Gunnell, K, Poitras, V and Tod, D

Questions and answers about conducting systematic reviews in sport and exercise psychology

Questions and answers about conducting systematic reviews in sport and exercise psychology

Katie E. Gunnell[1], Veronica J. Poitras[2], David Tod[3]

Word count: 8092

Disclosure Statement: The authors have no conflicts of interest to disclose

Corresponding Author:

[1]Katie Gunnell, PhD, Carleton University. A511 Loeb Building, 1125 Colonel By Drive

Ottawa ON, K1S 5B6. Katie.gunnell@carleton.ca, Phone: 613-520-2600 x2419

Co-Authors:

[2]Veronica J. Poitras, Independent Researcher

[3]David Tod, Faculty of Science, School of Sport and Exercise Sciences, Liverpool John Moores University

Questions and answers about conducting systematic reviews in sport and exercise psychology

Word count: 8092

Abstract

Systematic reviews are used to gain insight into the state of research on a given topic, theory, or process; or to inform the development of guidelines, interventions, and policy or public health strategies. Challenges associated with conducting a systematic review include the rapid increase in the variety of systematic review methods and the number of decisions that researchers must make during the process. The purpose of this paper is to provide succinct responses to common questions researchers face when conducting a systematic review. The manuscript is structured around 13 questions that arise during the systematic review process. The questions span the development stage (e.g., why and where should systematic reviews be preregistered; how to decide on inclusion and exclusion criteria), methodological stage (e.g., how to develop and execute a search strategy), and publication stage (e.g., what should be placed in online supplements). Each question was answered with a concise response with recommendations based on the scientific literature and current advances in systematic review techniques. Researchers who have never conducted a systematic review or who are wishing to reflect on their knowledge and practice in conducting a systematic review will benefit from the up-to-date procedures outlined herein.

Keywords: literature review; meta-analysis; synthesis; methodology; kinesiology

Researchers in sport and exercise psychology are increasingly conducting systematic reviews (Tod & Eubank, 2017). Although systematic reviews are thought to minimize bias and provide a balanced and rigorous account of a topic, they are conducted with varying degrees of robustness and the decisions made about the process are often subjective. The purpose of this paper is to provide sport and exercise psychology researchers with considerations and options to commonly asked questions that might arise during the systematic review process. Our aim is to provide up-to-date information to enhance rigour, transparency, and replicability. The target audience for this paper is people who have never conducted systematic reviews as well as those wishing to reflect on their knowledge and practices so they can continue to refine or update their skills. The common questions we present herein are not exhaustive. The questions represent topics that we have found relevant while conducting or assisting others with systematic reviews, and reviewing or managing systematic reviews submitted to academic journals. The answers we provide should be viewed as considerations rather than prescriptive guidelines. We draw heavily from health and psychology disciplines, and as such, some procedures may require modification for sport and exercise psychology contexts (e.g., the population, intervention, comparator, outcome (PICO) method discussed in question 4 might require tailoring for studies using observational data). It is not our intent to provide step-by-step procedures for all aspects of the systematic review process, but rather to focus on key aspects of the systematic review process where decision-making is required. Readers interested in step-by-step guidance for all aspects of quantitative systematic reviews of randomized controlled trials are encouraged to consult the Cochrane Handbook for Systematic Reviews of Interventions (freely accessible: www.training.cochrane.org/handbook) (Higgins, Thomas, et al., 2019). Further, the Systematic Review toolbox (http://systematicreviewtools.com/) serves as a searchable resource with tools

for supporting a systematic review across various disciplines (Marshall, 2018). Finally, it is important to note that a systematic review (i.e., a type of review) is not the same thing as a meta-analysis (i.e., a statistical procedure; see answer to question 10 below). Readers interested in conducting a meta-analysis can use the information provided herein when conducting a systematic review with meta-analysis to guide their systematic review process.

1. What is the difference between a review and a systematic review?

Reviews can be classified as either non-systematic or systematic (Ferrari, 2015).

Non-systematic Review

A non-systematic review (sometimes referred to as a "narrative review") is undertaken with the purpose of describing a particular aspect of the available evidence (or a subset of studies) to summarize information, explain how and why studies fit together, draw conclusions and suggest future research (Ferrari, 2015; Siddaway, Wood, & Hedges, 2018). Non-systematic reviews do not follow a standardized or reproducible methodology, are often broad, and involve selectively discussing the literature on a particular topic (Ferrari, 2015). Non-systematic reviews can be useful when resources to conduct a systematic review are limited or the research question is broad (Ferrari, 2015). For example, Sarkar and Fletcher (2014) conducted a narrative review on psychological resilience in sport performers with the explicit purpose of providing broad and extensive coverage of the topic. As is typically done with non-systematic reviews, Sarkar and Fletcher (2014) did not document reproducible study selection and search strategies, but rather, indicated that studies were selected based on their significance with respect to advancing knowledge of the topic.

Systematic Review

A systematic review is undertaken with the purpose of comprehensively synthesizing all available evidence that fits pre-specified eligibility criteria to answer a specific question (Chandler et al., 2019). The methodology used in a systematic review is unlike that used in a non-systematic review because it is characterized by explicit, systematic, and reproducible methods that attempt to capture all studies that meet set inclusion and exclusion criteria (Chandler et al., 2019). For example, Carson and colleagues (2016) conducted a systematic review to examine the effects of physical activity on cognitive development in children. In this review, a precise research question was posed and explicit methods for study selection, literature searching, and data extraction and analysis were provided. In recent years, review methods have diversified and this has led to the development of many different types of reviews that use systematic methods (e.g., scoping reviews). It is beyond the scope of this this manuscript to discuss them in detail.

2. Are there guidelines for conducting a systematic review?

Guidelines and reporting standards have been developed to outline best practices for conducting robust and replicable systematic reviews. Because many guidelines exist, researchers are advised to use a guideline that suits the focus of their systematic review, as described below and in Table 1.

A commonly used reporting standard in sport and exercise psychology is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009). PRISMA is comprised of evidence-based items to be considered when reporting reviews of quantitative primary studies (Moher et al., 2009). Although PRISMA is intended for reviews of intervention studies, it can be amended for reviews of observational studies. Many researchers in sport and exercise psychology (e.g., Carson et al.,

2016) implement the PRISMA checklist items and the PRISMA flow chart that is used to describe the flow of studies throughout screening (e.g., number of studies excluded after title screening, number of studies excluded with reasons after full-text screening; Moher et al., 2015). Several different PRISMA checklists exist to cover a range of reviews. For example, the PRISMA-Protocol (PRISMA-P; Moher et al., 2015) was developed to help researchers create robust and replicable protocols for their systematic reviews. Wurz and Brunet (2016) used the PRISMA-P to develop their systematic review protocol on physical activity and quality of life in adult cancer survivors.

The Cochrane Handbook for Systematic Reviews of Interventions (Higgins, Thomas, et al., 2019) provides in-depth guidance for researchers who are undertaking a Cochrane Intervention review (i.e., a systematic review that is prepared by a Cochrane Review Group). In conjunction with the Cochrane Handbook, guidance is available from the Cochrane Qualitative and Implementation Methods group (Cochrane, 2018) for conducting reviews to summarize qualitative studies. Sport and exercise psychology researchers can use both of these resources to guide their systematic reviews. At this juncture it is important to note that many of the recommendations within the Cochrane Handbook might need to be adapted for sport and exercise psychology research. For example, having two independent reviewers extract data might not be feasible (see question 6).

Other examples of tools include the Meta-analysis Of Observational Studies in Epidemiology (MOOSE; Stroup et al., 2000) for systematic reviews that focus on primary quantitative studies that use observational designs (Stroup et al., 2000) or the ENhancing Transparency in REporting the synthesis of Qualitative research (ENTREQ; Tong, Flemming, McInnes, Oliver, & Craig, 2012). ENTREQ was created to mirror the PRISMA guidelines but

for application to qualitative primary studies. In particular, ENTREQ encourages the use of the PRISMA flowchart but is influenced by the Standards for Reporting Literature Searches (STARLITE; Booth, 2006) insofar as it includes items that capture the sampling strategy (Booth, 2016). Additionally, A MeaSurement Tool to Assess Systematic Reviews (AMSTAR and AMSTAR-2; Shea et al., 2007, 2017) is an instrument for critically appraising systematic reviews that can also be used to guide the conduct of systematic reviews. The PRISMA-P is specifically designed for systematic review protocols; other checklists described above (e.g., MOOSE, AMSTAR-2) can also serve to guide the review process and ensure methodical and reporting quality.

3. Why and where should systematic reviews be preregistered?

Psychology, one of the parent disciplines to sport and exercise psychology, is currently going through a replication crisis (Maxwell, Lau, & Howard, 2015). In other words, findings from primary studies are unlikely to be replicable, raising concerns over false positives or false negatives, lack of statistical power, or insufficient transparency in research methods (Maxwell et al., 2015). Systematic reviews are not immune to replication issues. In particular, systematic reviews require numerous subjective decisions that, combined with poor reporting, could lead to systematic reviews that cannot be replicated, or at worst add misinformation to the literature (such as when systematic reviews have conflicting results) (Ioannidis, 2016).

To curb the replication crisis, researchers have advocated for more transparent research reporting (Munafò et al., 2017). “Open science” describes strategies to increase research transparency and replicability (Nosek et al., 2015) and has been recommended for sport and exercise psychology researchers (Tamminen & Poucher, 2018). As part of the open science framework, researchers are encouraged to preregister studies, publish protocols, share data and

materials, and publish in open access journals. Beyond the discipline of sport and exercise psychology, researchers have advocated that systematic reviews be preregistered to enhance transparency and reproducibility while reducing bias (Ioannidis, 2016; Munafò et al., 2017; Stewart, Moher, & Shekelle, 2012). A critical component of a good quality systematic review is the development of a protocol that outlines the main objectives, design features, and planned analyses *before* the review process begins (Stewart et al., 2012). Preregistration and protocols serve as guides for conducting systematic reviews and should contain sufficient information to allow for replication. Therefore, we recommend that preregistrations and protocols align with reporting guidelines (e.g., PRISMA-P, MOOSE, AMSTAR-2).

With respect to bias, the information provided in preregistered reports or published protocols allows readers to determine if the published review was completed as intended or if unintended or undocumented changes were made. This helps determine if selective reporting bias is present in a review (Shamseer et al., 2015). Preregistration should occur before study screening for eligibility has begun, to negate the possibility that viewing articles could change inclusion/exclusion criteria (Stewart et al., 2012). It is important to document if changes to the protocol are required after the review has begun; most preregistration platforms allow researchers to alter their protocol and create an audit trail with information explaining the revisions for complete transparency. Finally, PRISMA guidelines advocate for journals and funding agencies to require preregistration for publication or funding purposes (Moher et al., 2015).

A commonly used free preregistration platform for health research, and by extension some sport and exercise psychology research that specifically has a health-related outcome, is the International Prospective Register of Systematic Reviews (PROSPERO; Booth et al., 2012; see

Table 1). Within PROSPERO, researchers can document all protocol aspects including research question(s), databases searched and search strategies, selection criteria, and procedures for quality assessment and data extraction. Effective October 2019, PROSPERO only accepts registrations where data extraction has not commenced. Some of the information requested within PROSPERO might not be applicable to sport and exercise psychology research and may therefore require adapted responses. For example, PROSPERO requests information about the study PICO and it is possible that SPIDER may be a more amenable framework (see descriptions of PICO and SPIDER in question 4). In an applied example, a researcher interested in the effects of a psychological skills intervention on health-related outcomes in elite athletes might enter "not applicable" in the PROSPERO form under the category for comparator(s)/control if they have decided *a priori* to include all studies regardless of whether or not there was a control group. Researchers can complete these sections by adapting the fields as needed. Researchers can also use the Open Science Framework website (Open Science Framework, 2018; see Table 1) to register systematic reviews. Lastly, systematic review protocols can be published in various scholarly journals (e.g., Wurz & Brunet, 2016 for an example from sport and exercise psychology), some of which might require fees.

4. How are decisions about study eligibility made?

Arguably, one of the most important steps of a systematic review is to develop the research question (Thomas, Kneale, McKenzie, Brennan, & Bhaumik, 2019). Refining the research question involves determining the underlying objective of the review (e.g., general knowledge and understanding, guiding policy or practice) and developing an appropriate research question that follows from the objective. A good research question will focus the topic and guide study eligibility criteria and selection (Thomas et al., 2019). A research question that is

narrow may lead to a small and concise review that informs a particular context but has limited generalizability (Siddaway et al., 2018). In contrast, a broad research question may lead to a larger and more complex review that might be broadly relevant but difficult to apply to a particular context (Siddaway et al., 2018). There are trade-offs between the speed at which a review can be conducted and its scope (e.g., a more narrow scope translates to a more rapid review) and between scope and generalizability (e.g., a more narrow scope translates to good generalizability to that specific context but poor generalizability more broadly). Having knowledge about the topic and past research is useful when refining the research question and overall scope of the systematic review. The Cochrane Handbook recommends that the research question be a concise statement that specifies the population, intervention (and comparison) and outcomes of interest (Thomas et al., 2019). For example, Carson et al, (2016) had the broad objective "to comprehensively review all observational and experimental studies examining the relationship between physical activity and cognitive development during early childhood (birth to 5 years)" (p. 575). Conversely, Caddick and Smith (2014) had a more narrow objective focusing on a specific population of interest; their objective was to "evaluate the current evidence base surrounding the impact of sport and physical activity upon the well-being of combat veterans" (p. 10).

The Population, Intervention, Comparison, and Outcome (PICO) framework was created to help focus research questions and identify relevant evidence for clinical or intervention-based quantitative research (Schardt, Adams, Owens, Keitz, & Fontelo, 2007). Researchers have shown the PICO criteria to increase precision in the development of the search strategies (Schardt et al., 2007). Others have modified the PICO criteria to suit different bodies of literature. For example, Methley et al. (2014) produced PICOS where the "s" refers to study design such that PICOS can

be used to specify which types of quantitative (e.g., non-randomized trials, prospective cohorts, cross-sectional) and qualitative studies are eligible for inclusion. Another framework designed specifically for mixed-methods or qualitative studies is the Sample, Phenomenon of interest, Design, Evaluation, Research Type (SPIDER) tool (Cooke, Smith, & Booth, 2012). Example PICO and SPIDER criteria for sport and exercise psychology research questions are provided in Table 2. Readers are referred to Poitras et al., (2017), McGowan et al., (2018), and Woods et al., (2017) for studies that have employed PICO, SPIDER, and PICOS respectively, in sport and exercise psychology.

In some areas, when there is not a lot of direct evidence, researchers may choose to include indirect evidence. Researchers gather direct evidence when they observe variables reflecting their constructs of interest (e.g., clinical diagnosis of anxiety), whereas they gather indirect evidence when they observe variables that are proxies for the constructs of interest (e.g., self-reported scores on an anxiety symptom test). The difference between direct and indirect evidence may be extended to interventions and population samples. For example, evaluating an athlete's performance-related anxiety during a competition could provide direct evidence whereas evaluating an athlete's performance-related anxiety during a lab task might provide indirect evidence. Although it might be necessary to include indirect evidence, it is important to recognize that this represents a limitation to the systematic review because the results might not generalize to the specific research question of interest. For example, Kelley and colleagues (2017) were interested in the comparative effects of different types of exercise (e.g., aerobic, strength training, or both) on adiposity in overweight and obese children and adolescents. Recognizing that direct evidence comparing different types of exercise may not be available (e.g., in head-to-head trials), the authors decided *a priori* that they would include indirect

evidence, comparing any of the eligible types of exercise to a control group (e.g., placebo or wait-list control). Indirect evidence is "lower quality" by nature of its separation from the comparison(s) of interest; when included, the quality of the evidence should be clearly reported (see question 11). For example, in their review on prenatal exercise and depression and anxiety, Davenport (2018) and colleagues "downgraded" the quality of the evidence because some exercise interventions included other behavioural components (e.g., diet), thereby making it impossible to attribute the results to exercise directly. In all cases, justification for the decisions regarding inclusion of indirect evidence should be transparent within the manuscript and/or preregistration protocol.

Another consideration when deciding on eligibility is whether to include only studies from published peer-reviewed journals. Grey literature is literature that is typically not published in peer-reviewed scholarly resources such as books or journals (Lefebvre et al., 2019). It can include conference abstracts, dissertations, theses, government reports, or technical reports. Including grey literature may be important when reviewing interventions that are relatively novel since some evidence may be available that has not yet made it into the peer-reviewed journal domain. Excluding grey literature in health research can bias the results in a manner that exaggerates the effect size (i.e., yields larger overall effect sizes compared to when grey literature was included). This exaggeration could be because published trials are likely to show greater effects (Hopewell, McDonald, Clarke, & Egger, 2007; McAuley, Pham, Tugwell, & Moher, 2000).

Including grey literature, however, also has limitations. Including conference abstracts could result in a large increase in the number of records to screen, with little added value for including those that are identified. For instance, conference abstracts often contain interim data

or incomplete data (e.g., only the results that were statistically significant were presented which may be subject to selective reporting or publication bias), and there is insufficient information in abstracts to determine the level of confidence in the evidence they contain. Grey literature is typically not peer-reviewed and grey literature documents might be difficult to locate or access. Researchers must therefore weigh the advantages and disadvantages associated with grey literature in deciding on inclusion criteria. Readers are referred to Avugos et al., (2013) for an example of a review that included grey literature in sport and exercise psychology.

5. How is a search strategy developed?

A systematic approach must be used to search the literature for all studies meeting the inclusion/exclusion criteria. Given that most researchers in sport and exercise psychology are unlikely to have library science training, involving a librarian or research information specialist in the development of a search strategy can help to ensure the search strategy will capture relevant literature across databases. This is particularly useful given that each database has its own syntax and language that requires knowledge and expertise. Researchers unable to involve a library scientist can consult the Cochrane Handbook (Higgins, Thomas, et al., 2019) for suggested search strategies, consult similar published systematic reviews for their search strategies, and familiarize themselves with database searching through each databases online resources.

In addition to electronic searches, manual searches such as hand searching, reviewing reference lists, and asking researchers familiar with the literature to identify studies, should be conducted (Lefebvre et al., 2019). Manual searches are needed because electronic searches have been shown to capture only a subset of the relevant primary studies. For example, one study on clinical trials showed electronic retrieval rates ranging from 42-80% of the total number of

relevant studies, with the remainder identified through hand searching methods (Hopewell, Clarke, Lefebvre, & Scherer, 2007). Given that qualitative studies might use unique wording or have different indexing (Noyes et al., 2019), manual searches may be especially useful for systematic reviews of qualitative studies.

Depending on the research question, researchers in sport and exercise psychology may use a variety of databases to conduct their searches. For example, health-oriented research questions might warrant searching the MEDLINE database whereas non-health-related questions might omit this database. Databases have some degree of overlap in terms of the journals and publishers that are indexed. It is therefore important that each database be considered based on the number of potentially relevant studies it will yield relative to the total number of studies the search will retrieve to maximize coverage and minimize screening burden. Bramer and colleagues (2017) recently found that four databases (Embase, MEDLINE, Web of Science, and Google Scholar) should be searched, with PsyINFO being added for systematic reviews in the behavioral and health sciences. Although there are no concrete rules for the number of databases required to maximize identification of relevant studies, at a minimum researchers are encouraged to search multiple databases to ensure relevant studies are captured.

When a researcher decides to include grey literature, it is important they decide when to stop scanning that literature. For example, if Google is used to search for evidence, ‘stopping rules’, such as planning to stop screening after 100 consecutive non-relevant records, or after 10 pages of hits, should be specified *a priori*. Readers interested in more information and checklists for conducting good quality grey literature searchers are referred to the Canadian Agency for Drugs and Technologies in Health (2018).

6. How are search strategies executed?

When executing a search strategy, researchers should keep detailed records to assist in the creation of the PRISMA-type flow chart, to document the search procedure for transparency, and to allow other researchers to replicate or update the search (Lefebvre et al., 2019). Details to record and report for each database include the date the search was performed, the number of records returned, the number of duplicate records, and the exact name of the database (including the week of search within the database if the database lists this information (e.g., 1980 to 2019 Week 15).

Developing and executing the search strategy can be linear or iterative. An iterative search procedure is one wherein the search procedure is updated based on results from previous searches (e.g., new key words are identified; Lefebvre et al., 2019) whereas a linear search process is not. Iterative search procedures have been recommended within the Cochrane Handbook. Irrespective of the approach taken, it is unlikely that researchers will be able to retrieve every relevant primary study. As such, it is important that researchers acknowledge the search limitations and the implications for the results. By following robust search methods, and complementing database searches with manual searches, it is unlikely that critically important studies (e.g., that could substantially change the magnitude or direction of the findings) will be missed. To enhance transparency and replicability, all search execution decisions should be justified and documented within the manuscript.

Lastly, once the searches have been executed, the researcher will typically merge all records from each database together to be screened against inclusion/exclusion criteria. Given that many studies are indexed in multiple databases, it is important that researchers remove duplicate studies to reduce the burden of screening articles. De-duplication can be conducted

through reference management software, such as Reference Manager (Thompson Reuters, San Francisco, CA), or EndNote (Reuters, 2018).

7. Should a search be updated before the systematic review is published?

Systematic reviews can become quickly outdated, sometimes even before they are published (Borah, Brown, Capers, & Kaiser, 2017; Shojania et al., 2007). It may be necessary to run an 'update' to the search to capture new studies that were published while data were being screened and extracted. Ideally, screening, data extraction, and manuscript preparation should occur as soon as possible after the search such that it does not become outdated before publication. The Methodological Expectations of Cochrane Intervention Reviews (Chandler, Lasserson, Higgins, & Churchill, 2016) recommend all searches be updated within 12 months before publication. Decisions about whether an update is needed can vary by topic area. For instance, if a research question draws on a well-established body of literature then it may be unlikely that the addition of a handful of recent studies will dramatically alter the conclusions of a systematic review, rendering an update less critical. However, if a research question addresses a relatively novel area of research then failing to capture the most recently-published studies could result in reaching erroneous conclusions due to missing information. In some cases, researchers may choose to do targeted updates that aim to balance comprehensiveness with feasibility (e.g., if an original search was not limited by study design, an update may be limited to randomized controlled trials for expediency (e.g., Poitras et al., 2017). Decisions about whether an update is needed can also be at the discretion of the editor or editorial team for a particular journal. Readers interested in recommendations for updating a systematic review after it has been published are referred to a recent consensus statement and checklist (Garner et al.,

2016) or the Methodological Expectations of Cochrane Intervention Reviews guidelines (Chandler et al., 2016).

8. How should articles be screened for inclusion and what can be used to make the procedures more efficient and robust?

The decision to include or exclude a study during the screening process is critical because these studies will form the basis of the data and results of the systematic review (J. McGowan et al., 2016). The Cochrane Handbook (2011) recommends screening titles and abstracts as a first level to remove articles that are obviously irrelevant. At the next level, full text articles that remain should be examined against inclusion/exclusion criteria. Despite creating clear inclusion/exclusion criteria, the screening process may require subjective judgements (Lefebvre et al., 2019). As such, it has been recommended that at least two reviewers screen each full text article to ensure relevant studies are not erroneously excluded (Edwards et al., 2002).

Before screening begins, procedures to deal with conflicts between reviewers should be documented. Often times, when a conflict arises about whether a study meets inclusion or exclusion criteria, it can be settled through discussion between the reviewers. Sometimes conflicts are simple (e.g., one reviewer misinterpreted the information). Other times, they can be more complex (e.g., differences in operationalization of a variable). For example, one reviewer might operationalize “self-concept” as an indicator of “mental health” whereas another might not. If the conflict cannot be easily resolved, it is recommended that a third person be consulted to help resolve the conflict (Lefebvre et al., 2019). Other methods to help mitigate and/or resolve conflicts include creating code-books that provide reviewers with precise definitions, examples, and rules for making decisions, holding regular team meetings before and during screening, and contacting corresponding authors to obtain clarification when needed.

It is also essential that *a priori* decisions be made and documented about whether to include studies if the information provided is unclear or conflicts cannot be resolved. Some options for dealing with lack of clarity during screening include: (a) retaining the study and explicitly discussing the areas of uncertainty within the paper, (b) excluding the study and recording how many studies were excluded for this reason, (c) or contacting the authors to obtain more information. In the last case, the researcher will also need to determine what contingency strategy will be used if the authors do not respond. These are all decisions that should be made before screening begins.

Many researchers from sport and exercise psychology use Microsoft Excel to create forms or checklists to screen articles. Although useful for small systematic reviews, Excel does present challenges. For example, it is easy for one reviewer to accidently override a cell without knowing or tracking the change. Software is available to facilitate screening. DistillerSR (Evidence Partners, 2018; https://www.evidencepartners.com) and Covidence (Covidence, 2018; https://www.covidence.org) are examples of systematic review management software. DistillerSR and Covidence are not free but different pricing options are available and some institutions may support licenses. These software programs typically allow researchers to (a) set up the levels of screening (e.g., titles and abstracts, full text), (b) highlight key terms in abstracts, (c) upload search results, (d) calculate interrater agreement/reliability, and (e) track changes, duration spent screening, and progress of each reviewer. Additionally, software allows researchers to automate whether a study advances to the next level of screening (e.g., full text) based on the responses reviewers provide. For example, the researcher might require two reviewers to exclude a study during abstract and title screening but only one reviewer to pass the study to the next level of screening. Only requiring a single reviewer for inclusion at the title and

abstract level is referred to as ‘accelerated screening’ since it improves efficiency without compromising rigour. Further, software can be used to identify discrepancies in reviewer evaluations (i.e., conflicts) and how they are handled. The software typically has the ability to generate the PRISMA flow chart thereby reducing the potential for human error. Lastly, it is important to note that each software program has different capabilities. Therefore, researchers are encouraged to determine whether software is needed and if so which software is best for their circumstances. Most software programs provide free trials that might be useful in determining relevance.

9. How can data be extracted?

Data extraction forms should be thoroughly piloted to ensure comprehensiveness and clarity (Li, Higgins, & Deeks, 2019). During pilot testing, the forms are likely to be altered through consensus to avoid future conflicts. When forms are modified, they should be pilot tested in their newer iteration. Data extraction forms vary depending on the research question and type of review. Most forms will include fields for identifying the reviewer and the study (e.g., title of paper and authors or unique ID), and questions with simple tick box options (e.g., “yes”, “no”, “unclear”) or open-ended responses to expedite data extraction. Forms will typically query methodology (e.g., study design, subgroup analyses of interest), population (e.g., sample size, age, baseline characteristics), details of the intervention/exposure and comparator, results (e.g., means, standard errors), other relevant findings, and information used in risk of bias assessment (e.g., how participants were randomized; see question 11) alongside space for notes (Li et al., 2019).

The Cochrane Handbook recommends more than one reviewer extract the data to minimize bias and reduce errors (Li et al., 2019). Nonetheless, data extraction by two reviewers

is often not feasible from a logistical or resourcing perspective. In this case, it is possible for one reviewer to extract the data and a second reviewer to verify that extraction in totality (e.g., Poitras et al., 2017) or a subset of the studies (e.g., Harlow, Wolman, & Fraser-Thomas, 2018). Alternatively, one reviewer can extract study characteristics (e.g., demographic information) and two reviewers extract the substantive data (e.g., results) (Mathes, Klaßen, & Pieper, 2017). These alternatives involve trade-offs between robustness and feasibility.

Data extraction should occur independently by reviewers who have complementary knowledge and expertise (Li et al., 2019). Reviewers should have experience using the extraction forms and be appropriately trained. As with screening, there are likely to be discrepancies between reviewers during data extraction (e.g., determining which values of an outcome variable to extract), and it is essential that a plan be developed in advance for resolving these disagreements. Generally, discussion between the two reviewers is sufficient; otherwise, a third person may be needed or it may be necessary to contact the study's corresponding author for clarification (Li et al., 2019). It is possible to quantify the amount of agreement between reviewers (e.g., interrater reliability, percent agreement) to determine the reliability of the data extraction.

10. How are data analyzed in a systematic review?

The most appropriate type of data analysis in a systematic review depends on the research question and/or the nature of the available data (e.g., quantitative or qualitative data). Where feasible, it is advised that researchers involve methodological experts (e.g., statisticians, qualitative experts) to conduct the analysis or provide advice during data analysis.

Within a systematic review of studies that use quantitative data, a meta-analysis might be performed when the researcher wants to quantify an overall effect size of homogenous studies

(e.g., studies that use the same design, populations, intervention and comparisons, and outcome measures). Systematic reviews and meta-analyses are not the same thing; a systematic review is a type of review whereas meta-analysis is a type of statistical analysis used in any type of review (systematic or non-systematic) (Tod & Eubank, 2017) to combine the data from independent primary studies to obtain an overall summary of the effect (Deeks, Higgins, & Altman, 2019). A meta-analysis conducted within a non-systematic review may be inaccurate and/or biased (Tod & Eubank, 2017). Software such as RevMan (RevMan, 2014) or Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2013) can be used to conduct meta-analyses. RevMan is freely downloadable whereas Comprehensive Meta-Analysis is not. An overview of meta-analysis methods with application to sport and exercise psychology is provided by Ahn and colleagues (2016).

It is important to recognize that meta-analysis is not always possible or appropriate - even if the *a priori* plan was to use this statistical method. For example, sometimes there are not enough data from primary studies to obtain reliable effect size estimates via pooling in meta-analysis. Alternatively, sometimes data are not available in primary studies, or are available with an insufficient level of detail to be entered into meta-analysis (e.g., means, standard deviations, and correlation tables are not provided). Other times, the studies obtained in the searches are too clinically, methodologically, or statistically heterogeneous to combine. Clinical heterogeneity can come from differences in interventions (e.g., a writing intervention vs. verbal intervention), outcomes (e.g., anxiety vs. depression), or participants (e.g., elite athletes vs. recreational athletes) (Deeks et al., 2019). Methodological heterogeneity can come from differences in the design of a study (e.g., randomized controlled trial vs. longitudinal observational), the measurements used (e.g., different self-report measures that operationalize a conceptual

construct differently), and the risk of bias (e.g., group allocation concealed in one study and not concealed in another) (Deeks et al., 2019). Statistical heterogeneity occurs when there is more variability in the results across studies than expected due to chance alone (Deeks et al., 2019). Clinical and methodological heterogeneity can be assessed through subjective evaluations by the researchers conducting the review (e.g., whether the primary studies included participants with similar characteristics such as age and comorbidities) (Deeks et al., 2019). Statistical heterogeneity can be assessed using various statistical methods including chi-square or I^2. Readers are encouraged to consult the Cochrane Handbook (Deeks et al., 2019) or Borentsein and colleagues (2009) for more information on assessing heterogeneity and conducting meta-analysis.

When a meta-analysis is not possible or appropriate, researchers will sometimes use a technique to synthesize the evidence called “vote counting”. Vote counting occurs when the researcher tallies how many studies found a positive, negative, or null relationship (Siddaway et al., 2018). An example from sport and exercise psychology that used vote counting is provided in Gunnell et al., (2019). Although this appears to be a simple solution when meta-analyses cannot be performed, vote counting has been discouraged for a few reasons. For example, vote counting weights effects from all studies equally irrespective of characteristics such as sample size or the quality of the evidence. In other words, a study with only 10 participants would be given the same weight as a study with 1,000 participants. Furthermore, with vote counting there is no way to quantify the effect size or its limits of confidence (Bushman, 1994); the only information provided is how many studies found an effect and/or the direction of the effect. As an alternative, a narrative synthesis might be conducted, in which the researcher qualitatively summarizes the evidence (Siddaway et al., 2018). Note that a narrative synthesis (i.e., a type of data analysis) is

distinct from a ‘narrative review’ (i.e., a type of non-systematic review). With narrative synthesis, it is important to have rules that guide the synthesis strategy given that it is possible for researchers to implicitly vote-count. Caddick and Smith (2014) used narrative synthesis to analyze data in their systematic review on sport and physical activity and well-being in combat veterans. Although the decision to conduct a meta-analysis, vote-counting, or narrative synthesis is driven by the nature of the available data, it is important to have an *a priori* plan.

For qualitative data, meta-syntheses are conducted to summarize and synthesize evidence (Siddaway et al., 2018; Thorne, Jensen, Kearney, Noblit, & Sandelowski, 2004). Recently, the Cochrane Collaboration created the Cochrane Qualitative and Implementation Methods Group (https://methods.cochrane.org/qi/welcome) to advise Cochrane about policy and practice for qualitative synthesis, and to create and provide methodological guidance and training (Cochrane, 2018). Analysis and synthesis of qualitative data can be aggregative, summative or descriptive, or interpretative and theory generating (see Noyes & Lewin, 2011 for more details). A researcher interested in assembling and pooling data from qualitative independent studies will use the aggregation/summation approach and could include methods such as thematic analysis without theory generation, meta-aggregation, or meta-summary (Noyes & Lewin, 2011). Researchers interested in developing concepts and theories to understand themes from qualitative studies may use interpretive or theory-generating methods such as meta-ethnography, thematic analysis with theory generation, or grounded theory (Noyes & Lewin, 2011). In sport and exercise psychology, McGowan et al., (2018) used thematic analysis to analyze qualitative primary studies in their systematic review on the acceptability of physical activity for older adults. For more information on these methods, readers are encouraged to consult Noyes & Lewin (2011) and Cochrane (2018).

The synthesis methods described above are not exhaustive; other strategies for synthesizing quantitative and qualitative primary studies are available (Barnett-Page & Thomas, 2009; Deeks et al., 2019; Holt et al., 2017). Readers are encouraged to explore the various synthesis methods and select, with justification, the most appropriate method for their systematic review.

11. How should the quality and risk of bias of the primary studies be assessed?

It is important to recognize that conclusions from a high quality systematic review that follows guidelines and uses robust and replicable procedures are only as good as the quality of the studies on which they are based (Weir, Rabia, & Ardern, 2016). Researchers are encouraged to include quality assessments at both the level of the primary studies and the overall body of evidence (see question 12) to ensure the results of the review are situated in the quality of the evidence. For example, it would be misleading to report a strong effect size from a meta-analysis without also reporting that it was calculated using low quality data from primary studies.

"Quality" and "risk of bias" are often mistaken as interchangeable terms, when in fact they are distinct but related concepts (Higgins & Green, 2011). Quality refers to the degree to which studies have been conducted in alignment with the highest possible standards (e.g., including facets related to power calculations, ethical clearance). Conversely, bias refers to a systematic error, or deviation from the truth, that can lead to underestimation or overestimation of the true effect of a treatment intervention (Boutron et al., 2019). It is usually not possible to know to what extent potential biases have affected the results of a particular study (and results of a study may be unbiased despite methodological flaws), so judgments about bias are most appropriately termed 'risks' of bias (Boutron et al., 2019). Thus in simplest terms, "quality assessment" involves assessing whether the highest possible standards were met, and "risk of

bias assessment" involves appraising the degree to which potential biases may have led to underestimation or overestimation of an effect.

The operationalization of quality assessment is broad and there are various tools that can be used to assess study quality. For example, the National Institutes of Health (NIH; https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools) provides many quality assessment tools that can be selected based on the type of primary studies (e.g., intervention, case-control, observational and cross-sectional). These tools offer tailored questions about the methods of a study (e.g., sampling, measures, data handling) based on study type and culminate in an overall rating of "good", "fair", or "poor" (National Institutes of Health, 2018). Checklists for various types of primary studies are also available from the Critical Appraisal Skills Program (Critical Appraisal Skills Program, 2019; https://casp-uk.net/casp-tools-checklists/) and the Joanna Briggs Institute (Joanna Briggs Institute, 2019; http://joannabriggs.org/research/critical-appraisal-tools.html), and a compendium of critical appraisal tools for qualitative research has been compiled (Majid & Vanstone, 2018). As an example, the Critical Appraisal Program (CASP) contains 10 questions in three domains: (1) validity of results (2) results, and (3) local applicability of the results. Results from the checklist are used to determine if the studies were of low, medium or high quality (Critical Appraisal Skills Program, 2019).

Reviewers may enhance the rigor of their reviews by assessing both risk of bias and quality. A study can be conducted with high methodological quality, yet still have important risks of bias (Boutron et al., 2019). For example, it is often impossible to blind participants to intervention conditions in sport and exercise research (e.g., participants who volunteer for a study on physical activity will likely know if they were assigned to a physical activity vs. waitlist control group). In this case, the studies may have been conducted with sufficient methodological

quality, but still be at risk of performance bias that may lead to over- or underestimation of an effect.

The Cochrane risk of bias tool is commonly used for reviews focusing on randomized controlled trials (Higgins, Savović, Page, Elbers, & Sterne, 2019). The most recent Cochrane Risk of Bias Tool 2 (Sterne et al., in press) can be used to rate primary studies on five domains including risk of bias from (1) the randomization process (e.g., whether the allocation sequence was random) (2) deviations from the intended interventions (e.g., whether the participants were aware of their assigned intervention), (3) missing outcome data (e.g., whether data for the outcome were available for all, or nearly all of participants randomized) (4) measurement of the outcome (e.g., whether the outcome measure was appropriate), and (5) selection of the reported results (e.g., whether the data for the results were analyzed according to a pre-specified analysis plan) (Sterne et al., in press). The rating system for each domain is either "low", "high", or "unclear" risk of bias (Sterne et al., in press) and the researcher makes a judgment with supporting statements about the risks of bias (Higgins, Savović, et al., 2019).

Tools for assessing the risk of bias in non-randomized trials are also available, such as the Risk of Bias in Non-randomised Studies of Interventions (ROBINS-I; Sterne et al., 2016). The ROBINS-I tool domains include bias due to (1) confounding, (2) selection of participants into the study (3) classification of interventions, (4) deviations from intended interventions, (5) missing data, (6) measurement of outcomes, and (7) selection of the reported results (Sterne et al., 2016). Within ROBINS-I, risk of bias is assessed as "low", "moderate", "serious" and "critical" risk (Sterne et al., 2016).

Using Tools to Guide the Assessment of Risk of Bias and Quality

As described above, there are many tools and checklists available to assist researchers in assessing quality and risk of bias, all of which have strengths and weaknesses. Some checklists provide information on potential sources of bias and how to recognize them, but may be overly simplistic when it comes to evaluating the potential impact of the risk of bias. For example, a study may have only one source of bias, but this source might have critical implications in terms of interpreting the findings, whereas another study may have several areas that could introduce a risk of bias but other factors that enable confidence in the findings. Some tools also omit sources of bias, or require in-depth knowledge of study design and/or the content area in order to be used, and may have low interrater reliability. Items from multiple tools might be needed to consider all potential sources of bias. In fact, recent research has shown that the use of different tools could lead to opposite conclusions for a given set of studies, particularly when risk of bias is rated on either end of the high or low spectrum (Losilla, Oliveras, Marin-Garcia, & Vives, 2018). As such, researchers should be careful to reduce discrepancies and use domain-specific risk of bias assessments when available (Losilla et al., 2018). Further, it would be useful for researchers to pilot test the tools before starting assessment and explicitly document why that particular tool was selected (Losilla et al., 2018), including documenting how training was undertaken and if the tool had evidence of validity and reliability in similar contexts.

Finally, researchers should recognize that although scores and checklists allow authors to summarize sources of bias and indicators of quality within and across studies, they should be used as tools/guides to critically appraise the studies and evidence and to inform the researchers' level of confidence in the evidence. The Cochrane Handbook (2019) recommends four possible approaches to integrate the critical appraisals within analyses of a systematic review with meta-analyses. First, only results from studies that were rated as having a low risk of bias could be

presented. Second, results could be stratified based on the assessments of risk of bias (e.g., present results separately for studies with low or high risk of bias). Third, the results could be presented alongside narrative discussion of the risk of bias assessments; however, this should only be done when all studies have the same risk of bias (see Boutron et al., 2019, for further detail). Lastly, statistical methods could be used to adjust the effect estimate for bias (Boutron et al., 2019). For systematic reviews that do not use meta-analyses and rely on other quantitative or qualitative methods of synthesis, options 1 to 3 above can be used (Boutron et al., 2019).

12. How should the quality of the body of evidence be assessed?

The Cochrane Handbook (Higgins, Thomas, et al., 2019) recommends that the Grading of Recommendations Assessment, Development, and Evaluation (GRADE; Guyatt et al., 2008) framework be used to examine the quality of the evidence for the body of quantitative investigations (i.e., collection of studies) for a particular outcome of interest within a systematic review. Here, "quality" is defined as the degree to which users can be confident in the effect or association (Guyatt et al., 2008). Using this framework, the body of evidence is classified as "high", "moderate", "low", or "very low" quality. The quality of evidence begins with "high" for randomized experiments and "low" for observational studies. Other limitations across studies lead to downgrading the quality of evidence, such as serious risk of bias, indirectness (e.g., the measures used were not direct measures of the variable of interest), inconsistency of effects (e.g., the effects observed were not consistent across studies), and imprecision (e.g., large confidence intervals for the results) (Guyatt et al., 2011). For observational studies, which are common in sport and exercise psychology, if there is no cause to downgrade, the quality can be upgraded in the presence of certain conditions (e.g., large magnitude of effect, presence of dose-response).

When primary studies in a review are qualitative, the GRADE- Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) can be used (Lewin et al., 2018).

13. How can readers be provided with access to all the details needed to replicate or appraise the quality of the systematic review?

Publishing constraints limit the level of detail that can be provided in the body of the systematic review; however, supplementary materials can often be published online. Supplemental files should be provided so that all information about the review is available to readers, reviewers, and editors for complete transparency. Additionally, providing supplementary files aids with archiving, updating, and maintaining systematic reviews. In order for a review to be replicated or updated, the precise methodological details are needed. Types of information to place in online supplemental materials may include detailed search strategies, lists of excluded studies with reasons for exclusion, data coding manuals, data extraction documents, and justification for assessments of risk of bias.

Conclusion

Systematic reviews are characterized by explicit, systematic methods for comprehensively synthesizing evidence. Making informed decisions *a priori*, and reporting them explicitly and transparently, will enhance the robustness of the systematic review. We answered 13 commonly asked questions that arise during the systematic review process with the aim of highlighting current advances in systematic review techniques. Importantly, we document current perspectives on improving the rigour of systematic reviews through:

- following established guidelines (e.g., PRISMA, MOOSE, ENTREQ),
- *a priori* preregistering reviews and/or publishing protocols where possible,

- developing explicit eligibility criteria based on the research question(s) (e.g., using PICO, SPIDER),
- creating tailored search strategies, executing searches, and updating searches,
- screening citations and arriving at the set of included studies,
- extracting and analyzing data,
- assessing risk of bias and quality of primary studies and the body of evidence,
- providing readers with sufficient detail to enable replication

We are hopeful that readers find this useful for developing a platform from which to explore these topics in more depth before conducting systematic reviews. Ultimately, we urge researchers to enhance the rigour of their systematic reviews to advance the field.

References

Ahn, S., Lu, M., Lefevor, G. T., Fedewa, A. L., & Celimli, S. (2016). Application of meta-analysis in sport and exercise science. In N. Ntoumanis & N. D. Myers (Eds.), *An introduction to intermediate and advance statsistical analyses for sport and exercise scientists* (pp. 233–254). Chichester, UK: Wiley & Sons, Ltd.

Avugos, S., Köppen, J., Czienskowski, U., Raab, M., & Bar-Eli, M. (2013). The "hot hand" reconsidered: A meta-analytic approach. *Psychology of Sport and Exercise*, *14*(1), 21–27. https://doi.org/10.1016/j.psychsport.2012.07.005

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: A critical review. *BMC Medical Research Methodology*, *9*(1), 59. https://doi.org/10.1186/1471-2288-9-59

Booth, A. (2006). "Brimful of STARLITE": Toward standards for reporting literature searches. *Journal of the Medical Library Association*, *94*(4), 421-e205.

Booth, A. (2016). Searching for qualitative research for inclusion in systematic reviews: A structured methodological review. *Systematic Reviews*, *5*(1), 74. https://doi.org/10.1186/s13643-016-0249-x

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: An international prospective register of systematic reviews. *Systematic Reviews*, *1*(1), 2. https://doi.org/10.1186/2046-4053-1-2

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, *7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2013). Comprehensive Meta-analysis (Version 3). Englewood, NJ.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis—Northwestern Scholars*. Retrieved from https://www.scholars.northwestern.edu/en/publications/introduction-to-meta-analysis

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., & Hróbjartsson, A. (2019). Chapter 7: Considering bias and conflicts of interest among the included studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews, 6*. https://doi.org/10.1186/s13643-017-0644-y

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In *The handbook of research synthesis* (pp. 193–213). New York, NY, US: Russell Sage Foundation.

Caddick, N., & Smith, B. (2014). The impact of sport and physical activity on the well-being of combat veterans: A systematic review. *Psychology of Sport and Exercise, 15*(1), 9–18.

Canadian Agency for Drugs and Technologies in Health. (2018). Grey Matters: A practical tool for searching health-related grey literature | CADTH.ca. Retrieved November 23, 2018, from https://www.cadth.ca/resources/finding-evidence/grey-matters

Carson, V., Hunter, S., Kuzik, N., Wiebe, S. A., Spence, J. C., Friedman, A., … Hinkley, T. (2016). Systematic review of physical activity and cognitive development in early

childhood. *Journal of Science and Medicine in Sport*, *19*(7), 573–578. https://doi.org/10.1016/j.jsams.2015.07.011

Chandler, J., Cumpston, M., Thomas, J., Higgins, J. P. T., Deeks, J. J., & Clarke, J. J. (2019). Chapter 1: Introduction. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Chandler, J., Lasserson, T., Higgins, J. P. T., & Churchill, R. (2016). Standards for the planning, conduct and reporting of updates of Cochrane Intervention Reviews. In J. P. T. Higgins, T. Lasserson, J. Chandler, & R. Churchill (Series Ed.), *Methodological Expectations of Cochrane Intervention Reviews*. London: Cochrane.

Cochrane. (2018). Cochrane qualitative & implementation methods group. Retrieved December 8, 2018, from https://methods.cochrane.org/qi/welcome

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis. *Qualitative Health Research*, *22*(10), 1435–1443. https://doi.org/10.1177/1049732312452938

Covidence. (2018). *Covidence*. Retrieved from https://www.covidence.org/home

Critical Appraisal Skills Program. (2019). CASP checklists. Retrieved January 9, 2019, from CASP - Critical Appraisal Skills Programme website: https://casp-uk.net/casp-tools-checklists/

Davenport, M. H., McCurdy, A. P., Mottola, M. F., Skow, R. J., Meah, V. L., Poitras, V. J., … Ruchat, S.-M. (2018). Impact of prenatal exercise on both prenatal and postnatal anxiety and depressive symptoms: A systematic review and meta-analysis. *British Journal of Sports Medicine*, *52*(21), 1376–1385. https://doi.org/10.1136/bjsports-2018-099697

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2019). Chapter 10: Analysis data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Edwards, P., Clarke, M., DiGuiseppi, C., Pratap, S., Roberts, I., & Wentz, R. (2002). Identification of randomized controlled trials in systematic reviews: Accuracy and reliability of screening records. *Statistics in Medicine, 21*(11), 1635–1640. https://doi.org/10.1002/sim.1190

Evidence Partners. (2018). *DistillerSR*. Retrieved from https://www.evidencepartners.com/products/distillersr-systematic-review-software/

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing, 24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Akl, E. A., Beyene, J., … Schünemann, H. J. (2016). When and how to update systematic reviews: Consensus and checklist. *BMJ*, *354*, i3507. https://doi.org/10.1136/bmj.i3507

Gunnell, K. E., Poitras, V. J., LeBlanc, A. G., Schibli, K., Barbeau, K., Hedayati, N., … Tremblay, M. S. (2019). Physical activity and brain structure, brain function, and cognition in children and youth: A systematic review of randomized controlled trials—ScienceDirect. *Mental Health and Physical Activity*, *16*, 105–127. https://doi.org/10.1016/j.mhpa.2018.11.002

Guyatt, G. H., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., … Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of

findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Harlow, M., Wolman, L., & Fraser-Thomas, J. (2018). Should toddlers and preschoolers participate in organized sport? A scoping review of developmental outcomes associated with young children's sport participation. *International Review of Sport and Exercise Psychology*, *0*(0), 1–25. https://doi.org/10.1080/1750984X.2018.1550796

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. Retrieved from https://training.cochrane.org/handbook/archive/v5.1/

Higgins, J. P. T., Savović, J., Page, M., Elbers, R., & Sterne, J. A. C. (2019). Chapter 8: Assessing risk of bias in a randomized trials. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Higgins, J. P. T., Thomas, J., Chandler, J., Crumpston, M., Li, T., Page, M., & Welch, V. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook.

Holt, N. L., Neely, K. C., Slater, L. G., Camiré, M., Côté, J., Fraser-Thomas, J., … Tamminen, K. A. (2017). A grounded theory of positive youth development through sport based on

results from a qualitative meta-study. *International Review of Sport and Exercise Psychology, 10*(1), 1–49. https://doi.org/10.1080/1750984X.2016.1180704

Hopewell, S., Clarke, M., Lefebvre, C., & Scherer, R. (2007). Handsearching versus electronic searching to identify reports of randomized trials. *The Cochrane Database of Systematic Reviews*, (2), MR000001. https://doi.org/10.1002/14651858.MR000001.pub2

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *The Cochrane Database of Systematic Reviews*, (2), MR000010. https://doi.org/10.1002/14651858.MR000010.pub3

Ioannidis, J. P. A. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Joanna Briggs Institute. (2019). Critical Appraisal Tools—JBI. Retrieved January 9, 2019, from http://joannabriggs.org/research/critical-appraisal-tools.html

Kelley, G. A., Kelley, K. S., & Pate, R. R. (2017). Exercise and adiposity in overweight and obese children and adolescents: Protocol for a systematic review and network meta-analysis of randomised trials. *BMJ Open*, *7*(12). https://doi.org/10.1136/bmjopen-2017-019512

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., … Wieland, L. S. (2019). Chapter 4: Searching for and selecting studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., … Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. *Implementation Science*, *13*(1), 2. https://doi.org/10.1186/s13012-017-0688-3

Li, T., Higgins, J. P. T., & Deeks, J. J. (2019). Chapter 5: Collecting data. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Losilla, J.-M., Oliveras, I., Marin-Garcia, J. A., & Vives, J. (2018). Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology*, *101*, 61–72. https://doi.org/10.1016/j.jclinepi.2018.05.021

Majid, U., & Vanstone, M. (2018). Appraising Qualitative Research for Evidence Syntheses: A Compendium of Quality Appraisal Tools. *Qualitative Health Research*, *28*(13), 2115–2131. https://doi.org/10.1177/1049732318785358

Marshall, C. (2018). Systematic Review Toolbox. Retrieved November 23, 2018, from /help/tools-and-software/systematic-review-toolbox

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: A methodological review. *BMC Medical Research Methodology*, *17*. https://doi.org/10.1186/s12874-017-0431-4

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet, 356*(9237), 1228–1231. https://doi.org/10.1016/S0140-6736(00)02786-0

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

McGowan, L. J., Devereux-Fitzgerald, A., Powell, R., & French, D. P. (2018). How acceptable do older adults find the concept of being physically active? A systematic review and meta-synthesis. *International Review of Sport and Exercise Psychology*, *11*(1), 1–24. https://doi.org/10.1080/1750984X.2016.1272705

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: A comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, *14*. https://doi.org/10.1186/s12913-014-0579-0

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

National Institutes of Health. (2018). Study Quality Assessment Tools | National Heart, Lung, and Blood Institute (NHLBI). Retrieved October 25, 2018, from Study Quality Assessment Tools website: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Noyes, J., Booth, A., Cargo, M., Flemming, K., Harden, A., Harris, J., … Thomas, J. (2019). Chapter 21: Qualitative evidence. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Noyes, J., & Lewin, S. (2011). Guidance on Selecting a Method of Qualitative Evidence Synthesis, and Integrating Qualitative Evidence with Cochrane Intervention Reviews. In J. Noyes, A. Booth, K. Hannes, A. Harden, J. Harris, S. Lewin, & C. Lockwood (Eds.), *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions. Version 1 (updated August 2011).* Retrieved from http://cqrmg.cochrane.org/supplemental-handbook-guidance

Open Science Framework. (2018). OSF | Home. Retrieved November 7, 2018, from https://osf.io/

Poitras, V. J., Gray, C. E., Janssen, X., Aubert, S., Carson, V., Faulkner, G., … Tremblay, M. S. (2017). Systematic review of the relationships between sedentary behaviour and health indicators in the early years (0–4 years). *BMC Public Health*, *17*(Suppl 5). https://doi.org/10.1186/s12889-017-4849-8

Reuters, T. (2018). *EndNote*. Philidelphia, PA: Thomson Reuters.

RevMan. (2014). Review Manager (RevMan) (Version 5.3). Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Sarkar, M., & Fletcher, D. (2014). Psychological resilience in sport performers: A review of stressors and protective factors. *Journal of Sports Sciences*, *32*(15), 1419–1434. https://doi.org/10.1080/02640414.2014.901551

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(1), 16. https://doi.org/10.1186/1472-6947-7-16

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *349*, g7647. https://doi.org/10.1136/bmj.g7647

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., … Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(1), 10. https://doi.org/10.1186/1471-2288-7-10

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., … Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or

non-randomised studies of healthcare interventions, or both. *BMJ*, *358*, j4008. https://doi.org/10.1136/bmj.j4008

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine*, *147*(4), 224–233.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2018). How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-010418-102803

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R., Blencowe, N., Boutron, I., … Higgins, J. P. T. (in press). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*.

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews, 1*, 7. https://doi.org/10.1186/2046-4053-1-7

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., … Epidemiology (MOOSE)Group, for the M. O. O. S. in. (2000). Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. *JAMA*, *283*(15), 2008–2012. https://doi.org/10.1001/jama.283.15.2008

Tamminen, K. A., & Poucher, Z. A. (2018). Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise*, *36*, 17–28. https://doi.org/10.1016/j.psychsport.2017.12.010

Thomas, J., Kneale, D., McKenzie, J. E., Brennan, S. E., & Bhaumik, S. (2019). Chapter 2: Determining the scope of the review and the questions it will address. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Retrieved from www.training.cochrane.org/handbook

Thorne, S., Jensen, L., Kearney, M. H., Noblit, G., & Sandelowski, M. (2004). Qualitative metasynthesis: Reflections on methodological orientation and ideological agenda. *Qualitative Health Research*, *14*(10), 1342–1365. https://doi.org/10.1177/1049732304269888

Tod, D., & Eubank, M. R. (2017). Conducting a systematic review: Demystification for trainees in sport and exercise psychology. *Sport and Exercise Psychology Review, 13*. Retrieved from http://researchonline.ljmu.ac.uk/id/eprint/6033/

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

Weir, A., Rabia, S., & Ardern, C. (2016). Trusting systematic reviews and meta-analyses: All that glitters is not gold! *British Journal of Sports Medicine*, *50*(18), 1100–1101. https://doi.org/10.1136/bjsports-2015-095896

Woods, D., Breslin, G., & Hassan, D. (2017). A systematic review of the impact of sport-based interventions on the psychological well-being of people in prison. *Mental Health and Physical Activity*, *12*, 50–61. https://doi.org/10.1016/j.mhpa.2017.02.003

Wurz, A., & Brunet, J. (2016). A Systematic Review Protocol to Assess the Effects of Physical Activity on Health and Quality of Life Outcomes in Adolescent Cancer Survivors. *JMIR Research Protocols, 5*(1). https://doi.org/10.2196/resprot.5383

Table 1

Preregistration and systematic review guideline resources

Pre-register Review Methods	
PROSPERO	https://www.crd.york.ac.uk/prospero/
Open Science Framework	https://osf.io/
Guidelines for Reviews	
Cochrane Handbook for Systematic Reviews of Interventions	www.training.cochrane.org/handbook
Cochrane Qualitative and Implementation Methods Group	https://methods.cochrane.org/qi/welcome
PRISMA	www.prisma.io/
MOOSE	https://doi:10.1001/jama.283.15.2008
ENTREQ	https://doi.org/10.1186/1471-2288-12-181
AMSTAR-2	https://amstar.ca/

Note. PRISMA =Preferred Reporting Items for Systematic Reviews and Meta-Analyses, MOOSE = Meta-analysis Of Observational Studies in Epidemiology, ENTREQ = ENhancing Transparency in REporting the synthesis of Qualitative research, AMSTAR-2 = Assessment of Multiple Systematic Reviews. This list is not exhaustive and only contains examples discussed within the manuscript.

Table 2

Inclusion and exclusion criteria based on PICO and SPIDER tools

Quantitative Research Question: What is the relationship between physical activity and symptoms of depression and anxiety in children?		Qualitative Research Question: What are the experiences of children enrolled in extracurricular physical activity programs?	
Population	Children aged 5-12 years old	**Sample**	Children aged 5-12 years old
Intervention	Different levels of physical activity (e.g., duration, frequency, intensity). Physical activity is defined as any bodily movement produced by skeletal muscles that increase energy expenditure above resting levels (Caspersen, Powell, & Christenson, 1985)	**Phenomenon of Interest**	Extracurricular physical activity programs defined as physical activity programs offered outside of school
Comparison	Sedentary activity, defined as any waking behaviour characterized by an energy expenditure of less than or =1.5 METs while in a sitting, reclining or lying posture (Tremblay et al., 2017)	**Design**	Interview
Outcome	Symptoms of depression and anxiety	**Evaluation**	Experiences

Study Design	Cohort studies or cross-sectional studies	**Research Type**	Qualitative or mixed method