# Guidance on Conducting and Reviewing Systematic Reviews (and Meta-Analyses) in Work and Organizational Psychology

Kevin Daniels

Editor, European Journal of Work and Organizational Psychology

**Acknowledgements**: For feedback on earlier drafts, I am very grateful to Sara Connolly, Cigdem Gedikli, James Rumbold, Olga Tregaskis and David Watson.

To appear in *European Journal of Work and Organizational Psychology*. This is not the copy of record.

Abstract

Systematic reviews and meta-analyses are means of summarizing and synthesizing research evidence in a given topic area. They can be used: to define the current state of knowledge and how confident we can be in that knowledge; to identify evidence gaps; and to provide recommendations for policy and practice based on the best available evidence. At EJWOP, our editorial stance is explicitly to encourage the conduct of systematic reviews and meta-analyses. The purpose of this editorial is to provide some guidance to authors and journal referees on the (technical) features of good systematic reviews.

*Keywords*: Systematic review methods; work and organizational psychology; evidence based decision making.

Before embarking on the main part of this editorial, I must first extend great thanks from all who work on EJWOP to Ana Cristina Costa for her outstanding work over the last seven or so years as an associate editor of the journal. Although we'll miss her as an associate editor, I am pleased to say that Ana Cristina will continue to work with the editorial team as a member of the editorial board. As ever as editor, I am grateful for the support of the editorial team: Ans de Vos, Beatrice van der Heijden, Despoina Xanthopoulou, Eva Demerouti, Florian Kunze, Karen Niven, Roni Reiter Palmon and Simon de Jong.

In two recent EJWOP editorials (Daniels, 2016, 2017), I have noted the benefits of systematic review and meta-analytic methods for research and practice in work and organizational psychology (WOP), and the editorial of 2017 in particular indicated a specific interest in encouraging the conduct and reporting of systematic reviews and meta-analyses. The main purpose of this editorial is to provide some guidance to authors and journal referees on the (technical) features of good systematic reviews and to act as a starting point for those new to reviewing to gather more detailed information. In so far that the principles are the same for meta-analyses and systematic reviews on deciding on research questions, search strategies, inclusion and exclusion criteria for studies and deciding on the quality of available evidence, then the guidance provided here is relevant to the non-statistical elements of meta-analyses.[1]

The term systematic review refers to approaches that collate and synthesize available research on a pre-specified research question that fits pre-specified eligibility criteria. By doing so, systematic reviews aim to minimize bias in selecting studies that do or do not support a specific prediction, expectation or position. Systematic reviews use explicit and systematic methods to find, sift, sort and summarize findings. By making the research

[1] There exist other sources on the technical aspects of meta-analyses, e.g. Cooper (2010) and Field and Gillett (2010).

questions, eligibility criteria and methods explicit before commencing the review process, systematic review methods aim to minimize bias, enable reproducibility and replication of the review and therefore produce more robust and reliable conclusions in respect of a given question. Systematic reviews can also be extended into meta-analyses, which are statistical summaries of evidence in relation to a specific research question (for definitions of systematic reviews, see e.g. Green, Higgins, Alderson, Clarke, Mulrow & Oxman, 2011; Hodgkinson & Ford, 2014; Snape, Meads, Bagnall, Tregaskis, Mansfield & MacLennan, 2017). Systematic review approaches are not confined to methods that simply summarize the evidence for and against a particular hypothesis or theory, but can encompass systematic searches and synthesis of theoretical accounts in any given area. Systematic reviews and meta-analyses are also accompanied in many cases by a narrative evidence synthesis that interprets research findings. Therefore reviews can be used to test or build theory, or both in some cases.

There are two main reasons why we are keen to encourage the conduct and reporting of systematic reviews at EJWOP.

First, as no single study is perfect, then it is problematic to rely on evidence from a single study for a definitive answer to a question.[2] Instead, systematic reviews of studies that report data relevant to any given research question can help identify regularities across multiple studies, rule out capitalization on chance or error in single studies, provide external validity for and/or boundary conditions on findings and provide clues on potential moderators where boundary conditions are established. When effects are not replicable or inconsistent, systematic reviews also bring to light important gaps in our understanding. Importantly, if the right data are extracted from studies and synthesized in reviews, it is possible to examine

[2] Of course, no single systematic review is perfect either, but systematic reviews, if conducted well, will be less imperfect than the primary studies they are based on.

whether any relationships found come about because of expected theoretical processes (e.g. reviewing studies that incorporate mediators, in meta-analyses, this would involve meta-structural equations modelling). If the review is a systematic review of controlled intervention studies, then the review also provides strong statements of causality: Replicable effects from similar types of interventions provide some of the best ecologically valid evidence of cause and effect – especially if the interventions are subject to multiple and replicated randomized controlled trails. Establishing causal relations, the plausibility of different theoretical processes and boundary conditions all help to provide us with evidence on whether our theories are reasonable approximations of how the world works.

Second, with concerns over the relevance of the research we do (Grote, 2017), systematic review methods help provide us with a means of making our research more relevant by summarizing evidence in a given area to help evidence-based or evidence-informed decision making (especially where that evidence is provided by intervention studies). Systematic reviews by and of themselves will not narrow the practice/research gap (cf. Rynes, Colbert & O'Boyle, 2018), but accumulated and synthesized evidence does form part of the evidence eco-system (see Shepherd, 2014). Accumulated and synthesized evidence then needs 'evidence translators'[3] to create amongst users a demand for evidence and capability to use evidence as well as provide and advertise summarized and synthesized evidence to users. For work and organizational psychology researchers, our primary users (practising work and organizational psychologists) are intermediaries between research and the ultimate beneficiaries of our work (various organizational and societal stakeholders). Our primary users are scientifically literate and want to use evidence (Bartlett & Francis-Smythe,

[3] It cannot be assumed that those with skill in conducting primary research or systematic review methods also possess the skills, other resources or motivation to translate for non-scientific audiences, and being able to translate research into a form useful and appealing to users may require a range of skills and knowledge beyond pedagogy and being able to write in non-technical language. Such skills and knowledge can include graphic design, animation, film making.

2016) to benefit ultimate end users (see Carter, 2018 on how work and organizational psychology practice can reach multiple audiences). Advanced/expert work and organizational practitioners may be able to understand systematic reviews and bivariate effect sizes derived from meta-analyses without much translation from scientific/technical writing to more accessible writing. Notwithstanding , the level of translation required from technical reports of systematic reviews and meta-analyses for most work and organizational psychology practitioners may be less than that required for translating primary research findings to the ultimate beneficiaries of our research.

In recent years, systematic reviews, reviews using formal search criteria - if not using full blown systematic review methods - and meta-analyses published in EJWOP form an important component of papers that we publish (see Boer, Deinert, Homan & Voelpel, 2016; Bouckenooghe, Schwarz & Minbashian, 2015; Bozer & Jones, 2018; Harari, Reaves & Viswesvaran, 2016; Joseph, Walker & Fuller-Tyszkiewicz, 2018; Kröll, Doebler & Nüesch, 2017; Lomas, Medina, Ivtzan, Rupprecht, Hart & Eiroa-Orosa, 2017; Maynard, Kennedy & Sommer, 2015; Monteiro, Pinto & Roberto, 2016; Posthuma, Campion & Campion, 2018; Potočnik & Anderson, 2016; Samsudin, Isahak & Rampal, 2018; Watson, Tregaskis, Gedikli, Vaughn & Semkina, 2018). Of these papers, six were systematic reviews, three were meta-analyses and four were reviews with some level of systematic structure. Reviews related to wellbeing formed the biggest cluster of reviews (five systematic reviews, one meta-analysis).

Snape et al. (2017) have developed guidance on how to conduct systematic reviews and meta-analyses. The guidance was developed from a number of other sources on review methods, primarily from medicine and healthcare. Their guidance appears pertinent to work and organizational psychology because it was developed for a topic area many of us research and also gets published in EJWOP (well-being, see Daniels 2017) and a topic area that deploys many of the research methods we use (quantitative field studies, qualitative methods,

some use of experimental methods) and where interventions tend to be complex and multifaceted – which is also typical of many workplace interventions. I went through each of these reviews published in EJWOP and looked at whether the reviews incorporated the high-level features of high quality reviews as outlined in Snape et al., as well as other features evident in some of the reviews published in EJWOP and that are consistent with approaches advocated in Snape et al. Table 1 adapts and summarizes Snape et al.'s features of high quality reviews, and gives an indication of how many studies published in EJWOP explicitly contained each feature. A feature labelled green was found in over two thirds of the reviews/meta-analyses published in EJWOP (i.e., a large majority of reviews adopted this feature of high quality reviews), amber was given to features found in one third to two thirds of the studies and red given to features found in less than one third of the studies.

INSERT TABLE 1 HERE

Table 1 shows that reviews tended to have explicit keyword search terms, used multiple databases and applied structured methods for summarizing data (however, this is necessarily the case for meta-analyses and there were a number of meta-analyses in the sample). However, Table 1 does indicate there are several areas for improvement of our conduct of systematic reviews and meta-analyses. The two reviews scoring highest on the frequency of features listed in Table 1 were Lomas et al. (2017) and Watson et al. (2018), and so these may be useful exemplars for those new to systematic review methods.

The features listed in Table 1 can act as guidance to both authors and journal referees on what to look for in a systematic review and the non-statistical aspects of meta-analyses. It certainly is the case that not all features listed in Table 1 are relevant to every systematic review and meta-analyses (e.g. consultation with non-academic stakeholders might be critical in a practice or policy oriented review, but less relevant to a review where the primary concern is theoretical). However, the majority of features listed in Table 1 would be relevant

in the vast majority of cases. Next, I will outline the major steps included in Snape et al.'s guidance in more detail.

*Before the review*

*Dummy searches and scoping reviews.* Although not explicitly stated in the Snape et al. guidance, dummy searches and initial systematic scoping reviews serve a number of useful purposes. First, dummy searches allow researchers to refine keywords and to get an idea of the size and scope of the literature. Where are large number of 'hits' are returned in dummy keyword searches, it may indicate to the researchers that search terms can be more restrictive (e.g. include only more recent studies, place more stringent criteria on methodologies – e.g. only review longitudinal, multimethod or intervention studies). However, where a small number of hits are returned, it may be that search terms are too restrictive and bias the review away from some useful studies. This can happen for example if search terms are restricted to narrow operational definitions of key concepts in areas where broader terms are used for key concepts.

Dummy and initial searches can be useful for detecting whether there are other similar reviews already published, and therefore whether the planned review is needed, what the added value of the planned review is, whether it is more appropriate to conduct a systematic review of existing systematic reviews and meta-analyses or whether review questions can be refined so that the planned review adds a new dimension to existing knowledge. The initial searches can also be used to scope the literature for major theories relevant to the review, and therefore to gain an idea of data to extract and synthesize in the main review on potential explanatory mechanisms. Dummy keyword searches are also useful for checking the keywords are detecting the right papers: If a known and relevant paper is not found in a search, then the search terms need to be modified.

*Explicit and registered protocol*. Registering a review protocol helps to ensure that the review strategy and methods are developed prior to commencing the review, that the review does not duplicate other reviews that are being conducted and increases transparency of the review process. It also provides advanced notice to others that a review is being conducted, and therefore offers the chance for collaboration with other teams planning the same or similar review. A number of review registers are available, such as Cochrane, Campbell and PROSPERO. However, prior registration of a review is not a requirement at most journals (and not at EJWOP currently) and the available registers may not be suitable for the intended review.

Notwithstanding, a review protocol developed in advance of the review should include the following elements: background to review – context and justification; review question(s); inclusion and exclusion criteria including PICOS/PECOS elements – see below; methods – identification of evidence; selection of studies; keywords and data-bases to search; data to be extracted from studies; approach to assessment of the quality of the evidence; methods for synthesis; processes for amending protocol if allowed by the register of reviews. Despite conducting dummy searches, it may be the case that during the real review process, the number of 'hits' in the searches turns out to be much lower or higher than anticipated. In this case, it may be appropriate to amend the review protocol to make inclusion criterion more or less stringent. To prevent bias and retain transparency, any amendments to the protocol should explicitly recorded and be made prior to extracting data from studies, so that amendments are not made in light of emerging findings which could then potentially bias the reviews' conclusions. Shamseer et al. (2015) provide more details on developing review protocols (PRISMA-P guidelines).

*Specifying PICOS/PECOS or other structured approach to bounding the review questions*. Structured frameworks like PICOS and PECOS are ways to bound questions and

get precision in keyword searches and inclusion and exclusion criteria. The PICOS framework was developed for intervention studies, and the components are:

Population: This is a statement of the groups to be considered in the review (e.g. all those in employment, teachers, healthcare professionals, self-employed, unemployed);

Intervention: The intervention or class of interventions to be investigated (e.g. job redesign, mindfulness training, mentoring programs);

Comparators: Against what is the intervention being compared (e.g., practice as usual, other interventions);

Outcome: The outcomes of interest (e.g. sickness absence, turnover, creativity);

Study designs: The study types to be included in the review (e.g. randomized control trials, interrupted time series, qualitative case studies).

The PECOS framework is similar, except the 'Intervention' category is replaced by an 'Exposure' category. Although the connotations and history of development in medicine and healthcare of 'Exposure' may imply exposure to a risk factor (e.g. job insecurity), the 'Exposure' category can be used to specify any interdependent variable that is not an intervention of some sort (e.g., leadership behaviors, attitudes, experience of HR practices).

*Consultation with experts/other stakeholders on review questions, keywords, databases*. Prior to finalizing (and registering) the review protocol, consulting with subject matter experts on proposed review questions, search terms and other aspects of the reviews is a useful way of getting feedback on and improving the proposed review by those who have a good knowledge of the research literature. Consultation with non-academic stakeholders may be useful for developing a review that addresses pressing practical problems, or ensuring that an academically motivated review has practical implications. Consultation can be done via interviews, focus groups or public meetings and seminars. It is important to consult a range of

stakeholders (e.g. senior managers and worker representatives) to ensure review findings do not provide an advantage to one group at the expense of another (cf. Carter, 2018).

*Searching the literature*

*Specific keywords.* A replicable and transparent review needs pre-specified keywords that are able to reflect the range of the concept(s) being reviewed and, if appropriate, a range of relevant study designs. Dummy searches, PICOS/PECOS frameworks and consultation with others can help develop keyword strings that are both precise enough to reflect the review questions but broad enough so that relevant studies are found.

*Specific electronic sources*. Similarly, replicable and transparent reviews require the specification of the bibliographic databases that will be searched. Most reviews follow good practice here and search several databases (e.g. PsycINFO, PubMed Central, Web of Science, Scopus). Restricting electronic searches to specific journals (e.g., defined by an impact factor threshold or one of the many lists that rank the quality of journals) (also see next section) does run a risk of bias because of the problems identified in more selective journals favoring novelty over replication and bias towards significant results (for a discussion and a proposed solution, see Woznyj, Grenier, Ross, Banks & Rogelberg, 2018).

*Plans for additional searches*. Some reviews will also include other means of tracking papers. Often these include hand searches of key journals and asking known experts or research groups for papers. Sometimes, reviews can be accompanied by a 'call for evidence' posted on various lists, website or social media that requests viewers supply links to existing evidence, usually with specified quality control criteria. Contacting experts and issuing a call for evidence can be a good way for soliciting studies that are in the grey literature (see below). As part of the dummy search process prior to finalizing a review protocol, comparisons between hand searches and papers supplied by experts and electronic searches can be useful for refining keywords. However, consulting specific journals through hand

searches or contacting experts for papers in the main review does run the risk of biasing the review towards specific types of evidence. I have seen examples in the literature where hand searches have been confined exclusively or largely to journals published in one geographic area: This runs the risk of skewing reviews towards research from that area and excluding potentially relevant studies from other geographic areas. Notwithstanding, the decision to engage in or not engage in additional searches beyond electronic searches needs some justification.

*Explicit discussion of any restrictions on searches*. If searches are restricted to specific geographical regions, types of economy, years or population sub-groups this would need some justification. Sometimes the justification will be made in the introduction to a systematic review (e.g., a review focused on a specific occupation would focus on why that occupation is of substantive interest). In other cases, justification may be based on the size of the literature that warrants a focus on more recent studies: A good justification for restricting reviews to more recent studies could be only reviewing studies conducted after some earlier review (to avoid double counting studies) or other landmark paper that would indicate a shift in methods or measures used. Restricting searches to specific languages may be justified by the linguistic capabilities of the research team.

*Criteria for including/excluding grey literature*. The grey literature comprises studies not published in peer-reviewed journals (Adams, Smart & Huff, 2017; Snape et al., 2017). Reasons for including grey literature in systematic reviews and meta-analyses include: a wish to avoid publication bias that might skew the review's conclusions to confirming a hypothesis; to supplement a small body of peer-reviewed literature to make it more robust or to fill gaps in the peer-reviewed literature; to help make reviews more relevant to policy/practice by providing contextual information, illustrations of how research can be used in practice or in helping to refine research questions (see Adams et al., 2017). However, as

Adams et al. note, there are shades of grey, with some grey literature being more "academic" and retrievable than others, including doctoral theses, conference papers, working papers, technical and official reports, books and book chapters written by academic or other professional researchers. Notwithstanding, there should be a justification for including or excluding grey literature in a review. Adams et al. give guidelines for making these decisions (see especially table 6 in Adams et al.). If grey literature is included in a review, it is recommended to report findings from the peer-reviewed and grey literature separately (Adams et al.).

*Searches into relevant bibliographic software.* Documenting the search process is a means of ensuring the review process is transparent and provides data on included and excluded studies to allow other researchers the chance to replicate the review. Saving the results to bibliographic software (or even just a spreadsheet or word processing package for grey literature) is a means of keeping track of papers retrieved in initial searches and then retained during sifting of studies for inclusion in the review. Doing so also facilitates the process of up-dating reviews as new studies appear. Placing the results of searches in data repositories allows others the chance to replicate and up-date existing reviews.

*Selecting studies*

*Procedures to ensure and demonstrate inter-rater reliability/consistency at all stages of sifting.* Making inclusion and exclusion criteria explicit, having pre-specified keywords and use of PICOS/PECOS frameworks helps remove subjectivity from selecting studies to include in the review. Nevertheless, at the time of writing, after electronic keyword searches, study selection is done by humans rather than artificial intelligence, bringing some subjectivity into study selection. Typically, there are two stages in selecting studies: Sifting by title/abstract and then through reading the full paper. It is important at both stages to demonstrate there is consistency (and hence reproducibility) in deciding which studies to

include or exclude. This can be done through inter-rater reliability calculations (percentage agreement between independent coders, Cohen's kappa). For reviews with a large number of studies to consider, independent coding of a reasonably sized sub-sample of papers would be sufficient. Piloting procedures for selecting studies through dummy exercises can help ensure consistency of selecting or rejecting papers for the review.

*Presentation of PRISMA flow diagram.* A PRISMA flow diagram is a graphical representation of the stages of the review (Liberati *et al*., 2009). It adds value because it both documents and summarizes the number of studies that progress through the various stages of sifting papers to determine their inclusion in the review and the extraction of data, starting with the number of papers that were identified initially as being potentially relevant. Figure 1 shows an example.

INSERT FIGURE 1 HERE

*Data Extraction*

*Explicit and piloted data extraction sheets*. Like an explicit protocol, an explicit data extraction sheet makes it clear to the researchers what data need to be extracted and summarized from a paper and also makes it easier to collate and summarize findings from across the studies in the review. Formalizing data extraction in this way also enables the research team to check the consistency of extraction by having all or a sample of studies coded by two or more members of the review team. Piloting data extraction sheets enables the research team to maximize the chances that the right data will be extracted consistently during the real data extraction process.

As well as bibliographical information, data extraction sheets can include information on where the study was conducted, the study population, size and characteristics of the sample, the hypotheses and theories used, independent variables, dependent variables, mediators, moderators, control variables, how the key variables were measured, study design,

findings in relation to review questions – including effect sizes and significance for quantitative studies, and any features relevant to the quality of the study (e.g. attrition rates, confounding, inadequacies in analyses). Information on research context, design, measures and theoretical motivation is useful to include in data extraction sheets: Should there be divergence in study findings, such information allows researchers to examine whether some findings emerge only in specific contexts, conditions or for specific measures. Even if not included in the data synthesis part of a review, recording such features can be useful for future reviews.

Features relevant to the quality of the study can be specified in advance in the data extraction sheet, and there are checklists available for quantitative and qualitative studies that can be incorporated into data extraction sheets (e.g., Early Intervention Foundation, 2015; Critical Appraisal Skills Programme, 2014). In some cases, research teams may wish to give a grading on the quality of each study: However, providing summary ratings of the quality of a given study can be problematic – such assessments may have validity issues and a simple summary of a study's quality may obscure pertinent features of a study that are relevant to a given review question or conflate design features that are irrelevant to a given question (see Higgins, Altman & Sterne, 2011). Notwithstanding the choice to grade individual studies, information on the quality of studies can be synthesized to allow statements regarding the strength of evidence underpinning review findings (see below) and specify gaps in available knowledge. For example, a body of studies using the strongest methods may concentrate on one particular sector, suggesting possible limitations of the external validity of the strongest causal statements. Data extraction sheets for intervention studies can also include space for data on how well the intervention was implemented and factors that may have affected implementation as intended (and whether the data on implementation were gathered through formal quantitative or qualitative research processes or whether information on

implementation was in the form of post hoc speculation by the authors for the pattern of results).

*Procedures to ensure and demonstrate inter-rater reliability/consistency of data extraction from full study*. Where numerical data are extracted (e.g. effect sizes), inter-coder consistency can be checked statistically. However, for information summarized qualitatively, which will be the case for most systematic reviews irrespective of whether the selected studies are primarily quantitative or qualitative, consistency may be better checked through double coding all or a sub-sample of papers and discussing the level of consistency amongst the research team. Snape et al. recommend a minimum of 10% of papers are double coded to check consistency, but the number depends on the number of selected studies: 10% would be inadequate for a review of 10 studies for example. Each member of the review team involved in data extraction should have one or more of the studies they reviewed double coded to check the consistency of data extraction. Where consistency is difficult to establish, it may be appropriate to have all studies double coded with differences resolved by discussion or by recourse to a third reviewer.

*Evidence synthesis*

*Use of meta-analysis or graphical or other structured methods for evidence synthesis*. Systematic reviews often contain evidence tables summarizing key features and findings from included studies (evidence tables for larger reviews are sometimes made available as supplementary material on-line). Although informative, evidence tables do not synthesize the evidence into a coherent summary of the state of knowledge about a review question. In meta-analyses, this summary is provided by calculation of meta-effect sizes as well as moderator analyses demonstrating the conditions under which effect sizes may vary. Meta-analyses may also be accompanied by forest plots that summarize effect sizes and confidence intervals for each individual study (Lewis & Clarke, 2001).

Although systematic reviews include a narrative synthesis of evidence, often with a more heterogeneous range of studies than typically seen in meta-analyses, harvest plots can be used to summarize evidence (Ogilvie, Fayter, Petticrew, Sowden, Thomas, Whitehead, & Worthy, 2008) and help guide the narrative synthesis of the evidence. In their simplest form, a harvest plot simply creates a column for each study that provides evidence of a positive effect on an intervention or other independent variable on a dependent variable, a column for each study that provides evidence of a null effect and a column that provides evidence of a negative effect. Inspection of harvest plots can then give an overall picture of variability or consistency in the evidence, and where there are inconsistencies, justify an inductive search for reasons for discrepancies between studies. Harvest plots can also be modified to reflect potential moderators and mediators. For example, in Watson et al. (2018, supplementary material), harvest plots were developed so that: a) studies with better methods for causal inference (e.g., randomized control trials) were given higher columns than studies with weaker methods for causal inference (e.g., non-equivalent control group designs), enabling a visual comparison of whether studies with stronger or weaker designs were more likely to return evidence for positive, null or adverse effects; and b) studies that assessed theoretical mediators of the effects of the independent variable were given columns with different colors to those studies that did not assess mediators, allowing a visual summary of whether underpinning explanations for the effects were viable.

It is also possible to use various structured methods familiar to qualitative researchers (content analysis, thematic analysis) to structure narrative syntheses. Reviews of heterogeneous evidence can develop inductive categories of studies (sometimes overlapping) to help structure review findings and explore differences between different categories of studies (see again Watson et al., 2018 for their categorization of different training interventions).

*Explicit evidence statements or other forms of summarizing into review findings.* Another means of summarizing the narrative synthesis of a systematic review is the provision of evidence statements. These are merely sentences or short paragraphs that summarize the evidence in relation to a research question, and list the studies from which the evidence statement is derived. Evidence statements can relate to direct effects, null effects, moderated relationships, adverse effects, mediated effects and statements in relation to theory. Examples could be:

(Direct effect) *There is evidence to support a positive association between X and Y.*

(Null effect) *There is no evidence to support a relationship between X and Y.*

(Moderated effect/boundary condition) *Interventions to improve X coupled with direct improvements in Z may improve Y.*

(Potential adverse effects, including boundary condition) *Interventions that use process Z to improve X have mixed effects on Y, including adverse effects in circumstances.*

(Mediated effects, plus statement in relation to theory) *Interventions that improve X have positive effects on Y by reducing levels of M, supporting predictions made in theory A.*

Evidence statements need to be agreed by the research team as reflections of the evidence presented in the evidence synthesis. This provides a check on the consistency of interpretation of the evidence.

Evidence statements not only provide a succinct summary of the evidence and the narrative synthesis, which can then be used as a basis to communicate review findings in plain language, but also provide a very useful platform to summarize the quality of the evidence in a given area and therefore the degree of confidence we can have in extant research (see next section). Of course, evidence statements are implicit in meta-analyses with

the statement of the meta-effect size and credibility/confidence intervals, although the meta-effect size will still need translation for many user groups.

*Explicit assessments of quality of evidence underpinning each review finding*. Although it is possible to rate the quality of each study, this is not as informative as rating the overall quality of the evidence contributing to each review finding. Each evidence statement can be given a quality rating. Quality ratings are useful for many reasons. For example, where an evidence statement gives an indication of a relationship between two or more variables, but the quality of evidence is low, then there is a clear need for more studies attempting replication but with stronger methods. Where the evidence is good, but lacking intervention studies, there is a case for developing intervention studies. Where the evidence is strong for a relationship between two variables and there are multiple replications from intervention studies with consistent effects, then it is time to move research onto something else. Similarly, where there is insufficient evidence of a relationship between two variables but the evidence base is judged to be strong (and all sorts of moderators have been tried but there is no evidence of moderation), then it might be time to conclude there is not really a relationship there after all. On the other hand, insufficient evidence to claim a relationship between two variables and a weak evidence base suggests further research is needed (assuming good reasons to suspect a relationship should exist). For work and organizational practitioners, quality ratings of evidence can serve as a 'buyer-beware' warning, letting practitioners know the confidence with which evidenced-based recommendations can be made, giving practitioners one source of comparison between different courses of action and guidance on the extent to which more detailed knowledge of a given context is required before an evidence based recommendation for action can be made.

There are various guidelines and rating scales that enable researchers to provide statements on the quality of evidence underpinning findings. The GRADE criteria (see e.g.

Guyatt et al. 2011) have been developed in medical sciences for quantitative evidence, and have four levels going from high quality evidence (generally multiple replications of {unbiased} randomized control trials with consistent effects) to low quality evidence (observational studies). Problems with randomized control trials can down grade evidence but consistent and large effects sizes across observational studies with sources of confounding controlled can up-grade evidence from observational studies. As such, multiple replications from large scale and well-designed epidemiological studies controlling for known sources of confounding could be up-graded to high quality evidence. In work and organizational psychology and similar areas, the GRADE criteria can be problematic.

In our area, interventions tend to be complex, differ from study to study and context to context and randomized control trials are relatively rare. This could then mean that evidence from multiple longitudinal and large scale surveys would be rated of higher quality than intervention studies. This is very problematic if the survey research has an unknown omitted variables problem, and observed effects in surveys only occur because of the conjoint presence of an unmeasured moderator that correlates with the independent variable and that serves as a catalyst for the effects of an independent variable in interventions. For example, job control correlates with performance (Spector, 1986), job control is typically present in jobs that also have high levels of training and performance management practices (Ogbonnaya & Daniels, 2017) and so any causal effects of job control on performance may only occur when adequate training and other employment practices are also present (Combs, Liu, Hall & Ketchen, 2006). Put another way, why would a manager allow workers to take decisions if they had not been trained to have the knowledge to take good decisions?

For qualitative evidence, the CERQual criteria (Lewin et al., 2015) also grade evidence according to four categories, based on the methodological limitations of studies, coherence of the review finding with the evidence from primary studies, adequacy of

supporting data and relevance to the context specified in the review question. Ratings go from high confidence that the review finding represents the phenomenon/phenomena of interest to very low confidence, where it is not clear where the review finding is a reasonable representation of the phenomenon/phenomena of interest.

In attempting to integrate the GRADE and CERQual approaches and to make quality assessments applicable to complex interventions and to be communicable to users, Snape et al. (2017) have presented 'plain language' quality ratings, ranging from:

Strong evidence – where there is more than one high quality intervention study (e.g. randomized control trials) showing replicated results;

Promising evidence – where there is a single high quality intervention study (e.g. randomized control trial) with some limitations or multiple studies with limitations (e.g. non-equivalent control group designs) showing replicated results;

Initial evidence – where there is a single intervention study with some limitations.

Unclear evidence – no consistent evidence across studies or studies have significant limitations.

A similarly flexible evidence grading system has been developed by Puttick and Ludlow (2013) for complex social interventions. This is the NESTA standards of evidence and has five levels:

Level 1. An intervention can be described and why it matters;

Level 2. The evidence of positive change for an intervention, but there is only weak causal inference (e.g. pre-post only designs; panel studies);

Level 3. Evidence of change using a controlled design (randomized or non-equivalent control group);

Level 4. More than one independent replication of evidence of change.

Level 5. Fully proceduralized systems and procedures for implementation of interventions alongside multiple and consistent controlled replications.

As with all stages in a systematic review, procedures need to be in place to demonstrate inter-rater reliability/consistency of quality assessments. This can be as simple as discussions within the research team and/or discussions with researchers external to the review team to ensure appropriate factors have been considered and the quality ratings are consistent with the quality of the evidence available.

*Assessment of sources of bias*. Systematic biases in an evidence base do limit confidence in review findings. In meta-analyses, failsafe N calculations can help researchers determine whether selective publication of significant results in journals has skewed the evidence base by estimating the number of unpublished null effect studies that would be required to reduce a meta-effect size to zero. Publication bias and selective reporting of significant effects in the expected direction can limit findings from systematic reviews and such biases are harder to detect in systematic reviews (although not impossible). Searching the grey literature can help meta-analyses and systematic reviews in these respects (especially where the number of published studies are low).

Although partially captured in evidence gradings, considering systematic sources of bias common to a number of papers helps to define how to take research forward either through methodological refinements or providing alternative conceptual accounts. For example, where there are few randomized control trials and/or participants are allowed to opt into interventions (which is usually the case for ethical reasons), then selection bias is a problem. Although this is a methodological nuisance, in some areas, selection bias provides an alternative explanation for intervention effects. For example, opting into a training intervention provides a participant with some control over what knowledge s/he will receive, potentially reflecting a motivational mechanism for any improvements in performance rather

than necessarily improvement in knowledge or improvements in knowledge in the absence of motivational effects.

*Stated evidence gaps.* Evidence statements summarize what is known and quality assessments summarize the confidence we can have in that knowledge. Narrative explanations of why quality assessments are made can provide a means of determining future research directions, primarily through methodological refinement and constructive replication. The absence of evidence statements or evidence statements that indicate that there is too little evidence to make a sensible conclusion also point to areas for future research. Systematic reviews then should also synthesize research in such a way that the research team can provide recommendations for future research on the basis of what is *not* in the knowledge base. As well as methodological refinements or the absence of studies looking at particular outcomes, other examples could be an absence of studies examining theoretically predicted mediators or preponderance of studies conducted in similar contexts hinting at problems with external validity and potentially boundary conditions. For intervention studies, researchers may ask whether there are sufficient numbers of studies examining the processes of interventions using appropriate research methods and theories/conceptual models.

*Other things to consider in systematic reviews*

It appears from reviews published in EJWOP and other reviews I have read (and conducted) that cost effectiveness/return on investment type analyses are often absent from many systematic reviews. This impression is borne out at least in the health and wellbeing field by a recent study indicating such economic considerations are a research priority identified by occupational health professionals (Lalloo, Demou, Smedley, Madan, Asanati & Macdonald, 2018). Going forward, it would be helpful if systematic reviews and primary intervention studies were to give more attention to these economic considerations. Placing

cost effectiveness ratings as well as quality of evidence ratings next to evidence statements gives practitioners another criterion on which to judge the attractiveness of options for action.

There may also be concerns over what happens if a review returns only a small number of studies to the data extraction phase. Dummy searches and scoping reviews should guard against this, but research teams might be constrained by the needs of funders or the desire to use only the most rigorous methods given other reviews have included less rigorous methods. In general, a small number of studies should not be used as a criteria for rejecting a review that is otherwise technically excellent and conceptually well grounded. To do so would be like rejecting a primary empirical paper that is methodological rigorous and conceptually sound for having too few hypotheses supported. In the case of systematic reviews, knowing there is little or no information on a question is important to know and can form the basis of a whole new program of research. In this respect, there are perhaps two solutions for the case where a review has a small number of studies. The first is the results blind review route, so that the review is judged solely on its technical adequacy and the importance of the topic. This is linked to the second solution, which is to ensure that the review addresses a compelling questions(s) to which the answer 'there is not much evidence' points to significant ways in which the knowledge base can be improved.

Consultation with a range of academic and non-academic stakeholders can help develop compelling research questions for reviews. Some stakeholders may want answers to questions as simple as 'did it work' or 'is there a relationship', but such questions can be made more compelling by asking why something might work or why there is a relationship (and under what conditions) – which links the review into theoretical questions. It may be that prior reviews find equivocal evidence for a set of interventions, hinting at gaps in knowledge of how things work: Finding theoretically grounded reasons why something may or may not work and then reviewing the evidence would also be compelling.

Grote and Cortina (2018, p 338) give some advice on how to do useful work and organizational psychology research, which is very applicable to systematic reviews and meta-analyses. The advice requires researchers to ask this question (my additions in italics to reflect the broader context of Grote and Cortina's paper): "would an answer to this question improve organizational functioning [*broadly defined to reflect interests and wellbeing of a broad range of organizational and societal stakeholders*] in a nontrivial way, and would it prompt other researchers to improve organizational functioning further still".

*Concluding comments*

Systematic reviews and meta-analyses can provide clear summaries of what we know about specific questions, the theoretical reasons for those answers and the confidence we can have in those answers. Such summaries are of clear value to the research community and various stakeholders connected to policy and practice, although for many organizational stakeholders, the review itself would require further translation to provide useful and usable knowledge. As a summary of the features of high quality reviews, Table 1 can provide a baseline for the conduct and evaluation of reviews in work and organizational psychology. I would also encourage authors to reflect on how to make their reviews stand out by being more than technically accomplished, through for example: Addressing compelling research questions that address issues of real concern amongst stakeholders; addressing theoretical reasons for uncovered patterns of relationships; identifying boundary conditions on relationships; and, for reviews of intervention studies and to encourage practical up-take of the knowledge we generate in our research, considering the (economic) costs and benefits of interventions.

**References**

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews, 19*, 432-454. DOI: 10.1111/ijmr.12102.

Bartlett, D., & Francis-Smythe, J. (2016). Bridging the divide in work and organizational psychology: evidence from practice. *European Journal of Work and Organizational Psychology*, *25*, 615-630. DOI: 10.1080/1359432X.2016.1156672.

Boer, D., Deinert, A., Homan, A. C., & Voelpel, S. C. (2016). Revisiting the mediating role of leader–member exchange in transformational leadership: the differential impact model. *European Journal of Work and Organizational Psychology, 25*, 883-899. DOI: 10.1080/1359432X.2016.1170007.

Bouckenooghe, D., M. Schwarz, G., & Minbashian, A. (2015). Herscovitch and Meyer's three-component model of commitment to change: Meta-analytic findings. *European Journal of Work and Organizational Psychology, 24*, 578-595. DOI: 10.1080/1359432X.2014.963059.

Bozer, G., & Jones, R. J. (2018). Understanding the factors that determine workplace coaching effectiveness: a systematic literature review. *European Journal of Work and Organizational Psychology*, *27*, 342-361. DOI: 10.1080/1359432X.2018.1446946.

Carter, A. J. (2018). Commentary on neoliberal ideology in work and organizational psychology. *European Journal of Work and Organizational Psychology*, *27*, 552-553. DOI: 10.1080/1359432X.2018.1517116.

Combs, J., Liu, Y., Hall, A. & Ketchen, D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. *Personnel Psychology, 59*, 501-528. DOI: 10.1111/j.1744-6570.2006.00045.x.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks, CA: Sage.

Critical Appraisal Skills Programme (CASP) (2014). *CASP Checklists (qualitative checklist)* Oxford: CASP.

Daniels, K. (2016). An editorial in four parts. *European Journal of Work and Organizational Psychology*, *25*, 329-334, DOI: 10.1080/1359432X.2016.1145669.

Daniels, K. (2017). Thanks, congratulations and publishing useful research. *European Journal of Work and Organizational Psychology*, *26*, 629-633, DOI: 10.1080/1359432X.2017.1352575.

Early Intervention Foundation (2015). *Translating the evidence. A brief guide to the Early Intervention Foundation's procedures for identifying, assessing, and disseminating information about early intervention programmes and their evidence*. London: Early Intervention Foundation.

Field, A., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665-694. doi: 10.1348/000711010X502733

Green, S., Higgins, J. P., Alderson, P., Clarke, M., Mulrow, C. D., & Oxman, A. D. (2011). *Cochrane handbook for systematic reviews of interventions version 5.1. 0: updated March 2011*. London: The Cochrane Collaboration.

Grote, G., & Cortina, J. M. (2018). Necessity (not just novelty) is the mother of invention: using creativity research to improve research in work and organizational psychology. *European Journal of Work and Organizational Psychology, 27*, 335-341. DOI: 110.1080/1359432X.2018.1444606.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., et al. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings

tables. *Journal of Clinical Epidemiology, 64,* 383-394. DOI: 10.1016/j.jclinepi.2010.04.026

Harari, M. B., Reaves, A. C., & Viswesvaran, C. (2016). Creative and innovative performance: A meta-analysis of relationships with task, citizenship, and counterproductive job performance dimensions. *European Journal of Work and Organizational Psychology, 25,* 495-511. DOI: 10.1080/1359432X.2015.1134491.

Higgins, J.P.T., Altman, D.G., Sterne, A.C., (2011). Assessing risk of bias in included studies. In J.P.T. Higgins & S. Green (Ed.s). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.* The Cochrane Collaboration, 2011. www.handbook.cochrane.org

Hodgkinson, G. P., & Ford, J. K. (2014). Narrative, meta-analytic, and systematic reviews: What are the differences and why do they matter? *Journal of Organizational Behavior, 35,* S1-S5. DOI: 10.1002/job.1918

Joseph, B., Walker, A., & Fuller-Tyszkiewicz, M. (2018). Evaluating the effectiveness of employee assistance programmes: A systematic review. *European Journal of Work and Organizational Psychology, 27,* 1-15. DOI: 10.1080/1359432X.2017.1374245.

Kröll, C., Doebler, P., & Nüesch, S. (2017). Meta-analytic evidence of the effectiveness of stress management at work. *European Journal of Work and Organizational Psychology, 26,* 677-693. DOI: 10.1080/1359432X.2017.1347157.

Lalloo, D., Demou, E., Smedley, J., Madan, I., Asanati, K., & Macdonald, E. B. (2018). Current research priorities for UK occupational physicians and occupational health researchers: a modified Delphi study. *Occupational and Environmental Medicine, 75,* 830-836. DOI: 10.1136/oemed-2018-105114.

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C.J., Gülmezoglu, M., et al. (2015) Using Qualitative Evidence in Decision Making for Health and Social

Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (GRADE-CERQual). *PLoS Med 12*: e1001895. doi:10.1371/journal.pmed.1001895

Liberati A., Altman D.G., Tetzlaff J., Mulrow C., Gøtzsche P.C., Ioannidis J.P.A., et al. (2009) The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med 6*: e1000100. doi:10.1371/journal.pmed.100010.

Lomas, T., Medina, J. C., Ivtzan, I., Rupprecht, S., Hart, R., & Eiroa-Orosa, F. J. (2017). The impact of mindfulness on well-being and performance in the workplace: an inclusive systematic review of the empirical literature. *European Journal of Work and Organizational Psychology, 26*, 492-513. DOI: 10.1080/1359432X.2017.1308924.

Maynard, M. T., Kennedy, D. M., & Sommer, S. A. (2015). Team adaptation: A fifteen-year synthesis (1998–2013) and framework for how this literature needs to "adapt" going forward. *European Journal of Work and Organizational Psychology, 24*, 652-677. DOI: 10.1080/1359432X.2014.1001376.

Monteiro, S., Marques Pinto, A., & Roberto, M. S. (2016). Job demands, coping, and impacts of occupational stress among journalists: A systematic review. *European Journal of Work and Organizational Psychology, 25*, 751-772. DOI: 10.1080/1359432X.2015.1114470.

Ogbonnaya, C., Daniels, K. (2017). *What is a Good Job? Analysis of the British 2012 Skills and Employment Survey*. London: What Works Centre for Wellbeing.

Ogilvie, D., Fayter, D., Petticrew, M., Sowden, A., Thomas, S., Whitehead, M., & Worthy, G. (2008). The harvest plot: a method for synthesising evidence about the differential effects of interventions. *BMC Medical Research Methodology, 8*, 8. DOI: 10.1186/1471-2288-8-8.

Posthuma, R. A., Campion, M.C., & Campion, M.A. (2018). A taxonomic foundation for evidence-based research on employee performance management. *European Journal of Work and Organizational Psychology, 27*, 168-187. DOI: 10.1080/1359432X.2018.1438411.

Potočnik, K., & Anderson, N. (2016). A constructively critical review of change and innovation-related concepts: towards conceptual and operational clarity. *European Journal of Work and Organizational Psychology, 25*, 481-494. DOI: 10.1080/1359432X.2016.1176022.

Puttick, R., & Ludlow, J. (2012). *Standards of evidence for impact investing*. London: Nesta.

Rynes, S. L., Colbert, A. E., & O'Boyle, E. H. (2018). When the "best available evidence" doesn't win: How doubts about science and scientists threaten the future of evidence-based management. *Journal of Management, 44*, 2995-3010. DOI: 10.1177/0149206318796934.

Samsudin, E. Z., Isahak, M., & Rampal, S. (2018). The prevalence, risk factors and outcomes of workplace bullying among junior doctors: a systematic review. *European Journal of Work and Organizational Psychology, 27*. DOI: 10.1080/1359432X.2018.1502171.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *British Medical Journal, 349*, g7647. DOI: 10.1136/bmj.g7647.

Shepherd, J.P. (2014). *How to Achieve More Effective Services: The Evidence Ecosystem*. Cardiff University.

Snape, D., Meads, C., Bagnall, A-M, Tregaskis, O. & Mansfield, L. (2016). *What works wellbeing: A guide to our evidence review methods*. London: What Works Centre for Wellbeing.

Spector, P.E. (1986). Perceived control by employees: A meta-analysis of studies concerning autonomy and participation at work. *Human Relations, 39*, 1005-1016. DOI: 10.1177/001872678603901104.

Watson, D., Tregaskis, O., Gedikli, C., Vaughn, O., & Semkina, A. (2018). Well-being through learning: a systematic review of learning interventions in the workplace and their impact on well-being. *European Journal of Work and Organizational Psychology, 27*, 247-268. DOI: 10.1080/1359432X.2018.1435529.

Woznyj, H. M., Grenier, K., Ross, R., Banks, G. C., & Rogelberg, S. G. (2018). Results-blind review: a masked crusader for science. *European Journal of Work and Organizational Psychology, 27*, 561-576. DOI: 10.1080/1359432X.2018.1496081.

*Table 1.* List of features of high quality reviews explicitly stated in EJWOP papers since 2015.

| Feature | Proportion of studies meeting criteria |
|---|---|
| Dummy searches and/or scoping reviews | Red |
| Explicit & registered review protocol | Red |
| Specifying PICOS/PECOS or other structured approach to bounding the review questions | Red |
| Consultation with experts/other stakeholders on review questions, keywords, databases | Red |
| Specific keywords | Green |
| Specified electronic sources | Green |
| Plans for additional searches | Amber |
| Explicit discussion of any restrictions on searches | Amber |
| Criteria for including/excluding grey literature | Amber |
| Searches into relevant bibliographic software | Red |
| Procedures to ensure and demonstrate inter-rater reliability/consistency at all stages of sifting (title/abstract/full study) | Red |
| Presentation of PRISMA diagram | Amber |
| Explicit and piloted data extraction sheets | Red |
| Procedures to ensure and demonstrate inter-rater reliability/consistency of data extraction from full study | Amber |
| Use of meta-analysis or graphical or other structured methods for evidence synthesis | Green |
| Explicit evidence statements or other forms of summarizing into review findings | Amber |
| Explicit assessments of quality of evidence underpinning each review finding | Amber |
| Assessment of sources of bias | Red |
| Stated evidence gaps | Amber |

Red – stated in < 1/3 of studies, Amber – stated in 1/3 to 2/3 of studies, Green – stated over 2/3 of studies

*Figure 1.* Example study selection diagram.