. Rep. Stand. 2008), and MOOSE (Meta-analysis Of Observational Studies in Epidemiology; Stroup et al. 2000). Some reporting guidelines have begun to emerge in relation to meta-syntheses (e.g., Tong et al. 2012). The PRISMA checklist

is the most widely applicable across different research areas. PRISMA recommends creating a systematic review protocol that describes the rationale, hypothesis, and planned methods of the review. PRISMA also recommends that this protocol be prepared before a review is started, made publicly available, and recorded in a registry such as PROSPERO.

Identification (Searching)

It is now time to conduct a methodical and comprehensive literature search.

Search at least two different electronic databases. When you conduct a literature search, your aim is to find all available published and unpublished work that addresses your research questions, operationalized through your search terms. The best way to find relevant published work is to carefully search at least two different electronic databases. We suggest that you do the following:

- Select databases that are relevant to your topic area (e.g., Medline, EMBASE, ISI Web of Knowledge.
- Consider which parts of articles you want to search (e.g., abstract, full text, title).
- Consider using limits and filters to search by article type (e.g., review articles and research syntheses, empirical articles), subject categories, subheadings, etc.
- Consider using AND and OR Boolean search operators. AND will search for all of your search terms (e.g., "cognition AND memory"). OR will search for at least one of your search terms (e.g., "cognition OR memory"). We strongly caution against using the NOT search operator (which excludes certain search terms; e.g., "cognition NOT memory") because it can have odd implications for search results. We also strongly caution against including research design terms (e.g., "longitudinal") in an AND or OR search string because study design information can be incorrectly labeled in databases.
- We strongly recommend using a truncation symbol (e.g., $ or *, depending on the database) to look for all words starting with a particular combination of letters. For example, "weigh$" or "weigh*" will retrieve variations on the word "weigh," such as "weighing" and "weight".
- Perhaps consider using a wildcard symbol (e.g., # or ? or $, depending on the database) to stand in for one character (e.g., "wom#n" will find "women" and "woman"). However, again, we caution that using this can potentially have unforeseen implications for search results.
- Perhaps consider using the truncation symbol (*) between words (e.g., "midsummer * dream" will return results that contain the exact phrase "a midsummer night's dream").
- Perhaps consider using parentheses because commands within these run first [e.g., "(smoke OR tobacco)"].
- Perhaps consider searching by proximity to search for one word within *n* number of words of another word (ADJn or NEAR/n, depending on the database). For example, "self-control ADJ3 behavior" will retrieve records where "self-control" and "behavior" appear within three words of each other.
- Consider searching by publication year, but only if you have a good theoretical or empirical reason for doing so (e.g., if updating a previous review).

Inter-rater reliability. Best practice guidelines for conducting systematic reviews argue that the literature search and sifting process is ideally conducted by two separate reviewers, who must both agree on work to be included. Some sort of quantitative measure of inter-rater agreement on studies to be included may be reported; the intra-class correlation coefficient or Cohen's kappa are the most appropriate statistics for this purpose. The process for resolving disagreements between assessors should be specified explicitly in the review. For example, disagreements over inclusion

should be discussed and, where possible, resolved by consensus after referring to the inclusion and exclusion criteria and relevant theoretical and empirical issues. Discussing specific examples (studies) in your method section, if needed, will illustrate this process.

Although best practice guidelines suggest that two separate reviewers are needed to perform literature searching and sifting, in practice, this often does not or cannot happen (e.g., because the systematic reviewer is a student). It is possible that a single individual could correctly conduct an extremely high-quality and publishable systematic review. Some flexibility may therefore be called for on this issue. What is critical is that the reviewer should provide sufficient information to reassure readers that the systematic review was conducted and reported in line with best practice.

Carefully inspect the search results. You have conducted your searches and the results are back. Examine the search results and read a few of the better quality and more recent relevant articles. Consider the following questions:

- Do the search results suggest that your inclusion and exclusion criteria are reliable and effective in identifying potentially relevant articles and balancing specificity and sensitivity?
- If not, do you need to revise your inclusion and exclusion criteria or search terms?
- Do the search results reveal new search terms that would make a useful addition to your existing search terms?

If your search results suggest that you need to modify your search terms and/or inclusion and exclusion criteria, you will need to return to the planning stage and rerun the search. For example, say that you searched for studies on the impact of physical exercise on well-being. Having read some relevant articles to familiarize yourself with the literature, you discover that you need to widen your search terms to include "lifestyle interventions" because several potentially relevant studies were missed in your first search.

Conduct additional searches to ensure that all potentially relevant published and unpublished work has been located. Once you are confident in your search terms and inclusion and exclusion criteria, you will need to widen your search process. Searching electronic databases provides a very solid start and will certainly help you decide upon the scope of your review. However, electronic databases are not totally comprehensive, and a minority of potentially relevant work may be missed. Thus, additional searches are required for published and unpublished work. A number of methods may uncover potentially eligible work that may have been missed at the database searching stage:

- Read the reference section of work located through electronic database searches that is suitable for inclusion. This will provide you with (*a*) a list of potentially relevant work and (*b*) a list of journals containing relevant studies that you can then specifically search if needed.
- If needed, manually search journals. This can sometimes identify articles and other works (e.g., letters) that have not been included in electronic databases, are not indexed, or have been indexed incorrectly.
- If your database (e.g., Medline and Web of Science) did not cover non-English-language articles and conference proceedings, search for these.
- Locate relevant book chapters, perhaps for data, but more likely for references of relevant work that you want to track down.
- If the information reported in a published study is insufficient to make a decision about inclusion, try to contact the author to enquire about additional details or data.

Publication bias. By their nature, systematic reviews aim to be comprehensive. One key component of the methodology of systematic reviews is therefore a concerted effort to search for

and include relevant unpublished work that meets the inclusion criteria to reduce the effects of publication bias. Publication bias [also called the "file drawer problem" (Rosenthal 1979) or "bias against the null hypothesis" (Cooper 2003)] describes the tendency to only publish research that demonstrates statistically significant ($p < 0.05$) results. This practice arose because statistical significance has generally been considered good or important, whereas the absence of a statistical difference has been considered bad or trivial (Cumming 2014, Rosenthal 1979). Publication bias therefore occurs because researchers who find statistically nonsignificant findings do not submit their results for publication, or if they do, their manuscripts are rejected by reviewers and/or journal editors (Rosenthal 1979).

Publication bias arises because of a flawed and simplistic understanding of what statistical significance (i.e., NHST) actually means. Many scholars over several decades have drawn attention to the fact that *p*-values are limited because they incorporate information about the size of an effect and the size of the sample. More concretely, a statistically significant *p*-value may reflect a large effect size or a small effect size that has been measured in a large sample; a statistically nonsignificant *p*-value may reflect a small effect size or a large effect size measured in a small sample; and two studies with exactly the same effect sizes could vary greatly in their significance level depending upon the sample size (see Cumming 2014 for a detailed discussion and illustration of these issues). The $p < 0.05$ cutoff is of course arbitrary, as beautifully illustrated by Rosnow & Rosenthal's (1989, p. 1277) words: "Surely, God loves the 0.06 nearly as much as the 0.05."

Returning to our discussions of vote counting and the replication crisis, determining successful replication based on whether or not a replication attempt achieved statistical significance is not good science; the fact that one study produces a statistical difference whereas another does not produce a statistical difference does not necessarily indicate that the results of the two studies are statistically different from each other (e.g., the different significance values may be attributable to sampling error).

Publication bias poses a potentially major threat to the validity of the conclusions of a systematic review because, if research is published depending on the statistical significance of the results, then it is likely that published studies will have more positive results (larger effect sizes) than unpublished studies. Therefore, if systematic reviews include only published studies, this will result in an inflated impression of the literature (and in meta-analyses, in an overestimation of population effect sizes). A biased impression of the literature could potentially lead to inappropriate conclusions, with potentially serious practical and policy implications (Lipsey & Wilson 2001; for real-world examples, see Chalmers 2007, Lau et al. 1992). Although awareness of the flaws of NHST is becoming more and more prominent (which may reduce publication bias; see Cumming 2014), the use of NHST will probably endure for the foreseeable future because it is the dominant zeitgeist and because dichotomous decision making based on statistical significance/nonsignificance has an alluring simplicity.

Another issue is that unpublished manuscripts are often thought to be of low quality and are therefore excluded from systematic reviews. Cooper (2003) has thoughtfully discussed this issue and its complexities in depth and warned against making this assumption. For these reasons, it is accepted practice that rigorous research syntheses include both published and unpublished research that meets relevant inclusion criteria (Borenstein et al. 2009, Cooper et al. 2009, Higgins & Green 2011, Lipsey & Wilson 2001).

Locating unpublished work. It can be difficult to locate and obtain unpublished work, which means that reviews are probably always affected to an extent by publication bias (Ferguson & Brannick 2012). We suggest several approaches to locating potentially relevant unpublished work. The most important of these involves contacting researchers with one or more publications on the

topic to ask for forthcoming data and further details of existing data, if needed. Decide in advance how much time to give the authors to reply, balancing a consideration of their busy workload with your own need to progress the review. You might adopt a one-month cutoff, which would involve the following steps: initially contacting authors for information (explaining who you are and why you are contacting them), waiting two weeks, sending a nice follow-up email if you do not receive a reply, waiting another two weeks, then stopping your efforts and recording the outcome. You will need to keep a clear record of correspondence as you will need to acknowledge the scholars who replied in the finished article.

Search for gray literature. Depending on the research question and topic, you may need to search for gray literature, which is any literature produced in electronic or print format that has not been controlled by commercial publishers. Examples include technical or research reports from government agencies, reports and working papers from scientific research groups and committees, doctoral dissertations, conference proceedings, and official publications. Different databases specialize in different types of unpublished work. Examples of gray literature databases include the following:

- OpenGrey (**http://www.opengrey.eu**), a European database compiled by different national libraries in various European countries that submit any gray literature they receive;
- *Open*DOAR (Directory of Open-Access Repositories; **http://www.opendoar.org/**), a website that searches the open-access repositories of thousands of universities all over the world;
- WorldCat, a database for dissertations and theses; and
- Google and Google Scholar, search engines that are reasonably effective in locating dissertations and work by societies and charities.

Screening

Search results need to be screened for potential inclusion.

Export references to a citation manager to collate the search results. Your electronic database searching will almost certainly reveal a large number of results. Exporting search results to a citation manager (e.g., EndNote, RefWorks) confers several advantages: It saves a massive amount of time as this task becomes an electronic rather than a manual process; your search results are saved, meaning that this valuable information cannot be lost; the citation manager can identify and delete duplicate versions of the same work; you can obtain and share full-text versions of many of the identified journal articles; and the citation manager will compile your reference list and format it in an array of referencing styles (which can be very useful if you need to change referencing style as you submit to different journals).

Read the title and/or abstract of identified work. Read the title and/or abstract of all work identified by your searches. Most work identified by your searches will not meet your inclusion criteria. If the title and/or abstract suggest that the work is potentially eligible for inclusion in your review, the next step is to obtain the full-text version and read that carefully.

At this stage, we recommend that you continue to err on the side of sensitivity (locating and sifting as many articles as possible) so that you do not miss anything. For record-keeping purposes during the screening stage, it is sufficient to make a list of the number of rejected articles (rather than noting the reasons for excluding each study).

Eligibility

Read the full text of articles to determine their eligibility for inclusion.

Sift the full-text version of potentially eligible articles. At this stage, your focus will finally shift from sensitivity to specificity. You now need to sift the full-text version of potentially eligible articles to see if each is indeed appropriate for inclusion. Even during this stage, you can rapidly reduce the pool of potential studies by focusing your reading on whether or not each published or unpublished work meets your inclusion and exclusion criteria. This often means, especially when you are conducting a meta-analysis, focusing on the method and results sections rather than the introduction and discussion sections.

Extract all potentially relevant information. Once you are certain that a particular study is to be included, you will need to carefully and thoroughly extract all relevant information. What is considered relevant information will depend on your research questions and topic, on whether you are using quantitative or qualitative information, and on the conventions of the journal you are going to submit to or your university's submission guidelines. The information you extract will predominantly relate to your inclusion criteria and will therefore likely cover definition or conceptualization, measures/key variables, research design, participants, year of publication, data/results, study design, study setting, etc. We encourage extracting all potentially relevant information and tabulating it at this stage. This is because we have found it less time consuming to do this now (and potentially include information that will not be not used in the final review) than to extract specific information further down the line (e.g., in response to a reviewer's comments) at a time when studies are less familiar and you cannot remember your exact decision-making process.

Study Quality

Your inclusion and exclusion criteria are designed to ensure that only relevant work is included in your systematic review. However, that work will of course vary in quality. You may therefore choose to consider and perhaps account for study quality (bias) so as to draw conclusions that closely fit the nature and quality of the available evidence. Certainly, it may be helpful to summarize methodological issues that could limit or bias a literature. If you are conducting a meta-analysis, the results of a quality assessment could be used to inform a sensitivity analysis that tests whether study quality systematically biases effect sizes.

Selecting a tool to assess study quality. A huge range of tools have been proposed for assessing study quality. In fact, a relatively recent review found that there are 86 tools for assessing the quality of nonrandomized studies alone, highlighting the lack of a single candidate tool for assessing the quality of observational epidemiological studies (Sanderson et al. 2007; see Olivo et al. 2008 for a systematic review of scales that assess the quality of randomized controlled trials). Most study quality methods encompass fairly intuitive considerations such as the following: appropriateness of study design and sample size for addressing the research objectives, generalizability (representativeness of the sample), participant or condition selection methods, response and attrition rate, measurement of study variables, control of confounding, appropriateness of statistical analyses, quality of reporting, quality of intervention/condition, and authors' conflict of interest.

Problems with study quality tools. Examining the impact of study quality on the results of a systematic review is neither as simple nor as straightforward as might initially be assumed. Several articles have discussed the limitations of examining study quality and drawn attention to the potential complexities involved (e.g., Garside 2014, Juni et al. 1999). For instance, calculating a summary score may involve assigning weights to the different items that make up a measure of study quality, and it may be difficult to justify the weights assigned. There is also great variation

in what researchers perceive constitutes quality (Valentine & Cooper 2005), and study quality scales have been shown to be unreliable assessments of validity (Juni et al. 1999). It is noteworthy that the use of scales for assessing quality or risk of bias is explicitly discouraged in Cochrane reviews. A study quality tool should not be used to determine inclusion or exclusion from your review. If study quality is somehow to be examined empirically, a considered approach is required. It is probably preferable to look separately at a few workable indicators of bias risk rather than calculating a summary score. These indicators can be used to estimate the degree to which the literature that forms the content of the systematic review may have been affected by bias.

We welcome further research on this issue and rigorous psychometric testing of study quality tools. Although it is often taken as a given that systematic reviews should account for study quality, numerous scholars have drawn attention to the fact that the existing tools and procedures for determining study quality need considerable revision.

HOW TO PRESENT A SYSTEMATIC REVIEW

There is no one right way to present a systematic review. How you organize the review should be logically dictated by the goals you have for the review and seem a clear and readable way to organize things (see Bem 1995, Cooper 2003). However, there are some general principles that are recommended, which we discuss below in relation to the different sections of a systematic review.

Before entering that discussion, we note that, as a general rule, you will have a good chance of publication in a field-leading journal if you do two things. First, you will not go far wrong if you adhere to best practice guidelines for conducting and reporting systematic reviews. Following well-regarded guidelines will ensure that your systematic review is comprehensive and that the conclusions you draw are convincing, robust, and reasonable. Second, we strongly recommend that you consult several systematic reviews published in top journals in your field and use those as templates for formatting and content. Some of the concepts described in best practice guidelines can seem a little abstract: Becoming familiar with how high-quality published systematic reviews are conducted and presented in practice will make the principles and guidelines concrete.

We now provide some principles and suggestions for presenting the different sections of a systematic review. Further guidance on presenting the methods and results of a systematic review can be found in Bem (1995) and Cooper et al. (2009).

Introduction

The introduction to a systematic review should clarify the topic, scope, and rationale for your review. Introductory sections—especially those of very long review articles—often begin with a few catchy paragraphs that succinctly introduce the topic or problem and the central questions that will be addressed through the review (perhaps highlighting one or two major concerns, theoretical issues, or debates in the literature). These are usually followed by a few paragraphs summarizing how the review will be structured and what it will cover, perhaps stating some goals for the review. The introduction section may need to include an explanation of key terms, definitions, and concepts that are essential to understanding the information in the review. Key developments during the history of the literature may also be briefly summarized, if relevant.

The bulk of the introduction will involve presenting a clear and convincing flowing argument for the existence of the review, justifying the type of review conducted (narrative, meta-analysis, etc.) and the benefits of conducting a review on the particular topic area. The rationale for your review may discuss important theoretical and empirical issues and debates in the literature that a

systematic review has the potential to address. In conducting the review, you are aiming to address one or more key issues and to have a substantive impact on the way that readers understand an area and on practice and/or policy. You therefore need to explicitly explain and argue how the review will achieve these aims.

If the review will be used to evaluate the evidence for one or more new or existing theories, these need to be described at this point to orientate the reader. As Baumeister (2013, p. 126) notes, "You cannot expect a reader to keep dozens of research findings straight in memory before finding out how they all fit together." Baumeister & Leary (1997) suggest two potential strategies for introducing readers to a particular theory and its importance and implications: (*a*) fully describe an existing or new theoretical conceptualization in the introduction section and then use the remainder of the manuscript to review the literature as it pertains to the theory, or (*b*) present a brief summary of an existing or new theoretical conceptualization at the start of the review and fully elaborate this after the literature has been reviewed.

Method

The method section needs to clearly describe every step of the methodology used to conduct the review (what you did), along with your reasoning (why). It needs to explain your search in detail, including how and when you searched particular databases, what publication years you searched, what search terms you used, and what inclusion and exclusion criteria you adopted and for what theoretical and/or empirical reasons. It also needs to describe what concerted efforts you made to locate and include all published and unpublished work on the topic and what comprehensive and systematic preventative steps were taken to minimize bias and errors in the study selection process.

Discuss borderline cases. As we discussed earlier, the method section may also need to discuss and explain borderline cases that readers might have expected to have been included, or which were included despite partially breaching the inclusion and exclusion criteria (e.g., see Trickey et al. 2012). Where relevant, you will need to explain and justify how particular borderline cases influenced your inclusion and exclusion criteria.

Present a flow diagram. Best practice for systematic reviews is to present a flow diagram to summarize the literature searching and sifting process (e.g., a PRISMA flow diagram). This will be separated into identification, screening, eligibility, and inclusion stages, and it should provide a succinct summary of the number of studies included and excluded at each stage of the process.

Results

The core of the systematic review entails summarizing and critically evaluating and integrating the results of your comprehensive search strategy using a clear, logical structure. The search results need to be presented in an unbiased, structured, clear, and straightforward way. If the main purpose of the systematic review was to evaluate the evidence for a new or existing theory, it can be useful to organize the research findings accordingly (e.g., in relation to their support for different theories or components of a single theory; see Baumeister 2013). Like flow diagrams, tables are an economical and clear means of summarizing and conveying key results. The characteristics of included studies can be described in detail in a table.

We have found it helpful to plan the structure of a review (e.g., what content will go in what tables) early on in the process using a review protocol; doing so provides a focus and a direction that keep the task tangible and manageable. This approach also helps overcome procrastination

or perceptions of being overwhelmed because a large task can be distinguished into ever smaller sections and subsections. When study results are represented as effect sizes, graphical displays such as a forest plot can be an economical way to display study effect sizes and their uncertainties (in the form of confidence intervals) so that the distribution of estimates can be evaluated. As discussed above, you may choose to describe any assessment of the scientific quality of included studies that has taken place and whether quality ratings were used in any analyses.

Do not simply summarize but offer a new, improved understanding of the phenomena. As we discuss above, a systematic review is about more than simply cataloging what already exists. It is not enough to summarize (describe); you also need to integrate and critique the results of the systematic review. Critical thought and reflection are required; systematic reviews need to interpret the cumulative evidence from individual studies and provide a critical synthesis to advance the field's theoretical understanding of some issue (Baumeister 2013). Baumeister & Leary (1997, p. 317) advise that "literature reviewers should also ask themselves whether they have presented each study in a way that makes its relation to the integrative themes clear and explicit." These things take time and careful thought, and of course more heads than one will introduce new perspectives and observations about the nature of the literature. How does this work in practice? The next two paragraphs provide some pointers.

Zoom out. The nature, consistency, quality, and methodological diversity of the findings should be considered, "keeping in mind that consistency across large quantities of methodologically diverse evidence is the best available substitute for having the proverbial direct pipeline to the truth" (Baumeister & Leary 1997, p. 318). A synthesis can be obtained by zooming out from individual studies to provide a conceptual overview (see Baumeister & Leary 1997), for example by (*a*) linking concepts or grouping findings, (*b*) critiquing, (*c*) drawing attention to methodological or conceptual issues, (*d*) noting variations in results and exploring their possible reasons, and (*e*) assessing the strength of the overall evidence for each main point.

Depending on your aims and the research context, you will need to balance conflict resolution (by identifying inconsistencies in study results) against bridge building (by identifying points of contention in theories, conceptualizations, and methods in the literature; see Cooper 2003). In discussing narrative reviews, Baumeister & Leary (1997) suggest section critiques rather than criticizing each individual study. Each section might involve a summary of the methods and results of a group of studies relevant to a point, along with a brief outline of major flaws of that evidence.

Presenting qualitative findings. If you are presenting a narrative review or meta-synthesis, it may be useful to cite a study's conclusion while also describing the sample, method, and specific findings to provide a context (Baumeister & Leary 1997). This avoids the practice of referencing authors in a manner that does not make clear whether those authors thought or hypothesized something or, in fact, reported data on that topic. For example, one might write "In a sample of A, method B produced result C (Reference), thereby supporting the view that X causes Y" (Baumeister & Leary 1997, p. 317). Adopting this presentation style allows readers to appraise the quality and nature of the evidence.

Counterexamples. Evidence that runs counter to the hypothesis or theoretical model being evaluated can be presented and critiqued in a specific section (Baumeister 2013, Baumeister & Leary 1997). The presence of this evidence is not a weakness of the review but instead points to the complex realities of life:

> If there are important exceptions to the general patterns and conclusions, the literature review is strengthened by acknowledging them, and theory can be built further by recognizing moderators and boundary conditions. If the exceptions are merely apparent and do not on close inspection contradict the main pattern, the manuscript is strengthened by pointing them out. (Baumeister & Leary 1997, p. 319)

Discussion

You will need to summarize and discuss the findings of the review in a balanced and impartial way in the context of previous theory, evidence, and practice. Your conclusions need to explicitly link to the evidence reviewed. Discuss the strengths and limitations of the literature, including a consideration of the scientific quality of the included studies and methodological problems in the literature (e.g., methodological rigor or lack thereof, the amount of evidence, its consistency, and its methodological diversity). Any conclusions and recommendations for practice or policy should be based on the evidence and be tempered by the flaws and weaknesses in the evidence. A good systematic review links the current state of evidence back to theory and may comment on, evaluate, extend, or develop theory (Baumeister 2013). You might propose a new conceptualization or theory to explain inconsistent findings (Baumeister & Leary 1997).

Summarize the literature's progress. The two main purposes of a systematic review are to establish to what extent existing research has progressed toward explaining a problem and to clarify the extent to which a new or existing theory explains the existing evidence. Quantitative or qualitative reviews may conclude that the available evidence suggests one of four possibilities (see Baumeister & Leary 1997 for a detailed discussion): (*a*) A hypothesis is correct, at least based on the present evidence; (*b*) A hypothesis, although not proven, is currently the best guess and should be assumed to be true until a convincing body of contradictory evidence emerges; (*c*) It is not clear whether a hypothesis is true or false; or (*d*) A hypothesis is false. You would then describe directions for future theory, evidence, and practice by pointing out remaining unresolved issues (Baumeister & Leary 1997).

Appendices. Depending on the journal, it is often customary to include appendices to ensure transparency and replicability. You might consider including a detailed reference list of studies that were excluded at the "potentially eligible" (full-text versions of articles) stage. For the purposes of a student thesis, you may want to include sample record-keeping forms and your completed records to show your working, but these are not required for journal submissions.

CONCLUSION

This guide describes how to plan, conduct, organize, and present a systematic review of quantitative (meta-analysis) or qualitative (narrative review, meta-synthesis) information. We argue that conducting a large-scale review has the potential to be satisfying and informative for the researchers involved and good for one's career. High-quality systematic reviews have a competitive chance for consideration in top-tier journals and are relatively likely to have a tangible and substantive impact on policy and practice. We welcome the proliferation of systematic reviews and will be interested to see whether and how they contribute to the replication crisis debate.

DISCLOSURE STATEMENT

A.P.S. and A.M.W. are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review. L.V.H. has authored some of the referenced textbooks on systematic reviews and meta-analyses.

ACKNOWLEDGMENTS

We are grateful to Julian P.T. Higgins and Sofie Kuppens for commenting on earlier versions of this manuscript.

LITERATURE CITED

APA Publ. Commun. Board Work. Group J. Artic. Rep. Stand. 2008. Reporting standards for research in psychology: Why do we need them? What might they be? *Am. Psychol.* 63:848–49

Baumeister RF. 2013. Writing a literature review. In *The Portable Mentor: Expert Guide to a Successful Career in Psychology*, ed. MJ Prinstein, MD Patterson, pp. 119–32. New York: Springer. 2nd ed.

Baumeister RF, Leary MR. 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* 117:497–529

Baumeister RF, Leary MR. 1997. Writing narrative literature reviews. *Rev. Gen. Psychol.* 3:311–20

Bem DJ. 1995. Writing a review article for *Psychological Bulletin*. *Psychol. Bull.* 118:172–77

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2009. *Introduction to Meta-Analysis*. New York: Wiley

Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. 2017. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res. Synth. Methods* 8:5–18

Braver SL, Thoemmes FJ, Rosenthal R. 2014. Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9:333–42

Bushman BJ. 1994. Vote-counting procedures. In *The Handbook of Research Synthesis*, ed. H Cooper, LV Hedges, pp. 193–214. New York: Russell Sage Found.

Cesario J. 2014. Priming, replication, and the hardest science. *Perspect. Psychol. Sci.* 9:40–48

Chalmers I. 2007. The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: the importance of systematic reviews. In *Treating Individuals: From Randomised Trials to Personalised Medicine*, ed. PM Rothwell, pp. 37–58. London: Lancet

Cochrane Collab. 2003. *Glossary*. Rep., Cochrane Collab., London. http://community.cochrane.org/glossary

Cohn LD, Becker BJ. 2003. How meta-analysis increases statistical power. *Psychol. Methods* 8:243–53

Cooper HM. 2003. Editorial. *Psychol. Bull.* 129:3–9

Cooper HM. 2016. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. Thousand Oaks, CA: Sage. 5th ed.

Cooper HM, Hedges LV, Valentine JC. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Found. 2nd ed.

Cumming G. 2014. The new statistics: why and how. *Psychol. Sci.* 25:7–29

Earp BD, Trafimow D. 2015. Replication, falsification, and the crisis of confidence in social psychology. *Front. Psychol.* 6:621

Etz A, Vandekerckhove J. 2016. A Bayesian perspective on the reproducibility project: psychology. *PLOS ONE* 11:e0149794

Ferguson CJ, Brannick MT. 2012. Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychol. Methods* 17:120–28

Fleiss JL, Berlin JA. 2009. Effect sizes for dichotomous data. In *The Handbook of Research Synthesis and Meta-Analysis*, ed. H Cooper, LV Hedges, JC Valentine, pp. 237–53. New York: Russell Sage Found. 2nd ed.

Garside R. 2014. Should we appraise the quality of qualitative research reports for systematic reviews, and if so, how? *Innovation* 27:67–79

Hedges LV, Olkin I. 1980. Vote count methods in research synthesis. *Psychol. Bull.* 88:359–69

Hedges LV, Pigott TD. 2001. The power of statistical tests in meta-analysis. *Psychol. Methods* 6:203–17

Higgins JPT, Green S, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0*. London: Cochrane Collab.

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23:524–32

Presents a thorough and thoughtful guide to conducting narrative reviews.

Presents a comprehensive introduction to meta-analysis.

Presents a comprehensive glossary of terms relevant to systematic reviews.

Presents a comprehensive introduction to research synthesis and meta-analysis.

Discusses the limitations of null hypothesis significance testing and viable alternative approaches.

Presents comprehensive and regularly updated guidelines on systematic reviews.

Juni P, Witschi A, Bloch R, Egger M. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282:1054–60

Klein O, Doyen S, Leys C, Magalhães de Saldanha da Gama PA, Miller S, et al. 2012. Low hopes, high expectations: expectancy effects and the replicability of behavioral experiments. *Perspect. Psychol. Sci.* 7(6):572–84

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.* 327:248–54

Light RJ, Smith PV. 1971. Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard Educ. Rev.* 41:429–71

Comprehensive and clear explanation of meta-analysis.

Lipsey MW, Wilson D. 2001. *Practical Meta-Analysis*. London: Sage

Matt GE, Cook TD. 1994. Threats to the validity of research synthesis. In *The Handbook of Research Synthesis*, ed. H Cooper, LV Hedges, pp. 503–20. New York: Russell Sage Found.

Maxwell SE, Lau MY, Howard GS. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70:487–98

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, et al. 2010. CONSORT explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c869

Comprehensive reporting guidelines for systematic reviews.

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339:332–36

Nelson LD, Simmons J, Simonsohn U. 2018. Psychology's renaissance. *Annu. Rev. Psychol.* 69:511–34

Noblit GW, Hare RD. 1988. *Meta-Ethnography: Synthesizing Qualitative Studies*. Newbury Park, CA: Sage

Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. 2008. Scales to assess the quality of randomized controlled trials: a systematic review. *Phys. Ther.* 88:156–75

Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349:943

Paterson BL, Thorne SE, Canam C, Jillings C. 2001. *Meta-Study of Qualitative Health Research: A Practical Guide to Meta-Analysis and Meta-Synthesis*. Thousand Oaks, CA: Sage

Patil P, Peng RD, Leek JT. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11:539–44

Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. *Psychol. Bull.* 86:638–41

Rosnow RL, Rosenthal R. 1989. Statistical procedures and the justification of knowledge in psychological science. *Am. Psychol.* 44:1276–84

Sanderson S, Tatt ID, Higgins JP. 2007. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int. J. Epidemiol.* 36:666–76

Schreiber R, Crooks D, Stern PN. 1997. Qualitative meta-analysis. In *Completing a Qualitative Project: Details and Dialogue*, ed. JM Morse, pp. 311–26. Thousand Oaks, CA: Sage

Shrout PE, Rodgers JL. 2018. Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* 69:487–510

Stroebe W, Strack F. 2014. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9:59–71

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, et al. 2000. Meta-analysis of observational studies in epidemiology (MOOSE): a proposal for reporting. *JAMA* 283:2008–12

Thorne S, Jensen L, Kearney MH, Noblit G, Sandelowski M. 2004. Qualitative meta-synthesis: reflections on methodological orientation and ideological agenda. *Qual. Health Res.* 14:1342–65

Tong A, Flemming K, McInnes E, Oliver S, Craig J. 2012. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Med. Res. Methodol.* 12:181–88

Trickey D, Siddaway AP, Meiser-Stedman R, Serpell L, Field AP. 2012. A meta-analysis of risk factors for post-traumatic stress disorder in children and adolescents. *Clin. Psychol. Rev.* 32:122–38

Valentine JC, Biglan A, Boruch RF, Castro FG, Collins LM, et al. 2011. Replication in prevention science. *Prev. Sci.* 12:103–17

Valentine JC, Cooper H. 2005. Can we measure the quality of causal research in education? In *Experimental Methods for Educational Interventions: Prospects, Pitfalls and Perspectives*, ed. GD Phye, DH Robinson, J Levin, pp. 85–112. San Diego, CA: Elsevier