Gold in, gold out. Quality appraisal and risk of bias tools to assess non-intervention studies for systematic reviews in the behavioural sciences: A scoping review

Authors/contributors - affiliation and ORCiD

Lucija Batinović, Department of Behavioural Sciences and Learning, Linköping University, Sweden, 0000-0002-1017-0025

Jade S. Pickering, Independent/unaffiliated, Manchester, UK, 0000-0002-7242-9207

Olmo R. van den Akker, QUEST Center for Responsible Research, Berlin Institute of Health at Charité, Berlin, Germany, 0000-0002-0712-3746

Dorothy Bishop, Department of Experimental Psychology, University of Oxford. 0000-0002-2448-4033

Mahmoud Elsherif, Department of Psychology, University of Birmingham, 0000-0002-0540-3998

Thomas Rhys Evans, School of Human Sciences and Institute for Lifecourse Development, University of Greenwich, UK 0000-0002-6670-0718

Melissa Gibbs, Nuffield Department of Clinical Neurosciences, University of Oxford, UK, 0009-0004-6031-2361

Tamara Kalandadze, Department of Education, ICT and Learning, Østfold University College, Norway, 0000-0003-1061-1131

Janneke Staaks, University of Amsterdam, Netherlands, 0000-0001-8406-3989

Marta Topor, Department of Behavioural Sciences and Learning, Linköping University, Sweden, 0000-0003-3761-392X

ABSTRACT

Systematic reviews depend critically on the methodological quality and bias levels of the studies they synthesise to provide the highest standard of evidence available for informing future research, practice, and policy. Despite the development of extensive methodologies for various fields, current tools may not fully capture the specific needs of the behavioural sciences (broadly defined) where there are unique challenges in assessing both risk of bias and methodological quality of primary studies, particularly in the context of the field's recent

paradigm shift towards more open scholarship to address issues of reproducibility and replicability.

This scoping review aims to map existing tools for assessing methodological quality and risk of bias, including the characteristics of these tools and their applicability to non-intervention quantitative primary studies in the behavioural sciences. In addition to general meta-data from each tool, the review will map out the tools' study design, purpose, item themes and how completing the items can inform the systematic review's conclusions. The review will provide a comprehensive overview of how current tools can be applied to the behavioural sciences, and identify gaps for future development.

ACKNOWLEDGEMENTS

Thank you to Amy Riegelman (University of Minnesota) for assisting with the search strategy. [the section will be updated further at Stage 2 submission]

INTRODUCTION

Systematic reviews are considered one of the highest levels of evidence. They aim to provide a comprehensive overview of the literature to answer specific research questions and are well-established in certain fields (e.g. medicine to create policy changes and recommend best practices in all aspects of healthcare). Through the work of large organisations as well as concerted and pioneering efforts by working groups and individuals, it is now easier than ever to conduct a rigorous systematic review (the Joanna Briggs Institute, JBI, Aromataris et al., 2024; for example guidance for conducting systematic reviews see: the Cochrane Collaboration, Higgins et al., 2019; and MECCIR, The Methods Coordinating Group of the Campbell Collaboration., n.d.) and to thoroughly report all its components~~elements~~ (e.g. Preferred Reporting Items for Systematic Reviews and Meta-Analyses, PRISMA, Page et al., 2021; Non-Intervention, Reproducible, and Open Systematic Reviews, NIRO-SR, Topor & Pickering et al., 2023).

Recently, steps have been taken to provide systematic review guidelines that move away from more traditional intervention-based terminology common in the clinical literature. These guidelines instead focus on terminology more applicable to the field of behavioural sciences (Appelbaum et al., 2018; K. Rogers & Seaborn, 2023; Topor & Pickering et al., 2023). There are, however, still significant hurdles regarding the assessment of methodological quality and/or risk of bias of empirical studies that are synthesised within a systematic review in the quantitative behavioural sciences. Here and throughout, we broadly define the behavioural sciences as inclusive of, but not limited to, psychology (all fields), anthropology, behavioural economics, or sociology. Although the NIRO-SR tool (Topor et al., 2023) aims to emphasise the importance of risk of bias and methodological quality assessment, there is currently no established risk of bias or methodological quality assessment tool specifically for non-intervention studies in behavioural sciences.

The term ‘critical appraisal’ could be considered more as a general process of assessing research for reliability and credibility of evidence (Al-Jundi, 2017), which encompasses assessing the risk of bias and methodological quality. The terms ‘risk of bias’ and ‘methodological quality’ are often used interchangeably and are closely related, although distinct, concepts. Generally, assessing a study’s methodological quality takes a broad stroke approach as to whether the study adheres to the highest standards in the field based on its methodological rigour, generalisability, and applicability of research findings (Armijo-Olivo et al., 2012). For example, whether there were clearly defined aims and hypotheses, whether

the methodology was appropriate to the research question, and whether the sample size was justified by power calculations. On the other hand, risk of bias refers to the likelihood that the study's results are over- or under-estimated due to systematic errors or problems in the methodology and measures the impact (Furuya-Kanamori et al., 2021; Higgins et al., 2019; J. A. C. Sterne et al., 2019); for example whether lack of blinding, the randomisation method, or study attrition have an impact on the results. However, both the study's risk of bias and its methodological quality contribute to the overall picture in a systematic review and, thus, whether the review conclusions are biased overall.

Since risk of bias and (methodological) quality assessment are interdependent in the overall evaluation of a study, it is understandable that these terms are often used interchangeably in the literature. For clarity of the further text, we will also use these terms interchangeably to describe key factors that influence and shape the interpretation of results.

Behavioural scientists ~~Researchers~~ seeking to account for the potential impact of biases must adapt tools that were designed for intervention research, leading to a general lack of consistency in systematic review methodology (e.g., Nitschke et al., 2019). Some other available tools, which could potentially cater for non-intervention designs, are not well-disseminated among behavioural sciences and are published in clinical and medical journals (QuADS, Harrison et al., 2021; Mixed Methods Appraisal Tool, MMAT, Hong et al., 2018; QATSDD, Sirriyeh et al., 2012). Numerous biases, as well as low methodological and reporting quality of individual studies, impede evidence synthesis and our ability to reliably assess the contributions of the literature to accumulated topic knowledge (Munafò et al., 2017). Systematic reviews can only be as good as their foundations. The output, or the conclusions, of a review should accurately reflect the input and take into account the biases ~~quality~~ of the individual study records that are synthesised, the overall biases that may exist in the published literature as a whole, and the beliefs and personal biases of the systematic review authors (Figure 1). In other words: gold in, gold out. ~~Whilst T~~the ~~focus of the current study is to find out what existing tools are available to guide the evaluation of the risk of bias and methodological quality within non-intervention, quantitative, individual studies for systematic reviews in behavioural sciences~~Here we outline these three types of bias that contribute to the evidence within a systematic review; however biases at the level of the individual study is the focus of the current paper, and our goal is to~~.the focus of the current~~ find out what existing tools are available to guide the evaluation of the risk of bias and methodological quality within non-intervention, quantitative, individual studies for systematic reviews in behavioural sciences

Many Different Biases Can Influence Systematic Reviews

Figure 1.

Three types of bias that influence systematic reviews.

Researcher (Reviewer) Bias. When conducting a general review of the literature, all review authors will have unavoidable, pre-existing beliefs about their topic of interest. These beliefs have the potential to influence the systematic review process at all stages. During study selection and evaluation of the evidence, the so-called availability bias could lead researchers to rely on the studies that they are familiar with and which can be easily accessed (Rothstein et al., 2005). Authors may also succumb to confirmation bias and have a tendency to subjectively select studies for inclusion in line with their positions and beliefs about the project (Bishop, 2017). For instance, a recent umbrella review of 24 meta-analyses covering the same topic of the effects of regular physical exercise on cognitive function found a generally low overlap between the studies that were included in each individual meta-analysis (Ciria et al., 2023).

Key questions and inclusion/exclusion criteria are also affected by researcher biases. According to Ciria et al. (2023), none of the included meta-analyses extracted data from all studies that were generated using their search strategy and met their inclusion criteria. Instead, these independent meta-analyses presented a selected sample of studies. A strong belief by the authors that they can accurately judge the existing literature (overconfidence effect, Costa et al., 2017) or that a particular field operates exclusively in a certain way (functional fixedness, Dusink & Latour, 1996), can lead authors to create restrictive questions (e.g., studying colour

blindness in men because of a belief that women cannot be colour blind, Schiötz, 1920). Confirmation bias could also come into play during data synthesis in the form of selective reporting of results regardless of study quality (i.e. cherry-picking), or compromised methodological quality assessment such as presenting only part of a result that supports a particular position (Shamseer et al., 2015). This issue has been referred to as 'paltering' (Rogers et al., 2017), that is, the researcher does not make a false statement, but omits key contextual information. These are just a non-exhaustive subset of the potential biases that could affect the review process and are intended to be illustrative. To circumvent these researcher biases, many existing tools provide guidance on how to develop and register a protocol before conducting a systematic review (e.g., Topor & Pickering et al., 2023; Van Den Akker et al., 2023). The general need for pre-registration and its benefits have been discussed extensively (Munafò et al., 2017; Wagenmakers, & Dutilh, 2016), and the importance of~~for~~ bias control is relevant when designing and conducting systematic reviews. The protocol should encompass the research question, search strategy, screening process, the process for resolving disagreements between systematic reviewers, the data extraction and synthesis strategy, and the method for assessing the risk of bias and ~~/~~quality of each study. Bias introduced from peer-reviewers during the systematic review publication process (e.g., requesting certain literature to be added, post-hoc changes to protocol, etc.) can be minimised with a comprehensive protocol. Recently, the current authors developed the Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR; Topor & Pickering et al., 2023) guidelines for writing a pre-registered protocol (Part A) and reporting the results (Part B)~~. NIRO-SR is more~~ tailored for ~~conducting a systematic review on non-intervention research in the~~ behavioural sciences.

Literature bias. Controlling for researcher bias with a thorough pre-registered protocol can be an effective way of contributing to a systematic review's integrity. However, it does not protect the ~~systematic~~ review outcomes from being influenced by ~~any~~ pre-existing literature bias i.e. when the body of available evidence does not accurately reflect all the research done on a particular topic. Literature bias includes selective reporting of positive findings (reporting bias) and selecting articles with positive findings to be published in a journal (publication bias; McGauran et al., 2010). It ~~(or reporting) bias~~ is difficult to combat as it often stems from the influence of structural issues within academia, such as publication bias. Research is far more likely to be published if the overall findings are statistically significant, even more so when they support the researcher's hypothesis as well (Bertamini & Munafò, 2012; DeVito & Goldacre, 2019; Dickersin, 1990; Dickersin & Min, 1993). The misplaced valuation of high-

impact publications by academic institutions, and of "novel" research by journal reviewers and editors, means that null results have historically been published markedly less than statistically significant and/or novel results, contributing to poor replicability, as evidenced in fields such as cancer research (Begley & Ellis, 2012), ecology (Jennions & Møller, 2003), and psychology (Open Science Collaboration, 2015). This publication bias leads to what Nissen et al. (2016) called the 'canonisation of false facts': unwarranted confidence in published findings.

De Vries et al. (2018) documented how the combination of citation bias (the cumulative effect of researcher biases such as cherry-picking and the availability bias) and publication bias can make an intervention seem far more effective than is actually the case in a clinical context. Publication bias could be reduced by pre-registering study hypotheses, methods, and statistical analyses for each individual study, especially if papers such as Registered Reports are already reviewed, and accepted or rejected before the results are known (Chambers et al., 2015). In general, a structural change is needed to encourage the publication of research regardless of the results and to make 'file drawer' archival data available (Franco et al., 2014; Joober et al., 2012; Lakens, 2019). Publication bias inevitably affects the outcome of a systematic review as it reinforces the impression that the literature is more consistent than is actually the case, and overestimates the size of an effect. A strict systematic search helps reduce the impact of bias within the published literature. However, researchers should always assume publication bias in their systematic review data, evaluate it, and draw conclusions accordingly. This is already a standard part of systematic review methodology and is, therefore, a core component of NIRO-SR, and many other guidelines. In summary, NIRO-SR designed for systematic reviews in behavioural sciences aims to support systematic reviewers in reducing their own researcher bias and in evaluating the impact of literature bias. The missing step is the evaluation of individual study bias, which is the focus of the current work.

Study bias. ~~The incentives that drive academic career progression do not align with the long-term objectives of scientific progress (Chambers et al., 2015). In the current academic culture, a high impact published research paper is expected to present a tidy story, to show no signs of methodological problems, and to report conclusive significant results (Giner-Sorolla, 2012; Kerr, 1998; Nosek et al., 2012; Simmons et al., 2011). The scientific process presents many challenges, and realistically the majority of research projects are unlikely to reach completion without some unforeseen hurdles, or surprising results, but we have not yet culturally normalised the reporting of these events in their entirety. Therefore, we must consider incomplete reporting a likely influence of bias.~~

-Within the lifecycle of individual studies - or primary studies~~, -~~ which might be included in a systematic review, biases can arise at multiple points that may, in turn, be heightened ~~worsened~~ by aforementioned researcher and literature biases. This includes decisions from the early planning stages of a study (e.g., flawed study design) to methods, including data collection (e.g. selection bias, Keiding & Louis, 2018; interviewer bias, West & Blom, 2016), analysis (e.g., inadequate use of statistic tests, analytical flexibility; Simmons et al., 2011), and the writing and publication process (e.g., selective reporting, McGauran et al., 20; citation bias, Gøtzsche 2022; file drawer problem, Munafò et al., 2017). A large body of research, predominantly on clinical trials, discovered that the results of primary studies including the distortion of results, reduced reliability, and generalisability are all influenced~~mediated~~ by~~has focused on the study of how~~ different biases such as performance bias, detection bias, attrition bias, reporting bias, amongst ~~and~~ others ~~can impact~~ the ~~results of primary studies~~ (Lundh & Gøtzsche, 2008; Schulz et al., 1995). ~~They can lead to distortion of results, reduced reliability and generalisability.~~ Crucially, biased studies are then likely to ~~may~~ lead to misleading outcomes in systematic reviews and meta-analyses, which can compound and produce a skewed perspective of the effect of interest despite the systematic reviewers' best intentions (Kvarven et al., 2019).

~~Biased research can also stem from questionable research practices motivated by the perverse incentives in academic culture. These include manipulating the *p* value to ensure it is below the alpha level (usually .05; i.e. p-hacking, Pennington, 2023, p. 130) and therefore "statistically significant", hypothesising after the results are known ('HARKing', Kerr, 1998) and selective reporting (Simmons et al., 2011). Biases are not necessarily the result of a premeditated attempt to mislead on the part of the researchers, but likely a product of the system in which researchers operate or even that they are simply unaware that such research practices are frowned upon. Biases in primary studies can lead to distortion of results, reduced reliability and generalisability. Crucially, biased studies may lead to misleading systematic reviews and meta-analyses, which can in turn compound and produce a skewed perspective of the effect of interest despite the systematic reviewers' best intentions (Kvarven et al., 2019).~~

It is therefore important that systematic reviews evaluate the risk of bias in each primary study. One element of such evaluation is the assessment of the validity of inferences. Each study should present with sound construct, internal, external, and statistical validity (Schiavone et al., 2023). Construct validity relates to the operationalisation of variables and the limitations of the measures used. Internal validity concerns sources of influence, which could alter the effect of interest and include confounding variables, sampling biases and/or insufficient control

conditions. External validity is threatened when authors make far-reaching claims about the observed effect's generalisability in the contexts of people, times, geographical location, settings, stimuli and/or measures used. Statistical validity refers to the appropriateness of the choice of statistical methods and adherence to statistical assumptions. Currently available tools often only assess some aspects of validity, for instance, Cochrane's Risk of Bias tool is predominantly focused on internal validity (Hartling et al., 2009). D'Andrea et al., (2021) evaluated 44 risk of bias tools commonly used in medical research and found that only 34% assessed internal validity, 25% assessed external validity and 34% assessed statistical validity.

Another way to assess ~~the~~ methodological quality~~risk of bias~~ is by paying attention to the rationale behind the study. Scientific inferences are likely to be misleading if project aims are based on flawed and unfounded claims, misinterpretation of previous research and/or narrow view of the literature. In extreme cases, this can result in poorly designed studies and even give rise to pseudoscientific treatments (e.g. Bishop, 2023). Many standard risk of bias tools used in medical and clinical research do not include any items on theoretical background or study aims (e.g., ROBINS-I, Sterne et al., 2016; Newcastle-Ottawa Scale, Wells et al., 2009; QUADAS-2, Whiting, 2011). Others do so to a limited extent (e.g., only aims mentioned in AXIS, Downes et al., 2016; theory and aims mentioned in QuADS, Harrison et al., 2021).

Recently, it has also become common to assess publications for errors that could have introduced bias, for instance, calculation errors. There already exist some tools that help reviewers find different types of errors (e.g. Statcheck by Nuijten & Polanin 2020), but these are not traditionally included as part of the risk of bias or ~~and~~ methodological quality assessment in systematic reviews, although there are some examples of~~. However, it is not uncommon to see~~ systematic review authors using these tools to supplement their bias assessments (see: Heirene et al., 2024; Sparacio et al., 2023).

In most cases, it is impossible to mitigate bias entirely~~the existence of biases is a normal element of scientific work~~, and thus~~,~~ the detection of certain biases should not be the sole indicator of (low) trustworthiness~~of low trustworthinesspoor quality~~. Instead, risk of bias and methodological quality assessment tools should allow for a comprehensive overview of each study allowing for a well-informed judgment of potential problems, their scale, severity and implications. Even more so, these tools should instruct authors of systematic reviews on how to conduct and interpret such assessments. Currently, it is rare to find tools which thoroughly assess primary studies for all types of biases. As most of the available tools come from clinical

fields of research, they - not unexpectedly - place a great deal of importance on patient recruitment and intervention implementation instead (Sterne et al., 2019). Still, even if the right tools were available, assessment of bias and methodological quality would be limited by reporting quality.

The incentives that drive academic career progression do not align with the long-term objectives of scientific progress (Chambers et al., 2015). In the current academic culture, a high impact published research paper is expected to present a tidy story, to show no signs of methodological problems, and to report conclusive significant results (Giner-Sorolla, 2012; Kerr, 1998; Nosek et al., 2012; Simmons et al., 2011). The scientific process presents many challenges, and realistically the majority of research projects are unlikely to reach completion without some unforeseen hurdles, or surprising results, but we have not yet culturally normalised the reporting of these events in their entirety. For instance, a recent study looking at the adherence to pre-registered protocols found that only two out of 27 projects did not deviate from the initial protocol. Out of the remaining 25, nine did not provide a clarification for the deviations (Claesen et al., 2021). Therefore, we must consider incomplete reporting a likely influence of bias.

Lastly, it is worth noting that biased research can also stem from questionable research practices which can often be motivated by the perverse incentives in academic culture. These include manipulating the p-value to ensure it is below the alpha level (usually .05; i.e. p-hacking, Pennington, 2023, p. 130) and therefore "statistically significant", hypothesising after the results are known ('HARKing', Kerr, 1998) and selective reporting (Simmons et al., 2011). Whilst it is often challenging to detect questionable research practices, ~~T~~this evaluation is made easier when the primary studies show high reporting quality, engage in open research practices and abide by integrity standards (e.g. state conflicts of interest). It is desired~~,~~ that each study should provide in-depth details on their precise methodology, data analysis (including decisions made and analysis scripts), and full reporting of planned hypotheses regardless of results. Although i~~I~~t is not currently expected that each record included in a systematic review should be assessed for adherence to the protocol or for p-hacking if these additional materials are available~~,. However,~~ this level of transparency allows researchers in the field to verify the claims in the study and report inconsistencies present due to honest errors or questionable practices. For example, it is possible that some of these issues are picked up during the peer

review process and resolved prior to publication. Therefore, even though the implementation of open and reproducible research practices is not a foolproof solution to avoid bias, it is a sign of good practice, which allows for further scrutiny and verification. For this reason, risk of bias tools should include items on open and reproducible research practices. It is not common to see this in the traditional risk of bias tools given the recent nature of the open research paradigm shift and, similar to the statistics check tools, systematic review authors increasingly resort to adding items relating to open scholarship to their risk of bias assessments (see examples: Cooper et al., 2022; O'Daffer et al., 2022). There also exist tools, which aim to guide authors of primary studies on the use of open and reproducible research practices to increase the utility of their studies in future systematic reviews (see examples: Chow et al., 2023; Fernández-Castilla et al., 2024). Finally, discussions in the field have recently turned to the problematic primary studies included in systematic reviews specifically on the grounds of integrity concerns and possible falsifications. At least one tool is being developed to support study assessments in future systematic reviews (Wilkinson et al., 2024).

Aim

Our aim is to identify and establish the current limits of the full range of different tools that are currently available for the critical appraisal, and assessment of risk of bias and methodological quality for primary studies. ~~TAlthough the term "critical appraisal" could be considered more as a general process of assessing research for reliability and credibility of evidence (Al-Jundi, 2017), which encompasses assessing the risk of bias and methodological quality, we acknowledge some tools may be called critical appraisal tools. The terms 'quality assessment' and 'risk of bias' are often used interchangeably, but they can and do have distinct definitions. Here, we refer to bias as any factor (such as methodological decisions or flaws in design, conduct, or analysis) that may affect the validity or reliability of the results and conclusions of a study. To assess bias, a methodological quality assessment *summarises* the methodological rigour, generalizability, and applicability of research findings (Armijo-Olivo et al., 2012). Finally, a risk of bias assessment focuses on the potential *impact* of the methodological quality of the paper and the likelihood of this leading to an over- or under-estimation of the true effect (Furuya-Kanamori et al., 2021; Higgins et al., 2019; NHMRC, 2019; J. A. C. Sterne et al., 2019)~~. Our research question is therefore:

"Which existing, available tools are suitable to assess the methodological quality and risk of bias of non-intervention primary studies included in evidence syntheses within the behavioural sciences?"

We will evaluate the tools' relevance to non-intervention quantitative study designs, their quantification of bias, and practical guidance for using the assessment outcomes in the ensuing evidence synthesis. Particular attention will be given to items which assess a study's openness and reproducibility, such as open data and open access practices, as well as integrity checks like funding sources and ethics approval. The scoping review will conceptually summarise the characteristics of existing tools and identify the gaps and directions for further developments of guidelines. A preliminary search of the literature (8th January 2024) suggested that no systematic or scoping review had been conducted on this topic before. A search of Scopus, Epistemonikos, Prospero, Open Science Framework (OSF) Registries, and OSF Projects yielded a total of 17 results, none of which were a systematic or scoping review answering the same research question as we propose here (full search and results available on the OSF).

METHODS

We will follow the Joanna Briggs Institute (JBI) methodology for conducting systematic scoping reviews (M. D. J. Peters et al., 2020), which is compatible with the PRISMA Scoping Reviews Checklist, and report our findings according to PRISMA-Scr reporting standards (Tricco et al., 2018). The study is designed and conducted as a Registered Report. The Stage 1 Registered Report can be found at [link].

Inclusion and Exclusion Criteria

To determine the inclusion and exclusion criteria we mapped out the concept and context of the scoping review as per the PCC (Population, Concept, Context) framework in the JBI Reviewers' Manual (Peters et al., 2020). As we are not interested in a specific population we did not include this element, and we also clarified the evidence sources that we were interested in.

Concept: We aim to map existing risk of bias, critical appraisal, and methodological quality assessment tools that are relevant to the field of behavioural science, regardless of their intended application. We want to know what tools already exist, what features they have, how

they apply to different methodologies (e.g., non-intervention) and how the methodological quality of primary research is quantified by these tools.

Context: The tools must be relevant to behavioural sciences, either because they were specifically created for the behavioural sciences, or because they are sufficiently broad/generic that they could be consistently applied in this domain.

Types of evidence sources: The entire tool may be published in largely any format as long as it is available and accessible to the research team, either open access or through university libraries. This can include, but is not limited to: journals, pre-prints, dissertations/theses, websites, downloadable documents, book chapters, or in manuals. We will only include papers/tools available in English for feasibility reasons, but we will place no restrictions on the publication date of included records.

The full screening procedure is available on the OSF, and through a two-step screening process (title and abstract, followed by full-text) we will include records from the systematic search that fulfil all of the following criteria:

1) Introduce, present, evaluate, validate, translate or update a checklist/guidelines/list of items for assessing risk of bias, performing critical appraisal, or determining the methodological quality of primary quantitative research reports. Usually, the research record will be a journal article providing an account of how the tool was developed, validation of the tool, or a tutorial for how to use it in practice, but we are not prescriptive as to the specific format of the research record.
2) Are relevant to the behavioural sciences (inclusive of, but not limited to, psychology (all fields), anthropology, behavioural economics, or sociology), either because they were specifically created for the behavioural sciences, or because they are sufficiently broad/generic that they could be consistently applied in this domain.
3) Are written as one of the following: i) an account of how the tool was developed, ii) validation of the tool, iii) a tutorial for how to use the tool in practice.
4) Allow us to access a tool, for example within the paper, online, or in the paper's supplementary materials.
5) Written in English.

We will exclude records from the systematic search that fulfil any of the following criteria:

1) Only report a systematic review or a meta-analysis that simply used a relevant tool as part of the methodology (unless it also includes a new tool itself).
2) Describe/use a tool that is designed for qualitative studies.
3) Describes a tool that is not findable or accessible.
4) It is clear from the items that it is only applicable to a topic outside of behavioural sciences and cannot be used more widely in behavioural science research.

Both excluded and included records will be available on OSF as separate reference files, together with original search files.

Search Strategy

We will perform searches in the following databases: Medline (Ovid), PsycINFO (Ovid), and Web of Science Core Collection.

An example search strategy in MEDLINE:

1. *bias/ OR (quality assessment OR risk of bias OR critical* apprais* OR quality of evidence OR evidence quality OR methodological quality OR appraisal tool* OR quality appraisal).ti. OR (risk of bias).ab./freq=3*
2. *checklist/ OR (checklist* OR tool* OR list OR criteria OR scale OR instrument OR worksheet*).ti.*
3. *(systematic review* OR meta-analy* OR ((develop* OR evaluat* OR improv* OR reliab* OR valid* OR consistency OR feasab* OR utility OR usabil*) ADJ7 (checklist* OR tool* OR list* OR worksheet*))).ti,ab,kf.*
4. *1 AND 2 AND 3*

Key: / = medical subject heading (MeSH), ti = title, ab = abstract, kf = author supplied keywords, ADJn = word distance of maximum n words, /freq=n = occurrence of a search term of at least n times

The search criteria was developed and validated by two research librarians (a co-author JS supported by AR mentioned in the acknowledgements) with a reference set of articles (see Appendices). The reference set was selected as known examples of papers that published tools relevant to the current project. We confirmed that all reference set articles were successfully

identified by our search. A pilot search was performed on 28th May 2024, which yielded 1.341 results; Medline (718 results), PsycINFO (75 results), and Web of Science Core Collection (548 results).

The full search will be performed after receiving In Principle Acceptance of the Stage 1 Registered Report, and updated two years later if the Stage 2 report has not yet been submitted for review. The full search strategy is available on the OSF. The nature of scoping reviews means that the search strategy process is iterative as we become more familiar with the evidence base (Aromataris et al., 2024). If we feel the search strategy can benefit from improvements, this will be done by a research librarian (JS) and the process will be transparently documented.

We will also perform a manual backward reference search in the reference lists of papers whose full text met the inclusion criteria ~~during the screening process~~. When performing the reference searches, records will be extracted if i) not already identified ii) the title suggests an introduction, evaluation, validation or update of a relevant tool. To supplement the search further, we will also search the RRID website, which includes databases of relevant tools (three separate searches of the following phrases with no filters applied: "risk of bias", "critical appraisal", "quality assessment").

From the retrieved papers, reviewers conducting the screening will tag publications that present a number of relevant tools, for instance, in a systematic review of risk of bias tools. These records will not qualify for data extraction but will be tagged and used for further manual search. At this stage, each tool mentioned will be checked and its reference will be extracted if i) the tool is properly cited, ii) the record is not already identified iii) the title suggests an introduction, evaluation, validation or update of a relevant tool. As the goal of this scoping review is to discover what tools are available for use, we will not be contacting authors of primary sources for details of their tools if we cannot find them ourselves, as this does not fit the criteria of "available" within this context.

Source of evidence selection

Note that part of this process has already taken place for the search conducted in May 2021 which was previously pre-registered and where the project reached the data extraction stage. The updated search will follow the same process. Search results will be imported into RStudio

and the duplicates will be removed using the *revtools* and *synthesisr* packages (Westgate, 2019).

We will use Rayyan (Ouzzani et al., 2016) for the screening process. At each stage of the screening process (titles and abstracts; full text) two independent reviewers from the research team will follow the screening instructions (see https://osf.io/fndsc), and any discrepancies will be resolved through a discussion between the two reviewers, or by a third reviewer from the team when consensus cannot be reached. These reviewers may not be the same individuals for each record. For clarity, we define 'full text' here to mean not only the published record as identified by the search, but associated supplementary material including the tool itself, which may or may not be accessible in the original paper.

We have already conducted a pilot screening process using a random selection of 25 records from the aforementioned pilot search. Each of these records was independently screened against the eligibility criteria specified in the screening instructions by two separate reviewers. As suggested by the JBI, an agreement of <75% warrants modifications of the eligibility criteria, whereas an agreement of >=75% suggests that the screening process is ready to be started. We had an agreement rating of 92% (i.e. these records were rated identically by each reviewer). The discrepancies were due to vague wording in the element under the title and abstract section, which led to ambiguity about paper format eligibility. We modified this in the instructions following a team discussion.

[Prisma style flow chart will go here]

Data Extraction

Before the data extraction begins, all identified~~screened~~ tools will be checked for possible updates of included tools. In case the dataset contains multiple versions of a tool, we will merge the records and only extract data from the latest version. If there are multiple versions of the same tool which assess different sub-fields or have different purposes, they will be treated as unique records. Data extraction will be performed in R (R Core Team, 2021) using the *metabefor* package (Peters, 2022). This package, developed as a precursor to the *metafor* package for meta-analyses (Viechtbauer, 2010), ensures transparent and reproducible data extraction. For each record, extracted data are arranged in an RMarkdown script ensuring standardised formatting of the data and enhancing reproducibility and machine-readability.

Subsequently, the tools will be consolidated into a data frame, and both the data frame and the markdown files will be accessible on Github and OSF repositories. The associated R files consist of a script that initiates the creation of the template and aggregation of tools into a data frame, as well as the markdown extraction template. Details of the data extraction items, the markdown extraction template, and instructions that have been piloted by two of the authors (LB and MT) are available on the OSF.

The two authors (LB and MT) conducted independent extractions (pilot extractions available on Github) of the same tool, which was deemed extensive enough to find potential issues in the data extraction process. The pilot extractions were then compared to check for rater consistency. The comparison indicated consensus on most items, yet there remained discrepancies on items that require nuanced interpretation (e.g., open science practices or items assessing validity). The discrepancies mostly emerged from the unclear statements in the tool (e.g., how we should count the number of items in the tool, or extract the data about support). To avoid subjective assumptions, we have outlined detailed guidelines in the extraction instructions. To further reduce any potential misunderstandings, the extraction procedure will be as follows: each tool will be extracted simultaneously by two reviewers who will discuss discrepancies and resolve them as they occur. ~~items from each tool will be extracted by one reviewer and items marked with an asterisk will be independently validated by a second reviewer, who will focus solely on these tagged items. Conflicts will be discussed between the two reviewers and, if necessary, resolved by a third reviewer.~~

We are particularly interested in what the tools suggest - if anything - systematic reviewers should do to interpret the final outcome of each record rated with the tool. We will note how many tools provide guidance on interpreting the overall outcomes of the ratings as well as the consequential implications for the conclusions of the systematic review, such as how to weigh poor methodological quality or high risk of bias papers differently to those of higher methodological quality and lower risk. We will additionally look at how many of the tools provide some form of overall quantitative scoring or rating system, such as that researchers can quickly identify the highest methodological quality or highest risk of bias records in their systematic review.

More specifically, we will extract items pertaining to the following domains: metadata and content. Items included in the metadata domain will describe the tool title, source and access information, type and format of the tool and available support channels for the tool's use. The

Content domain focuses on the items that evaluate the tool itself, i.e., number of items in the tool, presence of usage instructions or study designs it intends to assess. Additionally, we will extract information about the areas that the tool assesses, such as validity, integrity, open and reproducible scholarship, as well as information about scoring systems and existence of interpretation guides. A complete list of items and detailed description of each one is available on OSF: https://osf.io/ewm7x.

Due to the iterative nature of data extraction, further data items may be added to the extraction sheet during the extraction process, in which case, this will be transparently marked as "ad hoc items". Depending on when the new items get introduced, we will retroactively extract information for these items from all previously assessed tools.

Analysis of the evidence

We will descriptively map the extracted data from the included records and provide an overview of existing tools categorised as "risk of bias", "quality assessment" or "critical appraisal" tools (depending on how the creators of the tool classified them). We will also provide a narrative summary about the usability of these tools, i.e. whether the tool includes detailed information and instructions to aid the researcher to make a decision on the rating of each item, as well as practical aids~~steps~~ for rating items such as pre-built~~by providing some form of~~ completion forms/checklists (whether for online or offline use). Additionally, we will provide a summary of tools that could be appropriate for the assessment of non-intervention designs.

Finally, we will report the type of content that the tools encourage the systematic reviewers to assess in the original research record, such as the validity and issues of open and reproducible scholarship as this will help to map any gaps in the available tools that may have arisen over recent years within the context of the credibility revolution (Vazire, 2018).

Data visualisation will be decided upon once the results are obtained and will depend on the complexity and richness of the data to ensure a clear presentation of the results.

Al-Jundi, A. (2017). Critical Appraisal of Clinical Research. *Journal Of Clinical and*

Diagnostic Research. https://doi.org/10.7860/JCDR/2017/26047.9942

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Armijo-Olivo, S., Stiles, C. R., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *Journal of Evaluation in Clinical Practice*, *18*(1), 12–18. https://doi.org/10.1111/j.1365-2753.2010.01516.x

Armijo-Olivo, S., Stiles, C. R., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: Methodological research. *Journal of Evaluation in Clinical Practice*, *18*(1), 12–18. Scopus. https://doi.org/10.1111/j.1365-2753.2010.01516.x

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-24-01

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Bertamini, M., & Munafò, M. R. (2012). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science*, *7*(1), 67–71. https://doi.org/10.1177/1745691611429353

Bishop, D. V. M. (2017, December 4). *Standing on the shoulders of giants: Why your literature review should be systematic*. https://osf.io/uk92g

Bishop, D. V. M. (2023, November 25). Low-level lasers. Part 1. Shining a light on an

unconventional treatment for autism. *BishopBlog*. https://deevybee.blogspot.com/2023/11/low-level-lasers-part-1-shining-light.html

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

Chow, J. C., Sandbank, M., & Hampton, L. H. (2023). Guidance for Increasing Primary Study Inclusion and the Usability of Data in Meta-Analysis: A Reporting Tutorial. *Journal of Speech, Language, and Hearing Research*, *66*(6), 1899–1907. https://doi.org/10.1044/2023_JSLHR-22-00318

Ciria, L. F., Román-Caballero, R., Vadillo, M. A., Holgado, D., Luque-Casado, A., Perakakis, P., & Sanabria, D. (2023). An umbrella review of randomized control trials on the effects of physical exercise on cognition. *Nature Human Behaviour*, *7*(6), 928–941. https://doi.org/10.1038/s41562-023-01554-4

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Cooper, S. E., Van Dis, E. A. M., Hagenaars, M. A., Krypotos, A.-M., Nemeroff, C. B., Lissek, S., Engelhard, I. M., & Dunsmoor, J. E. (2022). A meta-analysis of conditioned fear generalization in anxiety-related disorders. *Neuropsychopharmacology*, *47*(9), 1652–1661. https://doi.org/10.1038/s41386-022-01332-2

Costa, D. F., De Melo Carvalho, F., De Melo Moreira, B. C., & Do Prado, J. W. (2017). Bibliometric analysis on the association between behavioral finance and decision making with cognitive biases such as overconfidence, anchoring effect and confirmation bias. *Scientometrics*, *111*(3), 1775–1799.

https://doi.org/10.1007/s11192-017-2371-5

D'Andrea, E., Vinals, L., Patorno, E., Franklin, J. M., Bennett, D., Largent, J. A., Moga, D. C., Yuan, H., Wen, X., Zullo, A. R., Debray, T. P. A., & Sarri, G. (2021). How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. *BMJ Open*, *11*(3), e043961. https://doi.org/10.1136/bmjopen-2020-043961

De Vries, Y. A., Roest, A. M., De Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: The case of depression. *Psychological Medicine*, *48*(15), 2453–2455. https://doi.org/10.1017/S0033291718001873

DeVito, N. J., & Goldacre, B. (2019). Catalogue of bias: Publication bias. *BMJ Evidence-Based Medicine*, *24*(2), 53–54. https://doi.org/10.1136/bmjebm-2018-111107

Dickersin, K. (1990). The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA: The Journal of the American Medical Association*, *263*(10), 1385. https://doi.org/10.1001/jama.1990.03440100097014

Dickersin, K., & Min, Y. I. (1993). NIH clinical trials and publication bias. *The Online Journal of Current Clinical Trials*, *50*.

Downes, M. J., Brennan, M. L., Williams, H. C., & Dean, R. S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*, *6*(12), e011458. https://doi.org/10.1136/bmjopen-2016-011458

Dusink, L., & Latour, L. (1996). Controlling functional fixedness: The essence of successful reuse. *Knowledge-Based Systems*, *9*(2), 137–143. https://doi.org/10.1016/0950-7051(95)01025-4

Fernández-Castilla, B., Said-Metwaly, S., Kreitchmann, R. S., & Van Den Noortgate, W. (2024). What do meta-analysts need in primary studies? Guidelines and the SEMI

checklist for facilitating cumulative knowledge. *Behavior Research Methods*, *56*(4), 3315–3329. https://doi.org/10.3758/s13428-024-02373-9

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Furuya-Kanamori, L., Xu, C., Hasan, S. S., & Doi, S. A. (2021). Quality versus Risk-of-Bias assessment in clinical research. *Journal of Clinical Epidemiology*, *129*, 172–175. https://doi.org/10.1016/j.jclinepi.2020.09.044

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science*, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Harrison, R., Jones, B., Gardner, P., & Lawton, R. (2021). Quality assessment with diverse studies (QuADS): An appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. *BMC Health Services Research*, *21*(1), 144. https://doi.org/10.1186/s12913-021-06122-y

Hartling, L., Ospina, M., Liang, Y., Dryden, D. M., Hooton, N., Krebs Seida, J., & Klassen, T. P. (2009). Risk of bias versus quality assessment of randomised controlled trials: Cross sectional study. *BMJ*, *339*(oct19 1), b4012–b4012. https://doi.org/10.1136/bmj.b4012

Heirene, R., LaPlante, D., Louderback, E., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. (2024). Preregistration specificity and adherence: A review of preregistered gambling studies and cross-disciplinary comparison. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2021.2909

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions* (1st ed.).

Wiley. https://doi.org/10.1002/9781119536604

Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the MIXED METHODS APPRAISAL TOOL (MMAT). *Journal of Evaluation in Clinical Practice*, *24*(3), 459–467. https://doi.org/10.1111/jep.12884

Jennions, M. D., & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, *14*(3), 438–445. https://doi.org/10.1093/beheco/14.3.438

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience*, *37*(3), 149–152. https://doi.org/10.1503/jpn.120065

Keiding, N., & Louis, T. A. (2018). Web-Based Enrollment and Other Types of Self-Selection in Surveys and Studies: Consequences for Generalizability. *Annual Review of Statistics and Its Application*, *5*(1), 25–47. https://doi.org/10.1146/annurev-statistics-031017-100127

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kvarven, A., Strømland, E., & Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lundh, A., & Gøtzsche, P. C. (2008). Recommendations by Cochrane Review Groups for

assessment of the risk of bias in studies. *BMC Medical Research Methodology*, *8*(1), 22. https://doi.org/10.1186/1471-2288-8-22

McGauran, N., Wieseler, B., Kreis, J., Schüler, Y.-B., Kölsch, H., & Kaiser, T. (2010). Reporting bias in medical research—A narrative review. *Trials*, *11*(1), 37. https://doi.org/10.1186/1745-6215-11-37

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

NHMRC. (2019, September 29). *Guidelines for Guidelines: Assessing risk of bias.* https://www.nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, *5*, e21451. https://doi.org/10.7554/eLife.21451

Nitschke, F. T., McKimmie, B. M., & Vanman, E. J. (2019). A meta-analysis of the emotional victim effect for female adult rape complainants: Does complainant distress influence credibility? *Psychological Bulletin*, *145*(10), 953–979. https://doi.org/10.1037/bul0000206

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

O'Daffer, A., Colt, S. F., Wasil, A. R., & Lau, N. (2022). Efficacy and Conflicts of Interest in Randomized Controlled Trials Evaluating Headspace and Calm Apps: Systematic Review. *JMIR Mental Health*, *9*(9), e40924. https://doi.org/10.2196/40924

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

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Pennington, C. R. (2023). *A student's guide to open science: Using the replication crisis to reform psychology*. Open University Press.

Peters, G. (2022). *metabefor: Modular, Extensible, Transparent, Accessible, Bootstrapped Extraction for Systematic Reviews.* (Version 0.3.0.) [R package].

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping reviews. In *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-12

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rogers, K., & Seaborn, K. (2023). The Systematic Review-lution: A Manifesto to Promote Rigour and Inclusivity in Research Synthesis. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3544549.3582733

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456–473.

https://doi.org/10.1037/pspi0000081

Schiavone, S. R., Quinn, K. A., & Vazire, S. (2023). *A Consensus-Based Tool for Evaluating Threats to the Validity of Empirical Research*. PsyArXiv. https://doi.org/10.31234/osf.io/fc8v3

Schiötz, I. (1920). COLOUR BLIND FEMALES: THE INHERITANCE OF COLOUR BLINDNESS IN MAN (concluded). *The British Journal of Ophthalmology*, *4*(9), 393–403. https://doi.org/10.1136/bjo.4.9.393

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical Evidence of Bias: Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. *JAMA*, *273*(5), 408–412. https://doi.org/10.1001/jama.1995.03520290060030

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & the PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *349*(jan02 1), g7647–g7647. https://doi.org/10.1136/bmj.g7647

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice*, *18*(4), 746–752. https://doi.org/10.1111/j.1365-2753.2011.01662.x

Sparacio, A., Ropovik, I., Jiga-Boy, G., Cem Lağap, A., & IJzerman, H. (2023). Stress Regulation via Being in Nature and Social Support in Adults, a Meta-analysis. *Collabra: Psychology*, *9*(1), 77343. https://doi.org/10.1525/collabra.77343

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, l4898. https://doi.org/10.1136/bmj.l4898

Sterne, J., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, i4919. https://doi.org/10.1136/bmj.i4919

The Methods Coordinating Group of the Campbell Collaboration. (n.d.). *Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards.* Retrieved 20 May 2024, from https://onlinelibrary.wiley.com/page/journal/18911803/homepage/author-guidelines

Topor, M., Pickering, J. S., Barbosa Mendes, A., Bishop, D. V. M., Büttner, F., Elsherif, M. M., Evans, T. R., Henderson, E. L., Kalandadze, T., Nitschke, F. T., Staaks, J. P. C., Van Den Akker, O. R., Yeung, S. K., Zaneva, M., Lam, A., Madan, C. R., Moreau, D., O'Mahony, A., Parker, A. J., … Westwood, S. J. (2023). An integrative framework for planning and conducting Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR). *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2021.2840

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR):

Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Van Den Akker, O. R., Peters, G.-J. Y., Bakker, C. J., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Moreau, D., Nordström, T., Pickering, J. S., Riegelman, A., Topor, M. K., Van Veggel, N., Yeung, S. K., Call, M., Mellor, D. T., & Pfeiffer, N. (2023). Increasing the transparency of systematic reviews: Presenting a generalized registration form. *Systematic Reviews*, *12*(1), 170. https://doi.org/10.1186/s13643-023-02281-7

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the **metafor** Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Wagenmakers, E. J., & Dutilh, G. (2016, October 31). Seven Selfish Reasons for Preregistration. *Association for Psychological Science.* https://www.psychologicalscience.org/observer/seven-selfish-reasons-for-preregistration

Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2009). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. The Ottawa Hospital Research Institute. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

West, B. T., & Blom, A. G. (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, smw024. https://doi.org/10.1093/jssam/smw024

Westgate, M. J. (2019). revtools: An R package to support article screening for evidence

synthesis. *Research Synthesis Methods*, *10*(4), 606–614. https://doi.org/10.1002/jrsm.1374

Whiting, P. F. (2011). QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine*, *155*(8), 529. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wilkinson, J., Heal, C., Antoniou, G. A., Flemyng, E., Alfirevic, Z., Avenell, A., Barbour, G., Brown, N. J. L., Carlisle, J., Clarke, M., Dicker, P., Dumville, J. C., Grey, A., Grohmann, S., Gurrin, L., Hayden, J. A., Heathers, J., Hunter, K. E., Lasserson, T., … Kirkham, J. J. (2024). Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions. *BMJ Open*, *14*(3), e084164. https://doi.org/10.1136/bmjopen-2024-084164

APPENDICES

Reference set of articles

Medline accession numbers (PMIDs): (9764259 OR 27733354 OR 26993202 OR 26092286 OR 28935701 OR 12956787 OR 27932337).ui.

1. Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of Fields (https://doi.org/10.7939/R37M04F16)
2. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. (http://dx.doi.org/10.1136/jech.52.6.377)
3. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions (https://doi.org/10.1136/bmj.i4919)
4. Methodological index for non-randomized studies (minors):development and validation of a new instrument http://cobe.paginas.ufsc.br/files/2014/10/MINORS.pdf
5. Downes MJ, Brennan ML, Williams HC, et al. Development of a critical appraisal tool to assess the quality of crosssectional studies (AXIS). BMJ Open 2016;6:e011458. doi:10.1136/bmjopen-2016-011458 http://bmjopen.bmj.com/content/bmjopen/6/12/e011458.full.pdf

6. Rosella L, Bowman C, Pach B, Morgan S, Fitzpatrick T, Goel V. The development and validation of a meta-tool for quality appraisal of public health evidence: Meta Quality Appraisal Tool (MetaQAT). Public Health. 2016;136:57-65. Available from: http://www.publichealthjrnl.com/article/S0033-3506(15)00437-0/abstract
7. Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE) risk of bias assessment tool for assessing animal studies http://www.ncbi.nlm.nih.gov/pubmed/24667063
8. Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., ... & Churchill, R. (2016). ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of clinical epidemiology, 69, 225-234. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687950/
9. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017; 358 doi: https://doi.org/10.1136/bmj.j4008 (Published 21 September 2017) Cite this as: BMJ 2017;358:j4008