METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

The Development of a New Generic Risk-of-Bias Measure for Systematic Reviews of Surveys

Gabriel Nudelman [a], Kathleen Otto [b]

[a] *School of Behavioral Sciences, The Academic College of Tel Aviv-Yaffo, Jaffa, Israel.* **[b]** *Faculty of Psychology, Philipps University of Marburg, Marburg, Germany.*

Methodology, 2020, Vol. 16(4), 278–298, https://doi.org/10.5964/meth.4329

Received: 2018-11-04 • **Accepted:** 2020-08-10 • **Published (VoR):** 2020-12-22

Corresponding Author: Gabriel Nudelman, School of Behavioral Sciences, The Academic College of Tel Aviv-Yaffo, Rabenu Yeruham St., P.O. Box 8401, Jaffa 6818211, Israel. Tel.: +972 (0)36803365, E-mail: gabrielnu@mta.ac.il

Abstract

It is important to evaluate risk of bias of the primary studies included in systematic reviews and meta-analyses. Since tools pertinent to surveys are scarce, the goal of the current research was to develop a measure to address this need. In Study 1, an initial list of 10 relevant topics was compiled from previous measures. In Study 2, the list was refined into an eight-item risk-of-bias measure via discussion and a pilot study. In Study 3, experienced researchers used the measure to asses 70 studies, demonstrating high interrater agreement (weighted Kappa = .82). Inexperienced raters also utilized the measure to code 26 different studies included in a prior meta-analysis, which resulted in adequate interrater agreement (weighted Kappa = .64) and excellent convergent validity (r = .66). Thus, the new measure, designed to be accessible and flexible, can increase standardization of risk-of-bias evaluations and contribute to the interpretation of systematic reviews and meta-analytic findings.

Keywords

risk of bias, systematic review, meta-analysis, survey quality, correlational studies

Systematic reviews and meta-analyses are ubiquitous means of integrating past research in order to construct orderly and reliable knowledge, as well as to highlight important unresolved questions (Cooper, 2010). However, their results are only as valid as the studies they include. Tools designed to evaluate systematic reviews and meta-analyses acknowledge this and assign favorable ratings to reviews that assess the scientific quality of the studies they incorporate (Shea et al., 2007). Meta-analysts, in turn, increasingly perform such quality assessments and take into account risk-of-bias measures of primary studies that reflect the degree of confidence in the findings they synthesize (Johnson, Low, & MacDonald, 2015). Yet, many explicit indicators of scientific quality—and of

risk of bias in particular (see below for a distinction between the two concepts)—were originally developed for use in experimental studies (La Torre et al., 2006). While these measures serve their purpose well, they might be less applicable to the synthesizing of research based on surveys, which are designed to examine relations among psychological constructs and focus on means and simple correlations. Despite the increasing prevalence of surveys, tools to assess their quality have not been developed yet. Consequently, researchers conducting meta-analyses of studies using such research designs are often limited in their capacity to find a risk-of-bias measure that is compatible with their needs, which leads to disregarding this risk or evaluating it on the basis of simple decision rules (e.g., whether or not the measures utilized in a study are reliable). For example, a recent study that examined quantitative systematic reviews and meta-analyses in the field of Industrial and Organizational Psychology found that only 2 out of 120 randomly selected articles reported any assessment of study quality (Schalken & Rietbergen, 2017). This may lead to a detrimental effect in which the results of meta-analyses misrepresent the true phenomenon they seek to investigate (La Torre et al., 2006). Therefore, the development of a tool fit for surveys should be considered a priority (Protogerou & Hagger, 2019). To address this problem, the current study sought to develop an accessible, reliable, and valid generic tool to assess risk of bias in surveys and studies that do not include experiments or interventions.

Risk of Bias and Existing Tools for Its Assessment

Although sometimes used interchangeably, the terms "quality of study" and "risk of bias" represent different concepts. While the quality of a study may include numerous aspects related to its development, implementation, and presentation, the bias in this investigation refers only to systematic error in a particular study's methodology (Higgins & Green, 2008). For example, Downs and Black's (1998) checklist for assessing the quality of healthcare studies includes an item assessing the clarity with which the study's aims are reported, which might not be relevant for possible bias related to the numerical aggregation of effect sizes. Others have also stressed the need to discriminate reporting quality from methodological validity of studies (Shamliyan, Kane, & Dickinson, 2010). Thus, when addressing risk of bias of primary studies, emphasis should be placed on validity considerations, since other elements of scientific quality may be less likely to have implications for the meta-analytic results (Hoy et al., 2012).

Even tools developed explicitly to address risk of bias, and not quality of study, are not relevant to all methodological approaches. Many of these tools were designed for the evaluation of intervention-based studies and refer to research design aspects that are not necessarily present in other types of studies. For example, addressing the adequacy of the selection of the non-exposed cohort is a critical issue in assessing risk of bias in non-randomized medical studies (Sterne et al., 2016; Wells et al., 2000), whereas it is not pertinent to studies on the associations between personality characteristics (Nudelman,

2013). The limited applicability of existing risk-of-bias measures to other types of research designs might explain their low utilization levels in certain research domains. However, by not incorporating such measures, systematic reviews and meta-analyses effectively assume a low level of risk of bias across primary studies. Nevertheless, these studies are clearly not immune to the implications of systematic error. Currently, researchers wishing to incorporate broader risk-of-bias assessment into meta-analyses of survey studies need to create ad-hoc measures that are specific to their needs, as was done, for example, by Faragher and colleagues (2005) in the context of job satisfaction in organizations.

The Current Study

The aim of the current study was to develop an assessment tool that addresses key issues related to risk of bias, designed specifically to be relevant and accessible to researchers performing meta-analyses of surveys. The tool was planned to be based on and complement existing measures that are applicable to other research designs and questions. In addition, methodological validity provided the general framework for addressing bias (Shadish, Cook, & Campbell, 2002). Thus, in designing the new risk-of-bias tool, we focused on considerations related to appropriateness of sampling and data management, while disregarding topics such as the interpretation of statistical results and the congruence between hypotheses and operationalization. Although the latter topics may indicate low quality of research and reporting, they are not necessarily indicative of risk of bias related to the study design and implementation. We further sought to make the tool generic and flexible, in order to accommodate differences in certain criteria across research questions and designs (e.g., larger sample sizes are often required when utilizing between- compared to within- subjects designs). The process by which the risk-of-bias assessment tool was developed comprised three studies which goals were: 1) to create an initial list of potential topics based on previous measures and the literature; 2) to develop a final tool through discussion and the use of a pilot study; and 3) to evaluate interrater reliability among experienced and inexperienced raters who use the tool to code empirical studies, and to assess validity by using studies that were rated for quality in a previous meta-analysis.

Study 1: Creating an Initial List of Topics

In order to compile a list of possibly relevant topics for risk of bias in surveys, we examined prominent risk-of-bias measures and selected items that were relevant only to methodology (e.g., unrelated to the clarity of the research hypothesis) and to surveys and non-experimental designs (e.g., unrelated to manipulation differences between ex-

perimental and control group). In addition, a systematic literature search was conducted to ensure that potentially pertinent topics have not been overlooked.

Method

We reviewed the following common tools that are used for evaluating risk of bias: the Newcastle-Ottawa Scale (Wells et al., 2000), the Study Design and Implementation Assessment Device (Valentine & Cooper, 2008), the NIH Quality Assessment Tool (NIH, 2017), the STROBE Checklist (Vandenbroucke et al., 2007), the GRACE Checklist (Dreyer, Velentgas, Westrich, & Dubois, 2014), the MORE Checklist (Shamliyan et al., 2010), methodological quality assessment from the Cochrane handbook for systematic reviews of diagnostic test accuracy (Reitsma et al., 2009), the Risk of Bias In Non-Randomized Studies of Interventions Assessment Tool (Sterne et al., 2016), the revised tool to assess risk of bias in randomized trials (Higgins et al., 2016), and two other prominent scales (Downs & Black, 1998; Hoy et al., 2012).

The literature search was conducted using the ISI Web of Science database for articles in English in March 2017, and included all results up to that point. The terms used for the literature search were a combination (using the logical operator AND) of at least one key-word from each of three domains relevant to the current investigation: review, meta-analysis, or synthesis; risk of bias, quality assessment, quality of study, methodological quality, or validity of study; and correlational studies, observational studies, or prevalence studies.

Results

The lists were scrutinized by the first author for any items or topics that may be relevant for surveys. Most items from previous lists were not relevant for survey methods, such as whether a structured interview that was blind to case/control status was conducted (Wells et al., 2000) or whether the hypothesis/aim/objective of the study were clearly described (Downs & Black, 1998). This process resulted in an initial list of 10 topics addressing different elements of risk of bias: measurement reliability; sample representativeness; participant recruitment; exclusion rate; study settings; data management; sample characteristics (demographic variables); sample size; study reported in a peer reviewed journal or not; and response rate.

The literature search yielded 257 papers that fit our search. There were no exclusion criteria; each paper was examined by the first author for relevant items that had not yet been included in the initial list, in order from newest to oldest. However, this process ended after reviewing 150 papers, since no new topics were found and saturation was reached - no new information or themes were observed in the data (Guest, Bunce, & Johnson, 2006).

Discussion

Previous measures and studies related to risk of bias have provided a wide range of criteria for its estimation. While no tool addressed specifically surveys, many relevant items were retrieved, and a comprehensive yet manageable list of criteria was constructed. The literature review found no new topics, signifying that the key points pertinent to risk of bias have already been identified by and addressed in previous instruments. Nevertheless, the utility of some items was unclear in light of methodological changes over the past decades (e.g., distribution using social media platforms), and the exact phrasing of items and response scales often differed between measures.

Study 2: Developing and Finalizing the Tool

The purpose of this study was to create a risk-of-bias measure that addresses the main themes of methodological bias, that has a uniform answering scale, and that can be easily used and interpreted. Consistent with previous studies that developed risk-of-bias measures (e.g., Hoy et al., 2012; Jarde, Losilla, Vives, & Rodrigo, 2013), this was achieved by reaching consensus among the researchers concerning content, wording and format, and conducting a small pilot study.

Method

The list obtained in Study 1 was discussed between the authors (all Professors with a PhD in Psychology) until agreement was reached concerning relevance, phrasing and scope of the items.

This list was then used in a pilot study that entailed measuring the risk of bias of three reported studies, chosen due to the heterogeneity among them (e.g., different countries, measures, and recruitment methods). All the items regarding each empirical paper were rated by two judges (the authors), leading to a subsequent discussion for measurement refinement. An overall risk-of-bias score was calculated by summarizing the responses, with higher scores representing lower risk of bias.

Results

The initial list of 10 topics from Study 1 was narrowed down to eight topics and related items, each pertaining to a different aspect of risk of bias (two topics were excluded due to low relevance: whether the study was reported in a peer reviewed journal and response rate). It was also agreed that each item would be scored according to a binary approach (i.e., a “yes” or “no” answer for a given risk-of-bias criterion).

Following the pilot study, the reasons leading to the assessment of each item pertaining to each study were discussed and the exact phrasing of the items was refined for optimal clarity. For example, to retain both an item addressing recruitment method and an

item addressing the sampling frame while decreasing their overlap, the phrasing of these items as well as coding guidelines needed to be distinct and clearly defined. Therefore, one item of the measure examines whether the sampling frame is largely representative of the studied population (e.g., by using the phone directory for the general population) whereas another item assesses whether appropriate methods were utilized for participant recruitment (e.g., random sampling) regardless of the sampling frame. Consequently, the final tool included eight items that address eight topics (described in detail further on), each in a form of a question with two possible answers. For example, "were appropriate methods utilized for participant recruitment?" could receive the answer yes, explained in the coding guide as representing relatively low sampling bias (e.g., random selection) or the answer no, described as representing high potential for sampling bias (e.g., convenience sample) or not reported in the study. The new measure was labeled Risk of Bias Utilized for Surveys Tool (ROBUST) and includes a guide for the use of the tool with a corresponding coding sheet (see Supplementary Materials).

In what follows, we review the topics that the items address and their capacity to be adapted to the requirements of different systematic reviews and meta-analyses.

Sampling Frame

This item assesses the degree of correspondence between the theoretical population and the actual accessible population or list that is used to represent it, known as sampling frame (Warnecke, 2014). The question of adequate representativeness should be determined by the researcher conducting the meta-analysis, according to the assumptions concerning the population and investigated phenomenon. For example, in a study of employees in a given industry, the representativeness of a sample should be scored according to the extent to which the sampled companies represent the industry, as opposed to the sample's similarity to the general population.

Participant Recruitment

The method by which participants were recruited can influence several aspects related to research validity. For example, a snowball method can produce biased results, since it is a non-probability sampling technique (Bornstein, Jager, & Putnick, 2013). We note that, although the recruitment method may be related to the resultant sample's representativeness (discussed above), the two concepts are distinct, and existing risk-of-bias tools differentiate between them (Hoy et al., 2012). For example, while a stratified sample from a specific city might qualify as a recruitment method with low risk of bias, it might not necessarily be representative of the population of the country in which the city is located.

Acceptability of Exclusion Rate

In many studies, selected participants are excluded from the study or the analysis (Van Spall, Toren, Kiss, & Fowler, 2007). Exclusion might be due to reasons such as refusal to participate in the study or high percentage of missing data. Although this is an acceptable practice, a high exclusion rate may lead to biased results due to possible differences between excluded and included participants, and should therefore be treated with caution. It is important to note that this topic refers to exclusion unrelated to the required characteristics of the participants, which may be essential for properly representing the studied population (e.g., including only people with major depressive disorder).

Sufficiency of Sample Size

The meta-analytic calculation already takes the number of observations in a sample into consideration when assigning weights to effect sizes. Nevertheless, as the number of observations increases, the sample statistic becomes a better estimate of the true value corresponding to the overall population (Watt & van den Berg, 2002). This implies that a smaller sample is more likely to be affected by extreme and unrepresentative units from the population, and should thus be addressed in risk-of-bias measurements. Notably, the sample size that is considered sufficient may vary between systematic reviews. Therefore, while the question of whether a sample size is adequate should be addressed in risk-of-bias assessment, adequacy should be defined as a function of the effect size under investigation, statistical technique, and desired power (Guo, Chen, & Luh, 2019; Maas & Hox, 2005). These are often related to the research design, such as repeated measures necessitating fewer participants for an effect to be significant than a between-subjects design.

Demographic Variables

Although the characteristics of a given sample do not inherently constitute a source of bias, the absence of basic reporting associated with demographic variables may raise concerns pertaining to the validity of the results. For example, unreported properties concerning the participants' age distribution may lead to erroneous conclusions (Skelly, Dettori, & Brodt, 2012). It is not lack of reporting itself that increases the risk of bias, but the inability to assess the generalizability of the findings. For instance, if a study finds a significant result, but its sample consisted only of women, treating it as representing a phenomenon that exists in the entire population might be incorrect. Thus, demographic variables should be considered and specific criteria adopted when examining risk of bias. Moreover, it is difficult to establish precisely which demographic variables should be examined in risk-of-bias assessments. While age and gender are considered fundamental sample characteristics that should be reported, the relevance of additional variables depends on the research question and should be determined by the researcher (e.g.,

education, ethnicity, marital status). For example, reporting of religiosity might be of particular importance for certain research questions and in particular countries. Furthermore, the acceptable distribution within each characteristic can also be relevant to bias considerations. An approach that maintains high congruency between the sample and population characteristics is neither practical nor informative, since it would require statistically examining each characteristic of each primary study and it would be rare to find studies that are similar to the population on all demographic variables. Consequently, the acceptable distributions of the sample characteristics should be determined according to the research question, such as defining extreme distributions as indicative of risk of bias (e.g., when the proportion of men or women in the sample is under 20%).

Reliability of Measurements

Low reliability of a measure constitutes a barrier not only to its validity (Schutt, 2015) but also to the validity of the effect size under investigation. Consequently, results based on such a measure may be of higher risk of bias. It is important to note that reliability is a necessary but not sufficient indicator of measurement-related bias, since high reliability does not guarantee construct validity, i.e., that the proper construct is being measured (Reis & Judd, 2014). However, as is commonly practiced in research synthesis, we support establishing an inclusion rule for acceptable and valid measures as part of the systematic review process, leaving risk-of-bias measurement to depend solely on adequate internal consistency. This way, studies based on measures with unknown or inadequate construct validity will be excluded before the risk-of-bias assessment phase.

Setting

The setting in which a study is conducted can influence the risk of obtaining biased results. Controlled laboratory conditions generally represent the gold standard in terms of reducing external influences, whereas completion of surveys at an unsupervised location may increase the likelihood of careless responding (Meade & Craig, 2012). However, the complexity of the precise effect of the setting reveals that, in some cases, a sterile laboratory might decrease external validity (Berkowitz & Donnerstein, 1982). This discloses the need for a more adaptable answer to the question of appropriate setting, since studies included in systematic reviews may use a variety of settings and it may be difficult to determine which ones confer higher or lower risk of bias. One possibility for addressing this issue is to establish a general rule of thumb—e.g., that a monitored or more controlled location (such as a university classroom) has a lower risk of bias, whereas unknown locations have a higher one—and adjust it for specific circumstances. However, since online and paper-and-pencil surveys were sometimes shown to be essentially invariant (Davidov & Depner, 2011), while at other times differences were revealed for measures assessing sensitive content (Wood, Nosko, Desmarais, Ross,

& Irvine, 2006), it is important to bear in mind that bias may also depend on the nature of the phenomenon under investigation.

Data Management

Risk of bias related to managing data and reporting the procedures utilized in a given analysis can include several elements. In many contexts, addressing outliers, invalid data, and missing data can decrease the likelihood of obtaining biased results. However, in some cases, it may be superfluous to address certain issues, such as missing data when participants are presented with mandatory completion of closed-ended questions on a computer. Therefore, the researcher performing the meta-analysis should consider which elements need to be addressed in a primary study, and whether it is sufficient merely to report them (e.g., percentage of missing data) or whether it is necessary to utilize a specific technique that results in unbiased parameter estimates (Pohl & Becker, 2020) or takes these concerns into account, such as multiple imputations instead of mean imputations (Kleinke, 2018; Shrive, Stuart, Quan, & Ghali, 2006). This topic may be particularly relevant for longitudinal surveys and include an acceptable percentage of participants lost at follow-up.

Discussion

Consistent with previous studies, a consensus approach was applied to develop the new tool (e.g., Jarde et al., 2013). This led to a measure that included eight items, representing eight topics of risk of bias: the sampling frame, participant recruitment method, exclusion rate, sample size, demographic variables, measurement reliability, settings, and data management. In addition, a binary approach was implemented in the measure's response scale. This decision was based primarily on the importance given to ease-of-use, but also on the potential ambiguity accompanying judgments on more continuous scales (e.g., the difference between "yes" and "probably yes" related to a specific source of risk of bias) and on the fact that the inclusion of an intermediate option to code risk-of-bias items as "moderate" can hinder the selection of high or low risk (Hoy et al., 2012). Insufficient information in a given study concerning a particular item was regarded as conferring high risk of bias, since studies that utilize methods to minimize risk of bias are likely to report them. Additionally, many risk-of-bias tools that include an option to indicate insufficient reporting translate it into high risk of bias (Higgins & Green, 2008; Hoy et al., 2012). Following a pilot study, the new measure of risk of bias was finalized and labeled ROBUST.

Study 3: Evaluating Reliability and Validity

The purpose of this study was to empirically assess two fundamental attributes of a measure: reliability and validity. Reliability was assessed using concordance between raters, which is a standard means of measuring scale reproducibility (La Torre et al., 2006). Since agreement between raters changes as a function of experience (da Costa et al., 2017), data from experienced coders, whose discrepancies are mostly based on true error and not on inexperience in understanding risk of bias concepts or identifying them, should be used to provide detailed reliability analysis. Nevertheless, since researchers with varying degrees of experience may use the new measure, overall agreement among non-experienced raters was also examined. In addition, although there is no clear consensus regarding the conceptualization of validity (Camargo, Herrera, & Traynor, 2018), it was assessed by the common practice of calculating the association between overall ratings of studies using the new risk-of-bias measure and quality scores based on a different measure and coded by other raters.

Method

Two judges (the authors), both Professors with a PhD in Psychology and experience in methodology and meta-analyses, separately used the new tool to code the risk of bias of 70 empirical studies, collected for a meta-analysis in their field of expertise. Interrater reliability was assessed according to the percentage of agreement between raters and Cohen's Kappa coefficient, which takes into account the likelihood of chance agreement (McHugh, 2012).

To assess validity, a previous meta-analysis that examined associations between constructs using two time points was selected, due to similarities with the survey method for which the current measure is intended. The measure used in the meta-analysis was based on two checklists for assessing study quality (Theunissen et al., 2012). Two research assistants used the new risk-of bias measure to code three studies (with varying levels of quality) from the meta-analysis, followed by a discussion with the authors. Subsequently, the research assistants separately used the new tool to rate the risk of bias of 26 empirical studies. Discrepancies in their coding were resolved by the first author, and Pearson correlation coefficient was calculated between the final scores of the new risk-of-bias measure and those previously published in the meta-analysis.

Results

Interrater agreement among experienced researchers on individual items ranged from 91% to 100%, with Kappa coefficients ranging from .81 to 1.00, all significant at $p < .001$ (Table 1). The overall percentage of agreement between the independent raters across the eight items was 96% (538/560) with a Kappa coefficient of .93, $p < .001$, indicating an excellent agreement level (Fleiss, Levin, & Cho Paik, 2003). Interrater agreement on the

summary score was calculated using the Weighted Kappa coefficient for ordinal variables (Cohen, 1968) and yielded a high agreement of .82, $p < .001$, 95% CI [.75, .89].

Table 1

Interrater Agreement on Items of the New Risk-of-Bias Assessment Tool

No	Item [criterion used in current assessment]	Proportion of agreement	Kappa coefficient	95% CI of Kappa coefficient	
				LL	*UL*
1.	Sample frame representative? [yes = of general population]	0.96	0.88	0.74	1.01
2.	Appropriate participant recruitment? [yes = random selection or stratified sample]	0.94	0.82	0.66	0.99
3.	Adequate exclusion rate of participants? [< 20%]	0.97	0.90	0.76	1.04
4.	Acceptable final sample size? [> 100]	1.00	1.00	1.00	1.00
5.	Reporting of sample characteristics? [age and gender; yes = both reported]	0.99	0.96	0.89	1.03
6.	Measures with adequate reliability? [average r > 0.25, e.g., $\alpha > 0.7$ for 7 items]	0.93	0.81	0.65	0.97
7.	Controlled setting? [yes = controlled environment, e.g., lab]	0.91	0.82	0.68	0.96
8.	Acceptable data management? [addressing missing data, outliers, and invalid responses; yes = reporting of at least one of them]	0.99	0.93	0.78	1.07

Note. All Kappa values are significant at $p < .001$.

Interrater agreement among inexperienced raters across all individual items was 88% (184/208). Interrater agreement on the summary score was 58% with a Weighted Kappa coefficient of .64, $p < .001$, 95% CI [.44, .84], indicating adequate reliability. The correlation between the ratings of the new measure and those provided by the meta-analysis was $r = .66$, $N = 26$, $p < .001$, representing high convergent validity.

Discussion

The new risk-of-bias measure demonstrated excellent interrater agreement among experienced researchers, both across items and when using a summary score. This provides verification for the reliability of the measure and supports the probability of its results being accurately replicated. The agreement levels were comparable to those found in previous risk-of-bias measures, such as an overall Kappa of .75 (Jarde et al., 2013) and of .82 (Hoy et al., 2012). The lower reliability of inexperienced raters compared to experienced ones has been previously reported (da Costa et al., 2017). Nevertheless, the inexperienced

raters also reached substantial agreement (Landis & Koch, 1977). Moreover, their coding was used to assess validity, demonstrating a high correspondence between the scores of the new risk-of-bias tool and those of a previous measure.

General Discussion

Although meta-analysis is a reliable technique for synthesizing numerical results, it is only as valid as the primary studies it includes. However, there is a shortage of measures addressing possible bias concerning certain research questions and designs. Consequently, the assessment of survey research has been described as an art more than a science (Gauthier, 2001). Herein, we sought to develop a reliable, valid and accessible tool—the ROBUST—for assessing the risk of bias of primary studies in systematic reviews related to survey designs.

Measurement Validity, Reliability, and Utility

The items included in the current tool were collected from well-known and substantiated measures used to evaluate risk of bias or quality of study, and thus correspond to issues that have already been established as relevant for such an assessment. Specifically, measurement reliability can be found in the GRACE checklist (Dreyer et al., 2014); sample representativeness and participant recruitment in Hoy et al. (2012); exclusion rate in the MORE checklist (Shamliyan et al., 2010); setting and data management in the STROBE checklist (Vandenbroucke et al., 2007); and demographic variables and sample size in Faragher et al. (2005). This fact suggests high face validity (e.g., Conybeare, Behar, Solomon, Newman, & Borkovec, 2012), which is considered particularly important during the development phase of a measure (Menold, Bluemke, & Hubley, 2018). In addition, the empirical assessment of validity (i.e., criterion-related validity), which examines the correspondence between a given measure and an external criterion (Bhattacherjee, 2012), was also demonstrated. Together with the new tool's high interrater agreement, both its validity and reliability attests to the ROBUST's utility for risk-of-bias assessment.

A unique feature of the new tool, aside from being tailored to systematic reviews of surveys, is the emphasis placed on ease-of-use. This is achieved through the incorporation of a relatively small number of core elements to be assessed, the use of a direct binary approach for coding each item, and the derivation of one score for the general evaluation of a study. However, it should be noted that a low score might be related to issues such as the sampling of specific populations or lack of reporting, and therefore does not necessarily imply that the information provided by that study is not important. It merely represents the confidence in the study's findings in a particular meta-analytic frame of reference. The risk-of-bias scores can be used, for example, to evaluate changes in the summary effect due to possible bias considerations. Moreover, if

a researcher is particularly interested in evaluating the effect of a specific item from the tool on the meta-analytic results (e.g., setting), it is possible to use it as a moderator and test whether the summary effects differ significantly between sub-groups (Borenstein, Hedges, Higgins, & Rothstein, 2009).

General Scores, Items and Response Scales: Considerations and Suggestions

The purpose of the tool developed herein was to aid researchers conducting systematic reviews and meta-analysis of surveys. However, the items included in the tool should be viewed as guidelines, and not as strict rules. While several basic elements of validity are shared by all empirical studies, the precise methodological characteristics that need to be considered depend on the nature of the question under investigation and the types of research designs (Cooper, 2010). In an attempt to address this complexity, the ROBUST was designed as a flexible generic measure that addresses key validity elements and accommodates modifications to the items it includes as well as to the criteria for determining high or low risk of bias. Thus, researchers can easily create a domain-specific version of the tool, tailored to their needs.

The items currently included in the risk-of-bias tool encompass various core issues, since they are intended to be relevant to a variety of questions and research designs. However, with proper justification, research syntheses that include certain research designs or goals may require the estimation of additional elements of risk of bias. For example, when a large portion of the people invited to participate in a study fails to respond, it might indicate a variety of problems related to the topic, the phrasing, the timing, or the method. This phenomenon is known as non-response bias, and can result in a situation in which certain segments of the target population are under-represented (Berg, 2005). However, this bias is only relevant to studies that attempt to contact all participants from a particular list, and it does not apply to many common techniques used in the Behavioral and Social Sciences (e.g., posting an invitation to participate in a study on a university billboard or sending a collective request using online panels). Therefore, a meta-analysis of studies including known finite lists of possible participants (e.g., patients in a hospital in a certain time period) should add a response rate criterion. Another example relates to studies that typically favor two time points of assessment, such as surveys in Organizational Psychology. This might merit a criterion related to the use of more than a single assessment or to the time between assessments. However, it is important to note that adding criteria to the ROBUST might present a new element of bias when interpreting its results, and decrease the ability to reliably compare scores across meta-analyses. Consequently, such modifications should be made with caution.

Regarding the criteria for identifying risk of bias within each item, although some commonly acceptable guidelines are provided (e.g., random selection as a recruitment method with low risk of bias), we recommend that researchers adjust their criteria

to the scientific phenomenon under investigation and use exemplary studies to help determine the proper standards for their assessment. For example, if mean differences of two independent samples are examined, the minimal sample size for detecting a medium effect size (Cohen's $d = 0.5$) should be $N = 126$ (assuming equal group sizes, 80% power, and 5% significance level). However, if dependent samples are examined under the same conditions, the minimal sample size should be $N = 34$. In addition, to achieve high interrater agreement, it is imperative to establish clear and detailed cut-off points that address various potential decisions pertinent to each item. However, it may be difficult to determine dichotomous decision rules for topics that can include several components, such as data management considerations. A possible solution would be to break down the data management item into sub-items, such as outliers, invalid data, and missing data, and evaluate each one of them separately. A final score would then be assigned to the general item by either calculating the mean of the sub-item scores or employing some type of decision rule (e.g., the item is evaluated as indicating low risk of bias if at least two out of three sub-items are characterized by low risk of bias).

The current manuscript proposes using the final score related to a study as its risk-of-bias indicator. Nevertheless, different approaches regarding how meta-analyses should apply the results of risk-of-bias assessments exist, such as best-evidence synthesis, in which a meta-analysis only includes high-quality studies (Slavin, 1986). However, such an approach would not allow risk-of-bias scores to be incorporated interactively into meta-analytic modeling as a systematic sensitivity analysis (Johnson et al., 2015). Another option suggests evaluating grade of evidence: defining a "good study" as one that meets all criteria, a "fair study" as one that does not meet all criteria but is judged to have no fatal flaw, and a "poor study" as one that contains a fatal flaw (Harris et al., 2001). If deemed appropriate for the purpose of a given meta-analysis, the current tool can be altered to reflect whether the risk of bias associated with a given study exceeds a particular threshold for study inclusion, or to emphasize particular elements that would indicate major, minor, or no flaws.

Limitations

One shortcoming of the current research is that it was conducted only by researchers in the field of Psychology and Health. However, it is important to note that the measures that form the basis for the current risk-of-bias tool were developed by large teams of scientists from various fields, including epidemiologists and sociologists, suggesting proper representation of topics related to such fields. Another limitation refers to topics that are not represented in the new measure, particularly funding and conflict of interests. While some studies have found an association between funding source and methodological quality (Mandrioli, Kearns, & Bero, 2016), other have not (Jefferson, Di Pietrantonj, Debalini, Rivetti, & Demicheli, 2009). However, the current measurement was designed to assess only core methodological sources of risk of bias, and not non-methodological

issues related to conflict of interests that could influence the outcome of a study, such as personal, academic or political interests, which are rarely reported (Jarde et al., 2013).

In developing the new risk-of-bias tool, we recognized that a rigid and extensive measure would be neither accessible nor suitable in various circumstances. Therefore, we assigned high priority to ensuring ease-of-use and flexibility to adapt to the requirements of different research questions and designs. However, these considerations come at a price: ease-of-use leads to the inclusion of a limited number of items that employ simple decision rules, and flexibility can lead to changes in the assessment tool that increase uncertainty in the interpretation of its results. Comprehensive lists of items with continuous scales that employ uniform measures with predetermined criteria regarding likelihood of bias have the ability to address elements of risk-of-bias more accurately and with increased standardization. Nevertheless, there are also disadvantages associated with comprehensive, uniform measures with more continuous scales: familiarizing and applying them may require tremendous amounts of effort, and intermediate scale options can add random error and hinder the selection of high or low risk (Hoy et al., 2012). In addition, they might include assessment of irrelevant information while excluding information relevant for particular meta-analyses (Shamliyan et al., 2010). These points illustrate the difficulty in developing a measure that addresses all the shortcomings associated with the different options of assessing risk of bias, and may explain why such assessments are reported only in approximately 1% of meta-analyses in certain fields (Schalken & Rietbergen, 2017). Subsequently, we believe that the advantages of an accessible tool pertaining to issues relevant to surveys outweigh the disadvantages, and that the use of a valid and reliable tool that addresses core elements of risk of bias would be a considerable improvement over disregarding such risk or using tentative rules. Furthermore, while we acknowledge the imperfect nature of the proposed risk-of-bias measure, many well-known measures also utilize dichotomous scales (Harris et al., 2001), comprise short lists (Hoy et al., 2012), and leave important decisions up to the meta-analyst, such as determining the proportion of missing data that is considered "enough to be confident of the findings" (Higgins et al., 2016). Consequently, since different meta-analyses may have unique needs and considerations, this study follows previous recommendations for tools to provide both rigorous quality assessment and justified definitions of research-specific quality standards (Shamliyan et al., 2010). Although such an approach may lead to a certain degree of discrepancy, the current tool explicitly addresses this concern and advocates reporting of the assessment criteria, thereby allowing the scientific community to evaluate the appropriateness of the decisions related to risk of bias in the context of a particular meta-analysis. Nevertheless, future studies should properly weigh the benefits and costs of the new tool to identify the optimal balance between them, as well as to validate and examine its capacity to adjust to the needs of different meta-analyses. Moreover, the ROBUST should evolve alongside scientific understanding of research. For example, with a growing concern about replication

(Hedges, 2019) and p-hacking, i.e. collecting or selecting data or statistical analyses until significant results are reached (Head, Holman, Lanfear, Kahn, & Jennions, 2015), future versions of the risk-of-bias measure could include items addressing these issues as part of a primary study's risk assessment.

Conclusions

The current research sought to develop a generic tool to assess risk of bias in primary studies for systematic reviews and meta-analyses synthesizing effect sizes across studies based on surveys. Although the implications of bias in such studies or even the extent of its prevalence are unclear, it may pose a threat to the validity of meta-analytic conclusions (Johnson et al., 2015). Accordingly, it is important for meta-analysts to account for possible risk-of-bias effects, even if doing so adds another layer of complexity to the process (La Torre et al., 2006). To address this issue, the new tool proposed herein is specifically designed to be accessible and relevant to various research questions for which there is a scarcity of such measures. This would allow additional scientists to incorporate risk-of-bias measures into their systematic reviews, thus providing greater confidence in the products of statistical procedures of knowledge-building.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: The authors have no support to report.

Supplementary Materials

For this article the following supplementary materials are available (Nudelman & Otto, 2020):

- Risk of Bias Utilized for Surveys Tool (ROBUST) - coding guide.
- Risk of Bias Utilized for Surveys Tool (ROBUST) - coding sheet.

Index of Supplementary Materials

Nudelman, G., & Otto, K. (2020). *Supplementary materials to: The development of a new generic risk-of-bias measure for systematic reviews of surveys* [Code]. PsychOpen. https://doi.org/10.23668/psycharchives.4415

References

Berg, N. (2005). Non-response bias. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 865–873). London, United Kingdom: Academic Press.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *The American Psychologist, 37*(3), 245-257. https://doi.org/10.1037/0003-066X.37.3.245

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. Retrieved from http://scholarcommons.usf.edu/oa_textbooks/3

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, United Kingdom: John Wiley & Sons. https://doi.org/10.1002/9780470743386

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review, 33*(4), 357-370. https://doi.org/10.1016/j.dr.2013.08.003

Camargo, S. L., Herrera, A. N., & Traynor, A. (2018). Looking for a consensus in the discussion about the concept of validity. *Methodology, 14*(4), 146-155. https://doi.org/10.1027/1614-2241/a000157

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220. https://doi.org/10.1037/h0026256

Conybeare, D., Behar, E., Solomon, A., Newman, M. G., & Borkovec, T. D. (2012). The PTSD Checklist—Civilian Version: Reliability, validity, and factor structure in a nonclinical sample. *Journal of Clinical Psychology, 68*(6), 699-713. https://doi.org/10.1002/jclp.21845

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA, USA: SAGE publications.

da Costa, B. R., Beckett, B., Diaz, A., Resta, N. M., Johnston, B. C., Egger, M., . . . Armijo-Olivo, S. (2017). Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: A prospective study. *Systematic Reviews, 6*(1), Article 44. https://doi.org/10.1186/s13643-017-0441-7

Davidov, E., & Depner, F. (2011). Testing for measurement equivalence of human values across online and paper-and-pencil surveys. *Quality & Quantity, 45*(2), 375-390. https://doi.org/10.1007/s11135-009-9297-9

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377-384. https://doi.org/10.1136/jech.52.6.377

Dreyer, N. A., Velentgas, P., Westrich, K., & Dubois, R. (2014). The GRACE Checklist for rating the quality of observational studies of comparative effectiveness: A tale of hope and caution. *Journal of Managed Care & Specialty Pharmacy, 20*(3), 301-308. https://doi.org/10.18553/jmcp.2014.20.3.301

Faragher, E. B., Cass, M., & Cooper, C. L. (2005). The relationship between job satisfaction and health: A meta-analysis. *Occupational and Environmental Medicine, 62*(2), 105-112. https://doi.org/10.1136/oem.2002.006734

Fleiss, J., Levin, B., & Cho Paik, M. (2003). *Statistical methods for rates and proportions*. Hoboken, NJ, USA: John Wiley & Sons.

Gauthier, B. (2001, May 17-20). *Assessing survey research a principled approach*[Paper presentation]. The 2001 American Association for Public Opinion Research Annual Conference, Montréal, Canada. Retrieved from https://ssl.circum.com/textes/assessing_aapor_20010519.pdf

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82. https://doi.org/10.1177/1525822X05279903

Guo, J. H., Chen, H. J., & Luh, W. M. (2019). Optimal sample sizes for testing the equivalence of two means. *Methodology, 15*(3), 128-136. https://doi.org/10.1027/1614-2241/a000171

Harris, R. P., Helfand, M., Woolf, S. H., Lohr, K. N., Mulrow, C. D., Teutsch, S. M., & Atkins, D. (2001). Current methods of the U.S. Preventive Services Task Force: A review of the process. *American Journal of Preventive Medicine, 20*(3), 21-35. https://doi.org/10.1016/S0749-3797(01)00261-6

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology, 13*(3), Article e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V. (2019). The statistics of replication. *Methodology, 15*(Suppl. 1), 3-14. https://doi.org/10.1027/1614-2241/a000173

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, United Kingdom: John Wiley & Sons. https://doi.org/10.1002/9780470712184

Higgins, J. P. T., Sterne, J. A., Savović, J., Page, M., Hróbjartsson, A., Boutron, I. . . . Eldridge, S. (2016). A revised tool for assessing risk of bias in randomized trials. In J. Chandler, J. McKenzie, I. Boutron, & V. Welch (Eds.), *Cochrane methods 2016: Cochrane database of systematic reviews* (pp. 29-31). https://doi.org/10.1002/14651858.CD201601

Hoy, D., Brooks, P., Woolf, A., Blyth, F., March, L., Bain, C., . . . Buchbinder, R. (2012). Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *Journal of Clinical Epidemiology, 65*(9), 934-939. https://doi.org/10.1016/j.jclinepi.2011.11.014

Jarde, A., Losilla, J.-M., Vives, J., & Rodrigo, M. F. (2013). Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology, 13*(2), 138-146. https://doi.org/10.1016/S1697-2600(13)70017-6

Jefferson, T., Di Pietrantonj, C., Debalini, M. G., Rivetti, A., & Demicheli, V. (2009). Relation of study quality, concordance, take home message, funding, and impact in studies of influenza vaccines: Systematic review. *BMJ, 338*, Article b354. https://doi.org/10.1136/bmj.b354

Johnson, B. T., Low, R. E., & MacDonald, H. V. (2015). Panning for the gold in health research: Incorporating studies' methodological quality in meta-analysis. *Psychology & Health, 30*(1), 135-152. https://doi.org/10.1080/08870446.2014.953533

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology, 14*(1), 3-15. https://doi.org/10.1027/1614-2241/a000141

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

La Torre, G., Chiaradia, G., Gianfagna, F., Boccia, S., De Laurentis, A., & Ricciardi, W. (2006). Quality assessment in meta-analysis. *Italian Journal of Public Health, 3*(2), 44-50. https://doi.org/10.2427/5937

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92. https://doi.org/10.1027/1614-2241.1.3.86

Mandrioli, D., Kearns, C. E., & Bero, L. A. (2016). Relationship between research outcomes and risk of bias, study sponsorship, and author financial conflicts of interest in reviews of the effects of artificially sweetened beverages on weight outcomes: A systematic review of reviews. *PLOS ONE, 11*(9), Article e0162198. https://doi.org/10.1371/journal.pone.0162198

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. https://doi.org/10.1037/a0028085

Menold, N., Bluemke, M., & Hubley, A. M. (2018). Validity: Challenges in conception, methods, and interpretation in survey research [Editorial]. *Methodology, 14*(4), 143-145. https://doi.org/10.1027/1614-2241/a000159

NIH. (2017). *Quality assessment tool for observational cohort and cross-sectional studies.* Retrieved from https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools/cohort

Nudelman, G. (2013). The belief in a just world and personality: A meta-analysis. *Social Justice Research, 26*(2), 105-119. https://doi.org/10.1007/s11211-013-0178-y

Pohl, S., & Becker, B. (2020). Performance of missing data approaches under nonignorable missing data conditions. *Methodology, 16*(2), 147-165. https://doi.org/10.5964/meth.2805

Protogerou, C., & Hagger, M. S. (2019). A case for a study quality appraisal in survey studies in psychology. *Frontiers in Psychology, 9*, Article 2788. https://doi.org/10.3389/fpsyg.2018.02788

Reis, H. T., & Judd, C. M. (Eds.). (2014). *Handbook of research methods in social and personality psychology* (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.

Reitsma, J. B., Rutjes, A. W. S., Whiting, P., Vlassov, V. V., Leeflang, M. M., & Deeks, J. J. (2009). Chapter 9: Assessing methodological quality. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0.* Retrieved from https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/ch09_Oct09.pdf

Schalken, N., & Rietbergen, C. (2017). The reporting quality of systematic reviews and meta-analyses in industrial and organizational psychology: A systematic review. *Frontiers in Psychology, 8*, Article 1395. https://doi.org/10.3389/fpsyg.2017.01395

Schutt, R. K. (2015). *Investigating the social world: The process and practice of research* (8th ed.). Thousand Oaks, CA, USA: SAGE Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA, USA: Houghton Mifflin.

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology, 63*(10), 1061-1070. https://doi.org/10.1016/j.jclinepi.2010.04.014

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

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology, 6*, Article 57. https://doi.org/10.1186/1471-2288-6-57

Skelly, A. C., Dettori, J., & Brodt, E. (2012). Assessing bias: The importance of considering confounding. *Evidence-Based Spine-Care Journal, 3*(1), 9-12. https://doi.org/10.1055/s-0031-1298595

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*(9), 5-11. https://doi.org/10.3102/0013189X015009005

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

Theunissen, M., Peters, M. L., Bruce, J., Gramke, H.-F., & Marcus, M. A. (2012). Preoperative anxiety and catastrophizing: A systematic review and meta-analysis of the association with chronic postsurgical pain. *The Clinical Journal of Pain, 28*(9), 819-841. https://doi.org/10.1097/AJP.0b013e31824549d6

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130-149. https://doi.org/10.1037/1082-989X.13.2.130

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., . . . Egger, M. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLOS Medicine, 4*(10), Article e297. https://doi.org/10.1371/journal.pmed.0040297

Van Spall, H. G. C., Toren, A., Kiss, A., & Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *Journal of the American Medical Association, 297*(11), 1233-1240. https://doi.org/10.1001/jama.297.11.1233

Warnecke, R. B. (2014). Sampling frames: Overview. *Wiley StatsRef: Statistics Reference Online.* https://doi.org/10.1002/9781118445112.stat05715

Watt, J. H., & van den Berg, S. A. (2002). *Research methods for communication science.* Albany, NY, USA: Rensselaer Polytechnic Institute.

Wells, G. A., Shea, B., O'Connell, D., Petersen, J., Welch, V., Losos, M., & Tugwell, P. (2000). *The Newcastle-Ottawa Scale for assessing the quality of nonrandomised studies in meta-analyses.* Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Wood, E., Nosko, A., Desmarais, S., Ross, C., & Irvine, C. (2006). Online and traditional paper-and-pencil survey administration: Examining experimenter presence, sensitive material and long surveys. *The Canadian Journal of Human Sexuality, 15*(3), 147-155.