Methods reporting that supports reader confidence for systematic reviews in psychology: assessing the reproducibility of electronic searches and first-level screening decisions

Paul Fehrmann[1] and Megan Mamolen[2]

Abstract

Recent discussions and research in psychology show a significant emphasis on reproducibility. Concerns for reproducibility pertain to methods as well as results. We evaluated the reporting of the electronic search methods used for systematic reviews (SR) published in psychology. Such reports are key for determining the reproducibility of electronic searches. The use of SR has been increasing in psychology, and we report on the status of reporting of electronic searches in recent SR in psychology. In all, we used 12 checklist items to evaluate reporting for electronic strategies. Kappa results for most of the items developed from evidence-based recommendations, ranged from fair to almost perfect. Data for a stringent 'PRISMA' type of recommended reporting showed that only one of the 25 randomly selected psychology SR from 2009-2012 reported recommended information for all items in the set, and none of the 25 psychology SR from 2014-2016 did so. Results for a second less stringent set found that only 36% of the psychology SR reported basic information that supports confidence in the reproducibility of electronic searches. Using those two sets of checklist items found similar results for psychology SR published in 2017. Moreover, reporting was also very infrequent for a third supplemental set of 'confidence items'. Fuller and clearer recommended reporting of the electronic searches used in SR would provide a stronger basis for confidence in the reproducibility of searches. That reporting, in turn, would strengthen reader confidence more generally in the results and conclusions reached in SR in psychology.

Keywords

Systematic reviews, psychology, reproducibility, electronic searches, reporting, confidence

1. Introduction

1.1 Background

Recent discussions have shown a significant emphasis on the reproducibility of research in psychology (Pashler and Wagenmakers, 2012; Yong, 2013; Cooper and VandenBos, 2013; Novotney, 2014; Open Science Collaboration, 2015; Gilmore, Diaz, Wyble, & Yarkoni, 2017). Systematic reviews are viewed by many as a top level synthesis of research (Paul & Leibovici, 2014; Kisely et al., 2015; Ng & Benedetto, 2016), and a special emphasis on reproducibility and systematic reviews is evident in the recent *Psychological Bulletin* focus on the topic 'Replication and Reproducibility: Questions Asked and Answered via Research Synthesis'. Concerns for reproducibility pertain to methods as well

 Fehrmann, Paul & Mamolen, Megan (2020) Methods reporting that supports reader confidence for systematic reviews in psychology: assessing the reproducibility of electronic searches and first-level screening decisions, IASSIST Quarterly 44(1-2), pp. 1-26. DOI: https://doi.org/10.29173/iq968

as results, and the current paper contributes to discussions about the methods that are used for systematic reviews.

1.2 Rationale for our reproducibility research

Briefly stated, our research has been motivated by the following. First, in general, confidence in the methods used for systematic reviews provides a basis for confidence in the conclusions reached in systematic reviews. Second, and more specifically, electronic searches provide a crucial part of the methods used to identify data and build the evidence base that is then used for conclusions reached in systematic reviews. Third, the reproducibility of electronic searches has been a core recommendation in guidance on systematic review methods. Fourth, full and transparent reporting of electronic search methods is key for reader confidence in the reproducibility of electronic searches. Fifth, important reporting details also can support reader confidence in the reproducibility of first level screening of electronic searches. And, sixth, confidence in the reproducibility of searching and screening then supports the confidence that readers can have in the conclusions that are reached in systematic reviews (Cooper, 2017; Vazire, 2017: Golder et al., 2013; Pashler & Wagenmakers, 2012).

Meta-research has recently cautioned that '…even when research findings are reported, they can be undermined by a lack of transparency about how they were generated (Hardwicke et al., 2020), and the target of our study has been what is described by Goodman and others as 'methods reproducibility'. They argue that methods reproducibility '…refers to the provision of enough detail about study procedures and data so the same procedures could, in theory or in actuality, be exactly repeated (Goodman, Fanelli, & Ioannidis, 2016, p. 2). In response, to develop a picture of the basis readers have for confidence in the results of systematic reviews in psychology, we have been evaluating the reporting of electronic searches used for systematic reviews in psychology.

1.3 Systematic reviews, reproducibility, electronic searches, data management, reporting, and psychology

Systematic reviews (SR) have been widely recognized both as methods and as products that involve rigorous, transparent processes for identifying, analyzing, and synthesizing data from primary studies to draw conclusions relevant to a research topic (Cooper, 2017, p. 10; Gough & Oliver, 2012). SR are also given special recognition in a recent National Academy of Sciences (NAS) report 'Reproducibility and Replicability in Science' (Committee on Reproducibility and Replicability in Science et al., 2019), particularly in a chapter entitled 'Confidence' which has this conclusion "Multiple channels of evidence from a variety of studies provide a robust means for gaining confidence in scientific knowledge over time' (Committee on Reproducibility and Replicability in Science et al., 2019, p. 155).

Importantly, in that NAS chapter, SR are noted as a significant for gaining confidence in science. SR by design are 'the ensemble of research activities involved in identifying, retrieving, evaluating,

synthesizing, interpreting, and contextualizing the available evidence from studies on a particular topic'; and as such they address '…the central question of how the results of studies relate to each other, what factors may be contributing to variability across studies, and how study results coalesce or not in developing the knowledge network for a particular science domain' (Committee on Reproducibility and Replicability in Science et al., 2019, p. 144). Mirroring the NAS discussion, a recent introduction from the Joanna Briggs Association explains that SR are specifically designed to achieve results and reach conclusions based on analyzing and synthesizing "all" of the evidence or data relevant to a question (Aromataris & Munn, 2019).

Of course an important methological question for readers of SR is 'how did you get that data ?'. SR typically use a range of strategies to *find resources that have the data* that is extracted, analyzed and synthesized, including use of databases, grey literature, scanning reference lists of key articles, hand searching of journals, web sites, and contacting experts (Kugley et al., 2017). However, although different strategies are used for the comprehensive searching that is often stressed as a characteristic of SR, electronic searches have been noted as providing '…the largest portion of the evidence base for systematic reviews' (Sampson et al., 2009, p. 944), and a standard expectation is that the electronic searches used will be reproducible (Centre for Reviews and Dissemination, 2009; Liberati et al., 2009; Lefebvre et al., 2019; Agency for Healthcare Research and Quality, 2014; Kugley et al., 2017).

A follow up question for readers of SR is how to determine if electronic searches used in SR are reproducible. Readers might use what is reported about those searches in an attempt to rerun the searches (e.g., see Ali & Usman, 2018). Or, researchers might give their electronic search strategy report to a peer to have them execute the search, and then report the peer review results for readers (McGowan et al., 2016; Faggion, 2019). Of course, most readers of SR, including clinicians and policy makers, do not have the time or resources to attempt a rerun. Although an alternative is to see if information recommended for confidence in the possibility of reproducing the electronic search is actually reported, research indicates that greater transparency is needed to support that confidence (Campbell et al., 2019). As stressed recently, without 'clear signals' of *practices* that increase the 'trustworthiness of scholarly work', readers are challenged when 'ascertaining confidence' they might have in that scholarly work. A recommendation is that such signals would include *information that confirms* 'adherance to field-specific reporting requirements' (Jamieson, McNutt, Kiermer, & Sever, 2019).

Internationally accepted guidance on reporting SR search steps has been available for many years in the PRISMA Statement. As the authors of that guidance stated "…systematic reviews should be reported fully and transparently to allow readers to assess the strengths and weaknesses of the investigation" (Liberati et al., 2009)[3]. Importantly, that guidance includes descriptions and examples of what should be reported of SR electronic search methods, and the current study drew on that methods guidance to address our research objectives.

With respect to the field of psychology, as indicated in Table 1., there is clear evidence of growing use of SR.

Table 1. Search showing increase use of systematic reviews in psychology	
Date	October 8, 2018
Resource	PsycINFO (EBSCO)
Interface	Advanced Search
Search terms	None
EBSCO Search Limiters	• English language • Journal Article as Document Type • Systematic review as Methodology
Publication Year	Number of search results
• 2000 • 2005 • 2010 • 2015 • 2016	• 19 • 213 • 1006 • 2464 • 2306

Even though recent work has reiterated that electronic searches are insufficient for confidence in the comprehensiveness of searches (Delaney & Tamás, 2018), readers of SR still do have a key support for confidence in SR results if they see reporting that helps to ensure the reproducibility of electronic searches used. Actually, determining reproducibility is dependent on what is found in this report (Niederstadt & Droste, 2010; Rader et al., 2014; Mullins, 2014; Atkinson et al., 2015; Schalken & Rietbergen, 2017), and in our study of methods we evaluate the reporting of electronic search methods that are used for SR in psychology. To date we have not found this kind of assessment of SR methods in psychology[4].

1.4 Definitions and abbreviations

Other recent work has pointed to variation in how researchers define and assess 'reproducible searches' (Sayre & Riegelman, 2018; Koffel & Rethlefsen, 2016; Ali & Usman, 2018). Table 2 contains stipulated definitions for this paper, including definitions for the reproducibility of electronic searches and for confidence in the reproducibility of electronic searches. Additionally, in this paper we use the abbreviations that are indicated. For example, 'search' means electronic search.

Table 2. Definitions of terms related to the reproducibility of electronic searches	
Electronic resource	An electronic search engine, database, or search function that a researcher uses to discover sources that contain substantive information or data to be synthesized in systematic reviews (SR). Typical SR use more than one electronic resource.
Electronic search (search)	A researcher's choice of an electronic resource used for the SR, and the researcher's complete use of that resource (steps taken, terms input, etc.) to the point of seeing electronic search results to evaluate. Those choices and steps are often called electronic search elements (e.g., Maggio et al., 2011; Rader et al., 2012). Typical SR will have more than one electronic search.
Electronic search result (hits)	A listing of 'hits'. The number of items (hits) that correspond to the search terms entered is usually shown on a search screen after the electronic search is run. Each electronic search hit is brief information that represents a corresponding document that has the fuller content (and data) potentially of use for the SR. Each individual hit (or 'record') typically includes key information (e.g., title, author, journal name, etc.) for related full documents. Individual hits also can include abstracts or brief contextual information that gives additional insight or clarification on what is contained in the full document.
Electronic search report (report)	An electronic search report is defined as a researcher's published description of electronic search steps and choices (including the researcher's use of search results or 'hits' for inclusion/exclusion decisions). Such reports are seen as detailing the actions taken to discover and select items with data to be used for review purposes. Steps, that is, up to and including what is often called a first screening of 'titles and abstracts'.
Recommended report element	An electronic search choice or characteristic that should be used and then reported or evident as used in SR. These may be found in major SR manuals and publications.
Electronic search report item (item)	An explicitly worded statement or question that represents or addresses all or part of a recommended electronic search element. These may be found on checklists.
A reproduced electronic search	This is where a second researcher reads a report published by a first researcher, and the second researcher uses that report to develop and execute an electronic search that uses the same electronic resource(s), with the same steps, as those used by the first researcher.
Reproduced search results	This is seeing the same number of hits for a search that is reproduced.
Reproducibility of an electronic search (reproducibility)	The extent to which a second researcher might execute a reproduced electronic search. Presumably the reproducibility for a given SR is the extent to which reproduced electronic searches might be executed for every electronic resource used in that SR.
Confidence in search reproducibility	This is taken to be a reader's sense that their judgement on search reproducibility is well supported by the search report.

In addition to the stated definitions for the terms 'electronic resource' and 'electronic search', for this paper these terms refer to a range of approaches for finding the information sources that are eventually selected for use in SR.Electronic searches can involve using discipline-specific commercial resources (e.g., EBSCO's PsycINFO), public resources (e.g., PubMed), web search engines (e.g. Google Scholar), systematic review library databases (e.g., Campbell Library; Cochrane Library), electronic grey literature resources (e.g., ProQuest Dissertations and Theses), scholarly 'cited reference' databases (e.g., Scopus, Web of Science), and a range of other possibilities. Also, in our definitions for Electronic Resource, Electronic search result, and Electronic search report we refer to data. The extracting or selection of data from resources found using electronic searches is basic to SR, and in our definitions the term 'data' refers to the quantitative or qualitative information selected and extracted from multiple sources and then analyzed and synthesized to address research questions in SR.

1.5 Research objectives

1. We sought to compare reporting of electronic searches in psychology SR to the reporting of electronic searches in SR that are known to be completed with rigorous methods for reporting.

2. Most if not all research on reporting of SR has looked at reporting of individual items (individual steps) used for electronic searches. Assuming that higher levels of reproducibility are supported as more of a set of search steps are reported as recommended, we looked to document the extent that psychology SR provide reporting of electronic searches according to a set of widely accepted recommendations for what should be reported. As detailed in our Methods section, we called this a PRISMA set. Data and discussion our assessment of a "non-PRISMA" set are available in Supplementary files on this paper's OSF site[5] (hereinafter 'the OSF site').

3. There is information that is *not needed* to execute and see results for what we defined as a *reproduced electronic search,* but which nevertheless *can support reader confidence* in the reproducibility of searches used. We looked to document reporting for this kind of information in Campbell and psychology SR.

2. Methods

2.1 Checklist items used for this study

AMSTAR, PRISMA, and PRESS are three major resources providing guidelines to support the design, execution, and evaluation of SR, and we discuss those resources and the evaluation of reproducibility in Supplementary files on the OSF site. Other authors have also created checklists or reporting guidelines relevant to evaluating electronic searches (e.g., Booth, 2006; Yoshii et al., 2009; Maggio et al., 2011; Atkinson et al., 2015). Editors have also been urged to use checklist items to check reporting in order to 'protect the reporting process' and 'to signal the trustworthiness of science' (Jamieson, McNutt, Kiermer, & Sever, 2019).

For this study we used twelve items drawn from a set of 36 checklist items created during a search evaluation project pursued over a number of years. These items focusing on reproducibility were based on evidence based recommendations in publications and manuals of major SR organizations (e.g., Centre for Reviews and Dissemination, 2009; Liberati et al., 2009; Kugley et al., 2017; Lefebvre et al., 2011; Agency for Healthcare Research and Quality, 2014). Additional explanation for the eight basic and four additional confidence checklist items used is found in sections 2.3 and 2.4 below.[6]

2.2 Identification and selection of SR for this study

PsycARTICLES and the Campbell Library were used to identify SR for this study. PsycARTICLES is a resource for identifying 'peer-reviewed publications of the American Psychological Association (APA) and affiliated journals' that cover 'the science of psychology and behavior' (PsycARTICLES, n.d.). The Campbell Library is also a resource that consists of peer-reviewed SR publications. We used the reporting for Campbell Collaboration SR (Campbell SR) as a model or standard of comparison for assessing the psychology SR. Similar comparisons have been reported in the health sciences (Sampson et al., 2008; Yoshii et al., 2009; Popovich et al., 2012; Golder et al., 2013). The searches used with PsycARTICLES and with the Campbell Library are presented in Table 3.

Table 3. Electronic searches used with PsycARTICLES and the Campbell Library.	
Resource:	PsycARTICLES
Service provider:	American Psychological Association
Access:	Internet
Interface:	PsycNET
Search option:	Advanced
Date searched:	December 12, 2012.
Date range:	2009 through 2012
Limits:	Language = English; Methodology = Systematic Review
Results ('hits'):	29
Resource:	Campbell Library
Service provider:	Campbell Collaboration
Access:	Internet
Interface:	Campbell Library
Search option:	Advanced
Date searched:	April 14, 2013
Date range:	2009 through 2012
Limits:	Type of document: Review; Coordinating groups: Crime and Justice, Education, International Development, Social Welfare
Results ('hits'):	42
Resource:	PsycARTICLES
Service provider:	American Psychological Association
Access:	Internet
Interface:	PsycNET
Search option:	Advanced
Date searched:	April 3, 2016
Date range:	2014 through 2016 (on date of search)
Limits:	Language = English; Methodology = Systematic Review
Results ('hits'):	69
Resource:	PsycARTICLES
Service provider:	American Psychological Association
Access:	Internet
Interface:	PsycNET
Search option:	Advanced
Date searched:	November 11, 2017
Date range:	2017 (on date of search)
Limits:	Language = English; Methodology = Systematic Review
Results ('hits'):	45

The 29 hits retrieved in PsycARTICLES in December, 2012 were SR as defined by APA (APA Databases Methodology Field Values, n.d.), and for this study 25 articles were randomly selected to represent SR in psychology for the timeframe of 2009-2012. Each of those 25 SR contained electronic searches. The 29 hits from PsycARTICLES are listed in Appendix 1, with 25 included for this study marked with asterisks[7]. The 69 hits noted in the updated PsycARTICLES search in April 2016 were assumed to be

SR in psychology. This search was completed by the first author who examined all 69 to see if an electronic search was reported. Ten of the set of 69 articles did not report electronic searches, and a random sample of 25 was selected from the remaining 59 to represent SR in psychology for the timeframe of 2014-2016 (as of date of search). Appendix 3 lists all 69 of the second set of SR from PsycARTICLES; and there we indicate both the ten which did not provide electronic search reports as well as the 25 randomly selected SR that we used. Similarly, 25 Campbell Library articles were randomly selected from the 42 initial search results. The 42 Campbell Collaboration results are in Appendix 2, with 25 SR marked that were included for this study. In the remainder of this paper the Campbell Library systematic reviews may be referred to as 'Campbell SR' and the phrase 'Psych SR' may be used to refer to the PsycARTICLES assessed.

A search update was used in 2017 to collect data for the second research objective; and Appendix 4 shows the 45 SR articles identified along with indication of the 18 SR that were assessed for the current study. The 18 SR assessed were those papers that explicitly discussed PRISMA by name or cited and listed PRISMA in their references. Like other recent studies (Leclercq et al., 2019; Page et al., 2020), we took that discussion or citing of PRISMA to indicate that the authors of those SR had seen PRISMA as a guide for their reporting of electronic searches used.

2.2. Data collection

For the Psych SR, reports were in the article's method section, in appendixes, or in supplemental files. For the Campbell SR, the search reports were located either in a section entitled 'Search methods for Identification of Studies' or in an Appendix.

We used Qualtrics (Qualtrics XM - Experience Management Software, n.d.) to create a data entry tool for checklist scores for the 93 SR assessed. Before collecting data for this study, the authors pilot tested the 36 checklist items mentioned above using SR from psychology journals and SR from the Campbell Library. The SR used for pilot testing were not used as sources of data for addressing our research topics. Prior to reaching consensus scores, search reports for the sets of SR published from 2009-2016 were evaluated and scored independently by the authors.

Checklist items ask about electronic search elements that are documented in electronic search reports. Items are scored yes (Y), provisional yes (PS), not sure (NS), no (N), or NA. Y for an item means that the evaluator believes that what is noted in that item is clearly reported. N means the opposite of Y. PS means the evaluator feels confident they can guess what was done, and NS means the opposite of PS. NA for an item means that the evaluator believes that what is noted in that item is not applicable.

Figure 1 shows a copy of the first checklist item as it appeared on the data entry tool.

Figure 1. Sample item from survey tool used to assess the reproducibility of electronic/computer searches

1. Copy. Does the report state or show that it is providing a copy of the computer search strategy as a saved/downloaded/copy-pasted "search history" from the actual computer search session ? Discussion: Ideally, this format and detail allows a reader to reproduce the computer search activities to the point of seeing the same "hits" that the original researcher saw. That point is where the researcher can make their initial IE decisions (title/abstracts). It is also possibly just before combining the hits/results from a number of computer search resources used prior to removing duplicates. If this kind of information is online as a supplementary file, then the file access information should be in the article.

	Yes	Pretty Sure	Not Sure	No	NA
a. For any Resource	○	○	○	○	○
b. For every Resource	○	○	○	○	○

Comments on the above item.

As shown, two levels of assessment (a. and b.) were used for that search report item. A comment box allowed for qualitative observations used for our discussions as we reached consensus-scoring decisions for items.

2.3 Data analysis

We used and report below on three approaches to assess the electronic search reports in SR drawn from the psychology literature. Information and data for one additional approach is on the OSF site.

Campbell Collaboration Comparison. First, we evaluated Campbell SR to determine report frequencies for eight basic, individual search elements. SR in the Campbell Library, like those completed under guidance of the Cochrane Collaboration, are completed using guidance that increases the possibilities for higher quality reporting of searches (Kugley et al., 2017). Moreover, because the Cochrane Collaboration has been a leader for guidance and high quality SR, studies in the health sciences have compared the quality of reporting in Cochrane SR to that found in non-Cochrane SR (Moher et al., 2007; Sampson et al., 2008; Popovich et al., 2012). Mirroring the health studies research, then, we compared reports in SR published in the psychology literature to those SR from the Campbell Library. We used Fisher Exact Tests to evaluate the significance of difference in proportions between the report frequencies for the Psych and Campbell SR.

PRISMA set. As a key approach to assessing Psych SR for 2009-2012, 2014-2016, and 2017, a set of basic search element reporting recommendations that are found in the PRISMA statement were used. The PRISMA guidance urges the reporting of the 'full electronic search strategy' for at least one electronic resource used (Liberati et al., 2009). Above we noted our evaluation of the frequencies of reporting for *individual recommended search elements* related to reproducibility. However, when evaluating a specific SR paper, readers are likely to be concerned with reporting for *sets* of search elements relevant to that SR. To that end, we used the search elements listed below to see the extent to which a PRISMA type of full electronic search strategy has been reported in psychology SR. We also evaluated the Campbell SR. To date we have not found this kind of assessment of SR in psychology.

Below we briefy explain our choice of search elements. We then explain how we used those elements.

Our selection of four elements can be viewed as representing what is recommended in the PRISMA statement as a report of the 'full electronic search strategy' (Liberati et al., 2009). These elements also reflect relevent items in the American Psychological Association's 'Meta-Analysis Reporting Standards' (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008).[8] Moreover, this set corresponds to an overlap of elements provided in two papers which used extensive methods for identifying key elements, including consultation with groups of searching experts. See Table 3 in Mullins et al. (2014), and Table 1 in Rader et al. (2014). This list also overlaps with search reproducibility elements noted in other papers (Sampson, et al., 2008; Yoshii et al., 2009; Atkinson et al., 2015; Meert et al., 2017; Koffel & Rethlefsen, 2016), and corresponds to recommendations in the Campbell Collaboration guidelines (Kugley et al., 2017).

The 1-4 order of the list below tracks the order of percentages previously reported for the number of SR reporting those information elements (Mullins et al., 2014, see their Table 5). Those percentages are listed here in parentheses.

1. The names of all electronic resources used (94% of 102 SR reporting).
2. The publication time frames of articles to be included (34% reporting**).**
3. Copies of search strategies for any electronic resource used (13% reporting for at least one or all databases used).
4. The vendor names of any electronic resource used (8% reporting).

We assumed that higher levels of reproducibility are supported as more of a set of recommended reporting criteria are met, and we evaluated electronic searches using the following sequence. First, in SPSS we identified those SR which were given a consensus Yes score for item 1 just above. We then identified those SR that were given a consensus Yes score for both items 1 and item 2. Next we identified sets with positive scores for 1, 2, and 3. Finally, we identified a set with positive scores for 1-4. By using the basic percentage order indicated to sequence our analysis we looked to prevent premature elimination of those SR which had not reported the names of vendors but which had provided copies of search strategies of every database used.

Confidence items. During our current research project, we developed and used four items that go beyond the basic report information typically required by readers if they want to actually reproduce an electronic search. That is, as one reads and evaluates a systematic review (SR), if rerunning an electronic search is not feasible then these are supplemental report elements that may enhance confidence in search reproducibility and confidence in the potential use of searches.

Two supplemental search report elements involve reporting the final number of hits for *each* database search, and reporting the *final number of hits combined across the databases* used. This kind of reporting can be found in SR that use a PRISMA type flowchart (see the PRISMA Statement at prisma-statement.org; also see the 'Flow Diagram' in Gensby et al., 2012).

A third confidence element involves reporting the use of two researchers for inclusion/exclusion decisions when viewing the title and abstracts (McDonagh et al., 2008) and also reporting inter-rater agreement results (e.g., Kappa) for those inclusion/exclusion decisions (Liberati et al, 2009). As indications of potentially reduced selection errors, good reporting for either of these would support reader confidence in the possibility of reproducing the electronic search first level inclusion choices. That is, there could be support for increased confidence that the original use of electronic search results ('hits') could be reproduced.

A fourth confidence element would be reporting information that *identifies hits* actually chosen for potential inclusion as a result of evaluating the title and abstract. This is prior to evaluating the full text. If a reader reproduced an electronic search process and saw that the items they choose while screening the titles and abstracts match those identified by the original researcher, then that reader can have increased confidence that they are mirroring the original researcher's use of inclusion criteria with the electronic search hits. Actually, without rerunning a search, just seeing this information reported would support reader confidence in the possibility of reproducing the selection process as well as the searches used.

To briefly summarize, this kind of supplemental information in search reports can support the assumption of reproducibility for the search and selection process used for SR. As this paper was prepared, except for an increase in use of PRISMA type flowcharts showing the number of hits for searches, the *reporting* of information for these elements has been very infrequent. Additionally, with a recent exception (Schalken & Rietbergen, 2017), looking for reporting of this information also does not seem to be a part of *assessing reports* of searches (e.g., Atkinson et al., 2015; Booth, 2006; Golder et al., 2008; Golder et al., 2013; Maggio et al., 2011; Mullins et al., 2013; Niederstadt and Droste, 2011; Tunis et al., 2013; Yoshii et al., 2009; Koffel and Rethlefsen, 2016; Meert et al., 2016). As a contribution to discussions of the reproducibility of searches and confidence in the use of search results, we looked at reporting for these four supplemental elements across our set of 93 SR from Campbell and PsycARTICLES.

3. Results and discussion

3.1 Inter-rater agreement and potential use for basic checklist items

The disccussion and table below provide a look at data for the eight basic checklist items we developed as well as for our use of those items to address our first two research questions. Section 3.4 presents similar discussion for our four confidence items.

Similar to the item assessment reported for the development of AMSTAR (Shea et al., 2009), we looked at inter-rater agreement for individual checklist items. Some important studies that have evaluated reports of search strategies have not included inter-rater agreement results for the elements or items used to evaluate search reporting (Shea et al., 2002; Sampson et al., 2006; Moher et al., 2007; Yoshii et al., 2009; Golder et al., 2013; Mullins et al., 2014). Other recent related papers have provided this kind of checklist information (Willis and Quigley, 2011; Fehrmann and Thomas, 2011; Popovich et al., 2012; Pieper et al., 2015; Meert et al., 2016). That said, the current paper is the first to report this kind of reliability information for individual items on a checklist specifically

developed and worded to assess the reporting that supports the reproducibility of electronic searches.

Kappa coefficients were used to assess inter-rater agreement, and, following Landis and Koch (1977), those coefficients were interpreted with these categories: below chance considered poor; 0.01 to 0.20 slight agreement; 0.21 to 0.40 fair agreement; 0.41 to 0.60 moderate agreement; 0.61 to 0.80 substantial agreement; and 0.81 to 1 almost perfect agreement. The asterisks *** in the Table 4 indicate those items where we scored all SR articles as 'No' or 'Yes' and so Kappa could not be calculated. The asterisks indicate 100 percent agreement.

Table 4. Kappa and frequency results for individual basic recommended electronic search report elements related to search reproducibility

Search report items/sub-questions	Kappa (sig) Psych SR 2009-2012	Frequency Psych SR 2009-2012	Kappa (sig) Campbell SR 2009-2012	Frequency Campbell SR 2009-2012	Fisher's Exact	Kappa (sig) Psych SR 2014-2016	Frequency[1] Psych SR 2014-2016	Frequency[2] Psych SR 2017
1. Does the report state or show that it is providing a copy of the electronic search strategy as a saved/downloaded/copy-pasted 'search history' from the actual electronic search session ?								
a. For any resource	1.00 (.000)	1/25	.92 (.000)	14/25	.000	.359 (.019)	2/25	4/18
b. For every resource	1.00 (.000)	1/25	.76 (.000)	13/25	.000	***	0/25	2/18
2. Are names given for the electronic search resources used?								
a. For any resource	1.00 (.000)	24/25	***	25/25	1.000	***	25/25	18/18
b. For every resource	.65 (.001)	24/25	***	25/25	1.000	.359 (019)	24/25	17/18
3. A number of widely used electronic resources are provided by a number of vendors. Is a vendor, host, or publisher name listed?								
a. For any resource	.65 (.001)	2/25	.92 (.000)	10/25	.018	468 (.006)	3/25	10/18
b. For every resource	***	2/25	.48 (.005)	6/25	.247	***	2/25	7/18
4. Is a publication time frame listed?								
a. For any resource	.39 (.014)	20/25	.31 (.070)	21/25	1.000	.662 (.000)	14/25	16/18
b. For every resource	.39 (.014)	19/25	.05 (.656)	20/25	1.000	.662 (.000)	14/25	16/18
5. Is a date given for when the electronic search was last run?								
a. For any resource	.47 (.006)	3/25	.73 (.000)	10/25	.051	.754 (.000)	9/25	14/18
b. For every resource	.47 (.006)	3/25	.69 (.000)	7/25	.289	.754 (.000)	9/25	14/18
6. Does the report state (or show or make obvious) that all of the exact search terms (words or phrases) that were used are actually listed?								
a. For any resource	.34 (.065)	22/25	-.04 (.835)	24/25	.609	.603 (.001)	21/25	13/18
b. For every resource	.28 (.112)	22/25	-.07 (.656)	22/25	1.000	.429 (.022)	19/25	15/18
7. Does the report explicitly state or show that it is listing how all of the search terms used were combined?								
a. For any resource	.23 (.075)	18/25	.71 (.000)	22/25	.289	.762 (.000)	13/25	14/18
b. For every resource	.23 (.075)	18/25	.35 (.075)	20/25	.742	.606 (.002)	11/25	14/18
8. Are all search set combinations listed?								
a. For any resource	.18 (.114)	18/25	.69 (.001)	20/25	.742	.595 (.002)	8/25	12/18
b. For every resource	.18 (.114)	18/25	.40 (.026)	17/25	1.000	.415 (.032)	6/25	12/18

Note: 1. On date of search, April 3, 2016. 2. On date of search, November 11, 2017. Assessment data is for a select set of 18 items that had explicit reference to the PRISMA statement as explained in this paper.

As shown in Table 4, our testing with eight basic multi-level items showed them overall to have 'fair' to 'almost perfect' inter-rater scores (Kappa). Our findings do indicate challenges (e.g., item 6 Kappa results for the Campbell SR), and more training with our items might have improved our Kappa results. Others found repeated refining of and training with their 'keyword' item improved the Kappa to .64 for an item similar to item 6 in our Table 4 (Meert et al., 2016). Similarly, others reported a Kappa 'range' of .66 – 1 (Willis and Quigley, 2011), using items from the PRISMA Statement (Liberati et al., 2009), indicating that the PRISMA items relevant to reproducibility (6, 7, 8) had Kappas at or above the lower end of that range.

It does seem, then, that the current set of checklist items could usefully serve others as part of an approach to evaluating search reports to determine reproducibility. The use of items specifically worded to assess search report elements provides a consistent framework for assessing searches. This is similar to the specific items found in AMSTAR[9] and the PRESS tool[10]. That said, just as others have invited researchers to use their checklists (Shea et al., 2009; Tong et al., 2012; Meader et al, 2014), additional work with the checklist items we used could have value for research for assessing electronic searches, and for determining and possibly extending the value of these items. Also promising for checklists that researchers, readers, and editors can use is the related work underway both on the overall PRISMA update ("PRISMA," n.d.; Page et al., 2020), and on the PRISMA-S[11] extension that focuses more generally on the fuller set of searches done for SR (including electronic searches).

3.2 The status of recommended search reporting in psychology SR published from 2009-2017

Table 4 presents data on the reporting in psychology SR for widely recommended search elements, and here we focus on results for our checklist item 1. More detailed comments on the other checklist items are available on the OSF site.

It appears unanimous in SR guidelines that a copy of the full electronic search is desirable and even expected for at least one of the electronic search sources used (Liberati, et al., 2009; Higgins and Green, 2011; Kugley et al., 2017), and item 1. was used to assess the reporting or provision of what we call copies of the electronic search strategy. For this study we looked for the kind of copy that is possible using a search resource function for saving or printing a search history, or for a copy and pasted representation of the search steps. A typical printout may well have the terms used (free text and thesaurus terms), and show how terms were combined, the sequence of entering terms and their combination, the use of adjacency, and the use of truncation, along with other details. This checklist item is different from our other checklist items, because such copies can potentially include information for many of those other items. In fact, depending on the resource used, and the search run, either now or in the future such copies might include all of the information indicated in the other checklist items that would be relevant for providing a 'full electronic search strategy' for that resource. Providing such copies could be a straightforward way to efficiently and accurately show most if not all of the steps actually used with a given resource.

Our results showed that 14 of the 25 Campbell SR provided such a copy *for at least one electronic source* used (see item 1.a., column 5). These articles are identified in Appendix 2. Only one of our set of 25 Psych SR from 2009-2012 provided a copy for at least one electronic source used (see item 1.a., column 3, and article 24 in Appendix 1). Two of our Psych SR sample for 2014-2016 provided this in their reports, and these are identified in Appendix 3. Four of our select set of 18 Psych SR for 2017 provided copies for at least one resource used, and are indicated in Appendix 4.

Frequency results for copies of search strategies for *every electronic source used* (item 1.b) were similar (Psych SR, 1/25; Campbell, 13/25).[12] We did not find any in our 2016 Psych SR sample that

reported copies for all searches used, and 2 out of 18 Psych SR assessed in 2017 provided copies for all resources used.

Assuming that having copies of search strategies supports confidence in reproducibility as well as facilitating the actual reproducing of electronic searches, the results above suggest that there should be more frequent provision of copies of search strategies for *every* electronic source used. Others have similarly argued recently for this expanded reporting for electronic searches (Shokraneh, 2019).

However, even without such copies provided for every electronic resource listed in an SR, using the Campbell SR results as a comparison that indicates what might generally be possible and expected, in keeping with the PRISMA guidance, the psychology SR could and should more frequently provide a copy for at least one of the electronic resources used. SR in psychology routinely use PsycINFO, and such copies have been possible with PsycINFO available from vendors such as EBSCO, FirstSearch, Ovid, and with PsycARTICLES from the American Psychological Association. Although our results and interpretation will benefit from additional verification, there are concerns that approximately 90% of the 68 psychology SR we assessed did not provide copies for any of the searches used. Assessment of SR published more recently than those we looked at could show improvements in this kind of reporting.

As an alternative to providing copies as we defined them is not possible, researchers often list and report information for the individual search strategy elements that they used. We evaluated SR that used this approach to reporting, and we found that only 38% of those psychology SR reported basic information that supports confidence in the reproducibility of electronic searches . That assessment and data are discussed on the OSF site (see 'non-PRISMA' in Supplementary files).

In the past the space limits for journals have been a challenge to detailed reporting, and key guidelines have recognized such challenges, even while calling for fuller reporting (Liberati et al., 2009). Recent signs, such as the use of supplemental online files, indicate that the online environment will help to reduce the impact of space limits (Cooper & VandenBos, 2013; LeBel et al., 2013; Atkinson et al., 2015), and the journal *Archives of Scientific Psychology*, and the Open Science Framework[13] are two venues that support provision of needed reporting information. However, in addition to reducing space constraints, the need for fuller reporting to support reproducibility will be addressed significantly if what we have described as copies are required as a part of all electronic searches in SR.

3.3 A PRISMA set. The reporting of recommended information for recommended sets of search elements.

As our second approach to evaluating Psych SR, and reporting for searches, the results in Table 5 are for a set of elements/items that may be viewed as a fairly stringent PRISMA type of reporting

pertaining to reproducibility. These results show what we found as we assessed our samples of Psych SR and Campbell SR.

Table 5. Frequencies for a PRISMA set of elements/Items
1. Are names given for the electronic search resources used? (for every electronic resource) • PsycARTICLES 2009-2012 (24/25): 1- 8, 10-13, 15, 17-25, 27, 29 • Campbell (25/25): 1-4, 6, 8, 9, 11, 12, 18-20, 22, 24, 26, 27, 29, 31, 33, 34, 36-38, 41, 42 • PsycARTICLES 2014-2016 (24/25): 3, 13, 14, 17, 27, 30, 32, 33, 35, 37, 39, 44, 45, 50, 55, 56, 59, 61, 62, 65, 66, 67, 68, 69 • PsycARTICLES 2017 (17/18): 1, 4, 5, 6, 9, 16-19, 22, 30, 32, 36, 39, 40, 43, 44
2. All of the reports scoring positive for 1. that also listed a publication time frame for any electronic resource ? • PsycARTICLES 2009-2012 (19/25): 3-8, 11-13, 15, 17-20, 23-25, 27, 29 • Campbell (21/25): 1, 3, 6, 8, 9, 11, 18, 20, 22, 24, 26, 27, 29, 31, 33, 34, 36-38, 41, 42 • PsycARTICLES 2014-2016 (14/25): 3, 13, 14, 17, 32, 33, 45, 50, 56, 61, 61, 66, 67, 69 • PsycARTICLES 2017 (15/18): 1, 4, 5, 9, 16-19, 22, 30, 32, 36, 39, 40, 43
3. All of the reports scoring positive for 2. that also stated or showed that it was providing a copy of the electronic search strategy for any electronic resource as a saved/downloaded/copy-pasted 'search history' from the actual electronic search session. • PsycARTICLES 2009-2012 (1/25): 24 • Campbell (14/25): 1, 6, 9, 11, 18, 20, 24, 31, 33, 34, 36, 38, 41, 42 • PsycARTICLES 2014-2016 (2/25): 50, 56 • PsycARTICLES 2017 (4/18): 4, 22, 39, 40
4. All of the reports scoring positive for 3. that also listed a vendor, host, or publisher for any electronic resource. • PsycARTICLES 2009-2012 (1/25): 24 • Campbell (10/25): 1, 6, 9, 11, 33, 34, 36, 38, 41, 42 • PsycARTICLES 2014-2016 (0/25): • PsycARTICLES 2017 (3/18): 4, 39, 40
*Note: Numbers in parentheses show report frequencies for articles in PsycARTICLES SR and Cambell SR. Numbers following the colon correspond to articles in numbered lists of articles in Appendixes 1, 2, 3, and 4, respectively, for PsycARTICLES 2009-2012, Campbell, PsycARTICLES 2014-2016, and PsycARTICLES 2017.

Looking at item 3 in Table 5, our findings show that, for the randomly selected 25 Campbell SR that we examined, only 14 provided recommended information for the first three of our PRISMA set of recommended search report elements. In other words, forty-four percent did not provide this PRISMA search set information. The requirement of seeing vendor information reduced that number to 10 of 25 reporting desired information for our PRISMA set of reporting information elements. In comparison, the data show that the reporting for a PRISMA set in the 25 randomly selected Psych SR for 2009-2012 is much lower. Only 1 of those 25 Psych SR provided information for that set of PRISMA elements. The data for the randomly selected set of Psych SR for a 2014-2016 publication time frame showed that low reporting for that PRISMA set continued, and data for the 2017 set of SR was only slightly better. Overall, our assessment shows about half of the Campbell SR reporting for this PRISMA set, and reporting this information in the 68 SR from psychology is considerably lower. If electronic search reproducibility is viewed as dependent on reporting that is equivalent to what we call our PRISMA set, our findings suggest that readers would not have a strong basis for confidently assuming that the electronic searches in psychology SR are reproducible. This is a concern to address in the current studies and discussions of reproducibility in psychology.

3.4 Confidence items

There are search report elements/items that can provide added support for confidence in the reproducibility of searches, as well as for confidence in the use of those searches. The results in Table 6 present Kappa and report frequency data for these search elements based on our assessment of SR from psychology and the Campbell Library for the years 2009-2016, as well as for psychology SR for 2017.

Table 6. Items and sub-questions that support reader confidence in reproducibility

Items/sub-questions	Frequency PsycARTICLES 2009--2012	Kappa (sig) PsycARTICLES 2009--2012	Frequency Campbell 2009--2012	Kappa (sig) Campbell 2009--2012	Fisher's Exact	Frequency PsycARTICLES 2014--2016	Kappa (sig) PsycARTICLES 2014--2016	Frequency PsycARTICLES 2017*
1. Does the report show the total number of hits for each 'final strategy used'?								
a. For any resource	5/25	52 (.003)	4/25	.15 (.315)	1.000	5/25	.519 (.003)	NA
b. For every resource	4/25	.83 (.000)	3/25	.78 (.000)	1.000	4/25	.834 (.000)	3/18
c. Across all resources	17/25	.09 (.656)	15/25	.51 (.004)	.769	18/25	.088 (.656)	5/18
2. Does the report indicate that at least two researchers made independent choices (while looking at search results at the level of the title/abstract) to select the items for possible inclusion?								
a. For any resource	3/25	.43 (.009)	13/25	.52 (.009)	.005	7/25	.300 (.090)	7/18
b. For every resource	3/25	.78 (.000)	13/25	.52 (.008)	.005	5/25	.500 (.011)	NA
c. After de-duping	1/25	***	5/25	.26 (.184)	.189	5/25	.560 (.005)	NA
3. Does the report provide any inter-rater reliability results for the independent IE choices made at the level of the title/abstract?								
a. For any resource	2/25	.65 (.001)	1/25	***	1.000	5/25	.143 (.394)	1/18
b. For every resource	2/25	.65 (.001)	1/25	***	1.000	2/25	-.059 (.758)	NA
c. After de-duping	0/25	***	2/25	-.04 (.835)	.490	4/25	.400 (.050)	NA
4. Does the report identify which hits were included at the level of title/abstract?	3/25	.28 (.046)	8/25	.58 (.003)	.171	0/25	-.119 (.520)	0/18

* Note: Date of search - November 11, 2017. Assessment data is for a set of 18 items that had explicit reference to the PRISMA statement as explained in this paper.

Our Kappa results suggest challenges for assessing some of these confidence elements with our items; and, again, more training could give better agreement results. Additionally, in comparison to our item 2, other studies using the related but more general AMSTAR item for 'duplicate study

selection and data extraction' found Kappas of .93 (Popovich et al., 2012) and .77 (Pieper et al., 2015).

Relative to our third research question, our results also show low frequencies for clear reporting. Frequency of reporting for this item 1 was low, although reporting such numbers has been recommended by many including the PRISMA group (Liberati et al., 2009). Reporting information for our item 2 is encouraged, or even expected, for some projects (e.g. major health, psychology, or policy topics). This recommendation is seen in the IOM's Finding What Works in Health Care: Standards for Systematic Reviews (Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, Board on Health Care Services, & Institute of Medicine, 2011), and is reiterated by the Cochrane Collaboration in the MECIR standards for the reporting of new reviews of interventions (MECIR Manual, n.d.).The reporting of inter-rater agreement at the point of assessing the title and abstract (item 3) also is encouraged in the PRISMA guidelines for study selection. Our findings show some of this reporting for selection at the point of screening title and abstracts. Reporting for our item 4, the identification of items chosen or included at the title and abstract screening was infrequent, though it was evident in some SR. While space considerations could account for that finding, this would be information that supports readers who wish to be confident that they are reproducing the initial selection choices made as title and abstracts are screened.

Information for each of these confidence items is easily reported, and, going forward, such reporting will support confidence in conclusions of SR. That confidence in SR conclusions is important for all readers including other researchers, clinicians, and those who develop policies or share SR research information with the public.

Summary and conclusions

Recent discussions and research in psychology show a significant emphasis on reproducibility. Concerns pertain to methods as well as results, and this paper contributes to discussions about the methods that are used for systematic reviews. We specifically examined the reporting of electronic searches used for SR in psychology. Such reports are key for determining the reproducibility of electronic searches. Confidence in the reproducibility of electronic searches can also impact the confidence that readers have in the overall results or conclusions of systematic reviews.

In this paper we first discuss systematic reviews, reproducibility, electronic searches, transparent reporting, and the increased use of SR in psychology. Based on evidence-based recommendations, we developed and used 12 checklist items to evaluate electronic search reporting that supports reproducibility. Item Kappa results ranged from fair to almost perfect. Then, mirroring comparisons of reporting in Cochrane SR to that found in non-Cochrane SR, using those checklist items we compared reports in SR published in the psychology literature to those SR from the Campbell Library. Reporting of basic recommended electronic search step information that supports reproducibility was seen significantly less in psychology SR.

Additionallly, we found that 90% of the 68 psychology SR that we assessed did not provide what we defined as a *copy of a full search strategy* for *any* of the electronic search resources used. Moreover,

assuming that higher levels of reproducibility are supported as more of *a set of recommended reporting criteria* are met, we used a *set of checklist items* to represent a 'PRISMA' type of recommended reporting. We found that only one of the 25 randomly selected psychology SR from 2009-2012 reported recommended information for all items in the set, and none of the 25 psychology SR from 2014-2016 did so. Furthermore, although the set of 18 Psych SR from 2017 was used because each *referred to* the *PRISMA statement*, only 3 reported information for all in our 'PRISMA set' of search report elements.

We also looked at reporting for we view as 'confidence items' that can be a part of reporting of electronic searches in SR. Items covered reporting for the number of hits for every electronic search, the number of hits for all electronic searches combined, the use of two or more researches for independent title/abstract screening, the inter-rater agreement for two or more researches for independent title/abstract screening, and the identification of items selected for possible use at the end of title/abstract screening. About half of the 68 Psych SR we assessed reported the number of hits combined across all electronic searches; and reporting for the other search items was very low.

Based on our findings, we had six general conclusions.

1. Electronic search reporting in published SR in psychology shows that improvements should be made that support confidence in the reproducibility of electronic searches used.
2. As shown in our assessment data for the SR from the Campbell Collaboration, it does seem possible to report what we called our *PRISMA set* for *every* electronic resource used.
3. Reporting for what we called a PRISMA set should be seen more in SR published in psychology.
4. It does seem that reporting in SR could include more reporting for all of what we called confidence items. This information supports reader confidence not only in the searches run but also in the use of search results.
5. Findings from the current study could serve as a baseline for this kind of reporting in SR that are published in psychology.
6. The research checklist we developed and used had inter-rater agreement that suggests it might serve as a resource for those concerned with the reproducibility of electronic searches. That checklist, or some version, might be used for evaluating electronic searches, for research on the items themselves and/or for research on the reproducibility of electronics searches in SR.

Improving the reporting of electronic search strategies so that readers can be confident in the reproducibility of such searches can be challenging. However, we believe the improvements that we describe for SR are possible. And, going forward, improvements in the reporting of electronic searches used for SR can serve as 'clear signals' that provide a stronger basis for confidence in the results and conclusions of SR in psychology.

References

Agency for Healthcare Research and Quality. (2014). *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Agency for Healthcare Research and Quality, Retrieved from https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/cer-methods-guide_overview.pdf

Ali, N. B., & Usman, M. (2018). Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy. *Information and Software Technology*, *99*, 133–147. https://doi.org/10.1016/j.infsof.2018.02.002

APA Databases Methodology Field Values. (n.d.). Retrieved October 30, 2019, from https://www.apa.org website: https://www.apa.org/pubs/databases/training/method-values

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Aromataris, E., & Munn, Z. (2019). Chapter 1: JBI Systematic Reviews - JBI Reviewer's Manual - JBI GLOBAL WIKI. In *JBI Reviewer's Manual*. Retrieved June 29, 2020, from https://wiki.joannabriggs.org/display/MANUAL/Chapter+1%3A+JBI+Systematic+Reviews

Atkinson, K. M., Koenka, A. C., Sanchez, C. E., Moshontz, H., & Cooper, H. (2015). Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate. *Research Synthesis Methods*, *6*(1), 87–95. https://doi.org/10.1002/jrsm.1127

Booth, A. (2006). Brimful of STARLITE : toward standards for reporting literature searches. *Journal of the Medical Library Association*, (4), 421.

Campbell, M., Katikireddi, S. V., Sowden, A., & Thomson, H. (2019). Lack of transparency in reporting narrative synthesis of quantitative data: a methodological assessment of systematic reviews. *Journal of Clinical Epidemiology*, *105*, 1–9. https://doi.org/10.1016/j.jclinepi.2018.08.019

Centre for Reviews and Dissemination. (2009). *Systematic Reviews. CRD's guidance for undertaking reviews in health care*. Retrieved June 29, 2020, from https://www.york.ac.uk/crd/guidance/

Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Nuclear and Radiation Studies Board, Division on Earth and Life Studies, … National Academies of Sciences, Engineering, and Medicine. (2019). Confidence. In *Reproducibility and Replicability in Science*. https://doi.org/10.17226/25303

Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, Board on Health Care Services, & Institute of Medicine. (2011). *Finding What Works in Health Care: Standards for Systematic Reviews* (J. Eden, L. Levit, A. Berg, & S. Morton, Eds.). National Academies Press. https://doi.org/10.17226/13059

Cooper, H. M. (2017). *Research synthesis and meta-analysis: a step-by-step approach* (Fifth edition). Thousand Oaks, California: SAGE Publications, Inc.

Cooper, H., & VandenBos, G. R. (2013). Archives of scientific psychology: A new journal for a new era. *Archives of Scientific Psychology*, *1*(1), 1–6. https://doi.org/10.1037/arc0000001

Delaney, A., & Tamás, P. A. (2018). Searching for evidence or approval? A commentary on database search in systematic reviews and alternative information retrieval methodologies. *Research Synthesis Methods*, *9*(1), 124–131. https://doi.org/10.1002/jrsm.1282

Faggion, C. M. (2019). Should a systematic review be tested for reproducibility before its publication? *Journal of Clinical Epidemiology*, *110*, 96. https://doi.org/10.1016/j.jclinepi.2019.02.008

Fehrmann, P., & Thomas, J. (2011). Comprehensive computer searches and reporting in systematic reviews: Computer searches and reporting in reviews. *Research Synthesis Methods*, *2*(1), 15–32. https://doi.org/10.1002/jrsm.31

Gensby, U., Lund, T., Kowalski, K., Saidj, M., Jørgensen, A. K., Filges, T., … Labriola, M. (2012). Workplace Disability Management Programs Promoting Return to Work: A Systematic Review. *Campbell Systematic Reviews*, *8*(1). https://doi.org/10.4073/csr.2012.17

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience: Openness in cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1396*(1), 5–18. https://doi.org/10.1111/nyas.13325

Golder, S., Loke, Y. K., & Zorzela, L. (2013). Some improvements are apparent in identifying adverse effects in systematic reviews from 1994 to 2011. *Journal of Clinical Epidemiology*, *66*(3), 253–260. https://doi.org/10.1016/j.jclinepi.2012.09.013

Golder, S., Loke, Y., & McIntosh, H. M. (2008). Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *Journal of Clinical Epidemiology*, *61*(5), 440–448. https://doi.org/10.1016/j.jclinepi.2007.06.005

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, *1*(1), 28. https://doi.org/10.1186/2046-4053-1-28

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), null. https://doi.org/10.1146/annurev-statistics-031219-041104

Higgins, J. P. T., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

Jamieson, K. H., McNutt, M., Kiermer, V., & Sever, R. (2019). Signaling the trustworthiness of science. *Proceedings of the National Academy of Sciences*, *116*(39), 19231–19236. https://doi.org/10.1073/pnas.1913039116

Kisely, S., Chang, A., Crowe, J., Galletly, C., Jenkins, P., Loi, S., … Macfarlane, S. (2015). Getting started in research: systematic reviews and meta-analyses. *Australasian Psychiatry*, *23*(1), 16–21. https://doi.org/10.1177/1039856214562077

Koffel, J. B., & Rethlefsen, M. L. (2016). Reproducibility of Search Strategies Is Poor in Systematic Reviews Published in High-Impact Pediatrics, Cardiology and Surgery Journals: A Cross-Sectional Study. *PLOS ONE*, *11*(9), e0163309. https://doi.org/10.1371/journal.pone.0163309

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A. K., Hammerstrøm, K., & Sathe, N. (2017). Searching for studies: a guide to information retrieval for Campbell systematic reviews. *Campbell Systematic Reviews*, *13*(1), 1–73. https://doi.org/10.4073/cmg.2016.1

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science*, *8*(4), 424–432. https://doi.org/10.1177/1745691613491437

Leclercq, V., Beaudart, C., Ajamieh, S., Rabenda, V., Tirelli, E., & Bruyère, O. (2019). Meta-analyses indexed in PsycINFO had a better completeness of reporting when they mention PRISMA. *Journal of Clinical Epidemiology*, *115*, 46–54. https://doi.org/10.1016/j.jclinepi.2019.06.014

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., … Wieland, L. S. (2019). Searching for and selecting studies. In *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 67–107). https://doi.org/10.1002/9781119536604.ch4

Lefebvre, C., Manheimer, E., & Glanville, J. (2011). Chapter 6: Searching for studies. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Retrieved June 29, 2020, from www.handbook.cochrane.org

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Maggio, L. A., Tannery, N. H., & Kanter, S. L. (2011). Reproducibility of Literature Search Reporting in Medical Education Reviews: *Academic Medicine*, *86*(8), 1049–1054. https://doi.org/10.1097/ACM.0b013e31822221e7

McDonagh, M., Peterson, K., Raina, P., Chang, S., & Shekelle, P. (2008). Avoiding Bias in Selecting Studies. In *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Agency for

Healthcare Research and Quality (US). Retrieved June 29, 2020, from http://www.ncbi.nlm.nih.gov/books/NBK126701/

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *Journal of Clinical Epidemiology*, *75*, 40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021

Meader, N., King, K., Llewellyn, A., Norman, G., Brown, J., Rodgers, M., … Stewart, G. (2014). A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation. *Systematic Reviews*, *3*(1), 82. https://doi.org/10.1186/2046-4053-3-82

Meert, MLIS, D., Torabi, MLIS, N., & Costella, DDS, MSc, MLIS, J. (2017). Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *Journal of the Medical Library Association*, *104*(4), 267–277. https://doi.org/10.5195/JMLA.2016.139

MECIR Manual. (n.d.). Retrieved January 2, 2020, from https://community.cochrane.org/mecir-manual

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and Reporting Characteristics of Systematic Reviews. *PLoS Medicine*, *4*(3), e78. https://doi.org/10.1371/journal.pmed.0040078

Mullins, M. M., DeLuca, J. B., Crepaz, N., & Lyles, C. M. (2014). Reporting quality of search methods in systematic reviews of HIV behavioral interventions (2000-2010): are the searches clearly explained, systematic and reproducible? *Research Synthesis Methods*, *5*(2), 116–130. https://doi.org/10.1002/jrsm.1098

Ng, C., & Benedetto, U. (2016). Evidence Hierarchy. In G. Biondi-Zoccai (Ed.), *Umbrella Reviews* (pp. 11–19). https://doi.org/10.1007/978-3-319-25655-9_2

Niederstadt, C., & Droste, S. (2010). Reporting and presenting information retrieval processes: the need for optimizing common practice in health technology assessment. *International Journal of Technology Assessment in Health Care*, *26*(4), 450–457. https://doi.org/10.1017/S0266462310001066

Novotney, A. (2014). Reproducing results. *Monitor on Psychology*, *45*(8). Retrieved June 29, 2020, from https://www.apa.org/monitor/2014/09/results

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., … Moher, D. (2020). Mapping of reporting guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines. *Journal of Clinical Epidemiology*, *118*, 60–68. https://doi.org/10.1016/j.jclinepi.2019.11.010

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Paul, M., & Leibovici, L. (2014). Systematic review or meta-analysis? Their place in the evidence hierarchy. *Clinical Microbiology and Infection*, *20*(2), 97–100. https://doi.org/10.1111/1469-0691.12489

Pieper, D., Buechter, R. B., Li, L., Prediger, B., & Eikermann, M. (2015). Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties. *Journal of Clinical Epidemiology*, *68*(5), 574–583. https://doi.org/10.1016/j.jclinepi.2014.12.009

Popovich, I., Windsor, B., Jordan, V., Showell, M., Shea, B., & Farquhar, C. M. (2012). Methodological Quality of Systematic Reviews in Subfertility: A Comparison of Two Different Approaches. *PLoS ONE*, *7*(12), e50403. https://doi.org/10.1371/journal.pone.0050403

PRISMA. (n.d.). Retrieved October 31, 2019, from http://www.prisma-statement.org/

PsycARTICLES. (n.d.). Accessed June 29, 2020, from https://www.apa.org/pubs/databases/psycarticles/index

Qualtrics XM - Experience Management Software. (n.d.). Retrieved October 31, 2019, from Qualtrics website: https://www.qualtrics.com/

Rader, T., Mann, M., Stansfield, C., Cooper, C., & Sampson, M. (2014). Methods for documenting systematic review searches: a discussion of common issues. *Research Synthesis Methods*, *5*(2), 98–115. https://doi.org/10.1002/jrsm.1097

Sampson, M., & McGowan, J. (2006). Errors in search strategies were identified by type and frequency. *Journal of Clinical Epidemiology*, *59*(10), 1057.e1-1057.e9. https://doi.org/10.1016/j.jclinepi.2006.01.007

Sampson, M., McGowan, J., Tetzlaff, J., Cogo, E., & Moher, D. (2008). No consensus exists on search reporting methods for systematic reviews. *Journal of Clinical Epidemiology*, *61*(8), 748–754. https://doi.org/10.1016/j.jclinepi.2007.10.009

Sampson, M., McGowan, J., Cogo, E., Grimshaw, J., Moher, D., & Lefebvre, C. (2009). An evidence-based practice guideline for the peer review of electronic search strategies. *Journal of Clinical Epidemiology*, *62*(9), 944–952. https://doi.org/10.1016/j.jclinepi.2008.10.012

Sayre, F., & Riegelman, A. (2018). The Reproducibility Crisis and Academic Libraries. *College & Research Libraries*, *79*(1), 2–9. https://doi.org/10.5860/crl.79.1.2

Schalken, N., & Rietbergen, C. (2017). The Reporting Quality of Systematic Reviews and Meta-Analyses in Industrial and Organizational Psychology: A Systematic Review. *Frontiers in Psychology*, *8*, 1395. https://doi.org/10.3389/fpsyg.2017.01395

Shea, B., Moher, D., Graham, I., Pham, B., & Tugwell, P. (2002). A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. *Evaluation & the Health Professions*, *25*(1), 116–129.

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., … Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, *62*(10), 1013–1020. https://doi.org/10.1016/j.jclinepi.2008.10.009

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Medical Research Methodology*, *12*(1), 181. https://doi.org/10.1186/1471-2288-12-181

Tunis, A. S., McInnes, M. D. F., Hanna, R., & Esmail, K. (2013). Association of Study Quality with Completeness of Reporting: Have Completeness of Reporting and Quality of Systematic Reviews and Meta- Analyses in Major Radiology Journals Changed Since Publication of the PRISMA Statement? *European Journal of Radiology*, *269*(2), 413–426.

Vazire, S. (2017). Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology*, *3*(1), 1. https://doi.org/10.1525/collabra.74

Yong, E. (2013). Psychologists strike a blow for reproducibility. *Nature*, nature.2013.14232. https://doi.org/10.1038/nature.2013.14232

Yoshii, A., Plaut, D. A., McGraw, K. A., Anderson, M. J., & Wellik, K. E. (2009). Analysis of the reporting of search strategies in Cochrane systematic reviews. *Journal of the Medical Library Association : JMLA*, *97*(1), 21–29. https://doi.org/10.3163/1536-5050.97.1.004

Williams, M., Bagwell, J., & Nahm Zozus, M. (2017). Data management plans: the missing perspective. *Journal of Biomedical Informatics*, *71*, 130–142. https://doi.org/10.1016/j.jbi.2017.05.004

Endnotes

[1] Paul Fehrman, MA, MLS, is a Research and Instruction Services Librarian with the University Libraries at Kent State University, pfehrman@kent.edu

[2] Megan Mamolen, PhD, MLIS, is PhD, MLIS is a Reference, Instruction, E-Resources Librarian at Lakeland Community College, mmamolen1@lakelandcc.edu

[3] With respect to *transparent reporting of searches to identify resources with data* for SR, recent related discussion of data management planning also argues for clear documentation of such 'upstream activites that determine data quality' (Williams et al., 2017).

[4] We want to express appreciation to Matthew Cox, MLIS. His contribution was significant for the work on reproducibility presented at the Annual Campbell Collaboration Colloquium, held May 21-23, 2013 in Chicago. That poster is on the OSF site noted just below.

[5] OSF site for this paper - https://osf.io/g6x4k This site includes supplementary materials noted in the paper.

[6] In the Spring of 2016, we learned of work on 'PRISMA-Search'. The goal of that project has been to develop an extension to the PRISMA Statement to guide the reporting of components critical to a reproducible search. Information has been available in 'Reporting guidelines under development' on the EQUATOR Network website (www.equator-network.org/). Moreover, in the Spring of 2019, the authors of PRISMA-S shared that checklist and explanation documents were available for review. Information for PRISMA-S has been available in 'Reporting guidelines under development' on the EQUATOR Network website and more recently on the Open Science Framework - https://doi.org/10.17605/OSF.IO/YGN9W. We anticipate that when the extension work is complete it will provide additional support for our selection of items to assess electronic search reproducibility.

[7] The four appendixes noted in this paper, as well as additional background information about the checklists used for this study, are available on the paper's OSF site noted above.

[8] As an example of a paper using the APA guidance see this paper and supplemental file published in the *Archives of Scientific Psychology*. Youngstrom, E. A., Genzlinger, J. E., Egerton, G. A., & Van Meter, A. R. (2015). Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology*, *3*(1), 112–137. https://doi.org/10.1037/arc0000024

[9] https://amstar.ca/

[10] https://www.cadth.ca/resources/finding-evidence/press

[11] https://doi.org/10.17605/OSF.IO/YGN9W

[12] Of the 14 Campbell SR in item 1. a., one did provide a full copy of at least one electronic search, but not for all of the electronic resources used (see 11. in Appendix 2).

[13] https://osf.io