Methods in Psychology 3 (2020) 100031

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.journals.elsevier.com/methods-in-psychology

A checklist to assess the quality of survey studies in psychology

Cleo Protogerou [a,b,*], Martin S. Hagger [a,c]

[a] University of California, Merced, USA
[b] University of Cape Town, South Africa
[c] University of Jyväskylä, Finland

ARTICLE INFO

Keywords:
Study quality assessment
Psychology
Correlational studies
Survey studies
Evidence syntheses
Transparency

ABSTRACT

Study quality is emerging as an essential component of evidence syntheses. It allows practitioners and policymakers to make informed decisions based on the quality of the evidence reviewed. Study quality is typically assessed by checklists of pre-determined quality criteria. Few study quality checklists have been systematically evaluated, and none have been developed specifically for survey studies in psychology. The present study addresses this evidence gap by developing the quality of survey studies in psychology (Q-SSP) checklistusing an expert-consensus method. An international panel of experts in psychology research and quality assessment ($N = 53$) evaluated the inclusion and importance of candidate quality items and offered commentary. The resulting checklist was used to evaluate a set of survey studies and inter-rater reliability of checklist scores was computed. A preliminary test of criterion validity of checklist scores was conducted using on a sample of survey studies with 'known differences' in study quality verified by experts. Experts exhibited high agreement on inclusion and importance ratings of the candidate items. Minor adjustments were made to the candidate items based on experts' feedback. Inter-rater reliability of study quality scores using the checklist was high. Some evidence for criterion validity of scores using the checklist was obtained. Overall, we provide preliminary data to support the Q-SSP checklist as a potential means to evaluate the quality of survey studies in psychology. We recommend a future large-scale study using the Q-SSP checklist to assess study quality in studies with known differences in quality verified by experts.

As research evidence for psychological phenomena accumulates, scientific communities, stakeholders, and policymakers are becoming increasingly reliant on research syntheses, such systematic reviews and meta-analyses, to provide pithy summaries of effects of interest, and to inform evidence-based policy and practice. While innovation in methods and analytic techniques for evidence syntheses provides increasingly sophisticated means to summarize research and test effects of interest, these methods are highly dependent on the quality of the evidence included in the analyses. Ways to evaluate the quality of research evidence for evidence syntheses are therefore increasingly recognized as essential components of evidence syntheses (Greenhalgh and Brown, 2017; Higgins and Altman, 2008; Lipsey and Wilson, 2001). Coupled with the imperative of assessing study quality for syntheses of research, there is also an increased need to evaluate the quality of individual studies. Study quality assessment can facilitate comparisons across individual studies and optimize the quality of future studies and their replication.

Study quality is typically assessed using checklists, in which trained reviewers assess studies on a set of pre-determined quality components. Numerous study quality checklists or 'tools' exist (e.g., Higgins et al., 2011; Jadad et al., 1996; Oxman and Guyatt, 1988). However, to date, no tool has been developed for the expressed purpose of assessing the quality of psychological research, and researchers in psychology have fulfilled the need for study quality assessment by adapting existing quality measures originally developed in other disciplines (e.g., Hagger et al., 2017; Husebø et al., 2012; Protogerou et al., 2018). As these tools have not been specifically developed to evaluate psychological studies, they may lack validity and provide insufficient coverage of the appropriate study quality components.

The purpose of the present study is to fill this evidence gap by developing a study quality tool for psychological studies using survey designs. We focus on survey research as it is one of the predominant methods of research in psychology (Singleton and Straits, 2009). Furthermore, studies adopting survey methods are frequently the subject

* Corresponding author. Psychological Sciences and Health Sciences Research Institute (HSRI), University of California, Merced, USA.
E-mail address: cprotogerou@ucmerced.edu (C. Protogerou).

https://doi.org/10.1016/j.metip.2020.100031
Received 25 December 2019; Received in revised form 6 June 2020; Accepted 14 July 2020
Available online 17 July 2020

of research syntheses in psychology (Hunter and Schmidt, 2004). Specifically, the present study aimed to develop tool for researchers to assess psychology survey studies using an expert consensus approach. The primary focus of the study was to establish the Q-SSP checklist as a means to provide assessments of study quality with evidence for the face and content validity of its items, as well as inter-rater reliability, and a secondary focus was to provide preliminary evaluation of the criterion validity of Q-SSP checklist scores with a goal of establishing whether the tool was effective in differentiating between studies of acceptable (or higher) versus questionable (or lower) quality. We expect the tool to improve the precision of research syntheses by enabling researchers to incorporate assessment of study quality as a key component of the sample of studies under review, and test effects of study quality on findings of the syntheses. We also anticipate the tool will inform the development of higher quality survey studies and research syntheses, as well as replications, by highlighting deficiencies in currently-available studies.

Study quality: definitions and assessment

Study quality reflects the extent to which a study has taken appropriate measures to minimize bias and error from inception to reporting of findings (Khan et al., 2011). It has been estimated that only approximately 20% of published studies across fields of behaviour health research are of sufficient quality (Ciliska and Buffet, 2008). Assessment of study quality[1] – also known as *critical appraisal* – is the systematic evaluation of the degree to which a study has been conducted to the highest possible quality standards (Higgins and Green, 2008). A study of acceptable quality provides assurances that the research was conducted in line with a set of pre-specified discipline-appropriate standards, and that findings may be legitimately generalized to populations of interest and implemented in practice. Therefore, research that has been assessed as having ‘good’ quality based on a formal appraisal against specified quality standards, may allow researchers, clinicians, policy makers, and other interested stakeholders to make informed decisions based on the available evidence (Oxman and Guyatt, 1991). Other than providing a means to identify the strengths and weaknesses of a body of evidence, assessment of study quality also entails a number of other outcomes such as: selection of studies for inclusion in evidence syntheses; identifying potential sources of bias in the results of evidence syntheses; and gauging the impact of study quality on the results of a meta-analysis by incorporating study quality in subgroup and sensitivity analyses. Finally, quality assessment can also assist in improving research and publication standards by highlighting common deficiencies in the available evidence and possible means to improve the quality of subsequent studies (Greenhalgh and Brown, 2017; Greenhalgh, 2014; Johnson et al., 2014).

Numerous checklists or ‘tools’ designed to assess the quality of research study have been developed. Although there is idiosyncratic variability in content across the available tools, there is a degree of commonality in the general categories of quality criteria adopted. Typical categories of quality components relate to the population under investigation (e.g., sampling and recruiting strategies, sample size); study design (e.g., ‘appropriateness’ of methodology, ethical review procedures); data collection (e.g., validation of instrument/measures used, detailed descriptions of data collection process); data analyses (e.g., ‘appropriateness’ of statistical tests employed, dealing with attrition); and reporting and interpretation of results (e.g., completeness of results reported, suggestions for further research and practice) (see Crowe and Sheppard, 2011; Durant, 1994; Glynn, 2006 for examples of quality criteria used).

Reviews of the literature have identified nearly 200 tools used to assess study quality across health and social sciences research (Deeks et al., 2003; Katrak et al., 2004). Extant tools have been developed to appraise experimental studies (e.g., Jadad et al., 1996), systematic reviews and meta-analyses (e.g., Oxman and Guyatt, 1988, 1991), and qualitative studies (e.g., Zhang et al., 2019). Generic quality appraisal tools also exist (e.g., Glynn, 2006; National Institutes of Health, 2014). It has been argued, however, that most quality assessment tools have not been developed with sufficient scientific rigor (Crowe and Sheppard, 2011; Johnson et al., 2014; Katrak et al., 2004; Khan et al., 2011; Moyer and Finney, 2005). A long-standing argument against the standing of extant critical appraisal tools that has still yet to be resolved is that the tools omit key quality domains (Crowe and Sheppard, 2011; Deeks et al., 2003), and that no tool can be recommended without reservation (Alderson et al., 2003).

The most prominent criticisms of extant tools relate to the absence of validity and reliability checks in their development. For example, in their review of 44 published quality appraisal tools, Crowe and Sheppard (2011) found that only six tools had been tested for concurrent validity, only two for construct validity, and only 12 for reliability. A further 11 tools had not been tested for any type of validity or reliability, and 17 tools provided no explanation on how they were developed and did not include details on how they should be administered and scored. Recommendations for developing credible quality assessment tools have highlighted the need to systematically identify relevant domains of study quality, include appropriate validity and reliability checks, account for discipline-specific research principles, and provide a guide with precise explanations of the terms used and scoring strategies (Crowe and Sheppard, 2011; Moyer and Finney, 2005).

In addition, we note the absence of quality assessment tools designed specifically for survey research in psychology.[2] As survey research is one of the most frequently-used methods in psychology (Ponto, 2015; Singleton and Straits, 2009), a dedicated, fit-for-purpose quality tool is needed (Protogerou and Hagger, 2019). The lack of a tool to evaluate survey research has been noted by prominent methodologists in the field (Faragher et al., 2005; Hoffmann et al., 2017). Given the absence of relevant tools, researchers have adapted tools from disciplines outside psychology, or developed bespoke tools, in order to evaluate study quality (e.g., Faragher et al., 2005; Hagger et al., 2017; Hoffmann et al., 2017; Young et al., 2014). One problem with these adapted tools is that they reflect a ‘retrofitting’ of tool content designed to assess the quality of other types of research and in other fields (e.g., medicine, health sciences), and, as a consequence, they are often lacking in some way. Such adapted tools may omit essential criteria relevant to psychology or survey methods, or the criteria are not specified in such a way that they relevant to, or sufficiently tailored to, the particular discipline. For example, many research quality checklists include items relating to participant recruitment or sampling methods, but this is often in the context of randomized-controlled or cross-sectional designs in settings like medical research, few make explicit reference to the specific information necessary to judge the quality of the recruitment/sampling methods in psychology survey studies (e.g., rates of participants refusing an initial invitation to participate, rates of attrition from survey completion). This makes the development of a psychology discipline-specific tool that satisfies the specific criteria for studies adopting survey methods essential for comprehensive coverage of study quality assessment in this domain.

Development of discipline- and method-specific quality assessment tools is also important to ensure consistency in the criteria content and ratings of studies. The absence of a discipline- and method-specific tool that provides valid and reliable scores on study quality means researchers

[1] Study quality should be differentiated from *risk of bias*, a related concept which reflects the extent to which systematic error in research will lead researchers to draw incorrect conclusions from the findings (Higgins and Altman, 2008).

[2] Our definition of survey research is based on that provided by Check and Schutt (2012): “the collection of information from a sample of individuals through their responses to questions” (p. 160).

have fallen back on the use of multiple, diverse tools with idiosyncratic content to assess study quality. This presents a considerable challenge to researchers attempting to assess study quality across multiple studies, such as in the context of systematic reviews and meta-analyses, overviews of psychology survey research. The application of diverse tools to the same body of evidence may result in researchers arriving at different conclusions on the quality of the evidence, which can have ramifications for subsequent interpretations of the evidence. For example, variation in quality assessment scores may influence effect sizes across moderator groups defined by methodological quality scores in meta-analyses and affect conclusions drawn (Protogerou and Hagger, 2019). The imperative of precisely and reliably distinguishing between studies of acceptable and questionable quality in survey studies in psychology and the deficiencies of retrofitted tools highlights the need for a purpose-developed study quality tool.

Study overview

Recognizing the challenges presented by the lack of a dedicated tool to assess the quality of survey studies in psychology, we aimed to develop a tool to assess the quality of survey studies in psychology. The tool, the quality of survey studies in psychology (Q-SSP) checklist, was based on a comprehensive review of previous methodological quality assessment tools and checklists followed by an expert consensus method to evaluate and refine its content. The purpose of expert consensus methods is to define levels of agreement in a wide variety of settings, especially when insufficient or conflicting evidence exists (Fink et al., 1991; Jones and Hunter, 1995). Expert agreement pools the collective expertise of those with in-depth knowledge and training applied to the subject of interest (Hasson et al., 2000; Michie et al., 2017). Although it is acknowledged that variation and disagreements will occur in expert ratings, the consensus approach provides a summary of the convergence of expert knowledge. Furthermore, consensus approaches capitalize on the accumulated knowledge and practical experience of experts to obtain information that is culturally apt and rapidly implemented (Minas and Jorm, 2010; Stephens et al., 2017). For example, expert consensus studies have been used broadly across many disciplines to develop content of instruments and measures based on the pooled expertise in research (Herdman et al., 2002; Michie et al., 2005, 2013; Stephens et al., 2017; Velligan et al., 2010), including the development of quality appraisal tools (Burnett et al., 2005; Jadad et al., 1996; Pace et al., 2012).

Expert selection is a pertinent issue when it comes to using expert consensus to judge the validity of the content of measures and tools. While there is no established definition of an ‘expert’ in a particular field or discipline, or rule as to who should be included as an expert in an expert consensus panel, some published guidelines exist. Broadly, it is recommended that experts are pooled from “relevant, backgrounds, and experiences”, “pertinent specialties”, and, when appropriate, members of relevant advocate groups and general public (Fink et al., 1984; Hsu and Sandford, 2007). Other recommendations suggest identifying experts through their involvement in relevant research and authorship of relevant publications (Addington et al., 2013; Yap et al., 2014). In addition to the criterion of ‘relevance’ of experts' experience to the phenomenon under investigation, the ‘diversity’ of experts has also been proposed as important; extant consensus studies have aimed to include experts from diverse geographical locations, professional ranks, genders, and age groups (Jorm, 2015). A common theme in the literature is that expertise and selected experts should be guided by the questions, aims, and needs of the consensus study in question (Jorm, 2015). Our definition of ‘expertise’ is consistent with these extant practices in the literature using expert consensus methods.

The Q-SSP checklist was developed in four stages. First, we developed an initial set of candidate quality items for the checklist based on a review of existing study quality appraisal tools and recommendations of consensus statements on quality requirements in psychology (Appelbaum et al., 2018; Asendorpf et al., 2013; Finkel et al., 2017). Based on this review, we developed clear language descriptions and assessment criteria for each item. Second, we used an expert consensus method to provide external, independent evaluations of the candidate set of quality items. A panel of experienced researchers with expertise in survey research, evidence synthesis, and quality appraisal evaluated the initial item set, descriptions, and assessment criteria in terms of their necessity, appropriateness, and importance for inclusion. Third, the tool was refined based on the results of the expert consensus ratings and open-ended comments to produce a final prototype of the tool. This version was used by the two authors to evaluate a sample of survey studies from three meta-analyses (Hagger et al., 2017; Hoffmann et al., 2017; Young et al., 2014). In the fourth and final stage, we aimed to provide preliminary support for the criterion validity of scores produced by the Q-SSP checklist using a ‘known differences’ approach. We evaluated whether the tool could be used to effectively distinguish between groups of studies known to be of “acceptable” and “questionable” quality. First, a set of 20 candidate studies was identified from the aforementioned meta-analyses based on scores on the bespoke quality assessment tools used in the meta-analyses from which they were drawn. Next, a second expert panel provided independent appraisal of the studies and rated them as “acceptable” and “questionable” in quality based on their expert judgment. Subsequently, a final expert panel used the Q-SSP checklist to assess the quality of the same set of 20 studies. Ratings of the studies using the Q-SSP checklist the final panel were compared to the expert judgment ratings and scores from the previously used tools for the set of studies.

Method

Participants

Participants comprised research-active faculty members and research staff from psychology, behavioral science, and health science faculties, with expertise in the application of psychological methods, survey research, study quality appraisal, and evidence synthesis. Participants were primarily identified through their publications. Specifically, literature searches were conducted with Google Scholar search engine, using the terms “questionnaire”, “survey”, “correlation”, “psychology”, “social” “behavior*“, methodological quality”, “study quality”, “meta-analysis”, “review”, and “evidence synthesis”. Relevant authors of retrieved studies were entered as candidates on our initial list of experts. Relevant expertise of candidates was based on their overall scholarly profile and experience including, but not limited to, number and breadth of research articles in high impact peer-reviewed discipline-relevant journals, citation ratings, rigor of publication design, previous experience with study quality assessment using methodological, study quality, or risk-of-bias tools or checklists. As we aimed to include a diverse panel of experts, and we did not restrict our search to country or language. In order to ensure international coverage, we conducted a further search of psychology/social/health science departments in the countries that did not feature in our search to identify relevant experts, through lists of publications appearing on the departmental websites. Recommendations on the optimal sample size in expert consensus vary. It has been suggested, for example, that 10 to 15 experts are sufficient if there is sufficient homogeneity in the group (e.g., members have similar professional background, education, training, and expertise; Delbecq et al., 1986). Overviews of consensus studies indicate that most studies employ between 15 and 20 experts (Hsu and Sandford, 2007). Based on these guidelines, we aimed for a final sample of at least 20 experts. Given that response rates to online surveys are approximately 33% (Nulty, 2008), we aimed to contact at least 100 eligible University faculty members and researchers with the requisite expertise and diversity in country and discipline coverage.

Procedure

The Q-SSP checklist was developed in four stages. An overview of the

Table 1
Purpose and outcome/findings of each stage of development of the Q-SSP checklist.

Stage	Purpose	Outcome/findings
1	Development of Q-SSP checklist items and scoring scheme by study authors.	Shortlist of 20 candidate items and scoring scheme by study authors.
2	Expert consensus (agreement or disagreement) on (1) inclusion of shortlisted candidate items; (2) importance of candidate items; and (3) appropriateness of scoring system. Expert provision of feedback on any aspect of Q-SSP checklist.	High inter-rater agreement on inclusion of 18/20 of items and importance of 16/20 items. 82% of experts agreed on scoring system.
3	Q-SSP checklist refinement based on (1) goodness-of-fit analysis of experts' responses to the agreement with and importance of items; (2) content analysis of experts' feedback; and (3) quality assessment of survey studies with Q-SSP checklist and inter-rater agreement analysis.	Q-SSP checklist item refinement based on goodness-of-fit, content, and inter-rater agreement analyses.
4	Establishing the capacity of the Q-SSP checklist to distinguish between studies that vary in quality (criterion validity) based on (1) experts' quality assessments; (2) inter-rater agreement analyses; and (3) goodness-of-fit analyses.	Some evidence for the criterion validity of the Q-SSP checklist was obtained.

Note. Q-SSP = Quality of survey studies in psychology.

aims and outcomes of each stage are provided in Table 1.

Stage 1 – Development of Candidate Items. In the first stage of development of the Q-SSP checklist, we identified a set of candidate items for the initial version of the checklist based on previous research and recommendations (Asendorpf et al., 2013; Crowe and Sheppard, 2011; Durant, 1994; Moyer and Finney, 2005; Zeng et al., 2015). First, we searched the literature for existing study quality appraisal tools, and for overviews, systematic reviews, and meta-analyses, evaluating these tools. Second, we studied the content of the items in each tool and identified domains of study quality. Third, we studied reviews that appraised the rigor of existing quality assessment tools and took into account their recommendations for quality appraisal tool development. We also considered general published psychological research and publication standards (Appelbaum et al., 2018), and other recommendations for enhancing quality in psychological research (Asendorpf et al., 2013; Finkel et al., 2017). The procedure was conducted by the two authors who have extensive experience in evidence synthesis and quality assessment in the fields of social psychology, health psychology, and behavioral medicine.

Stage 2 – Refining Item Pool Using Expert Consensus. The expert consensus study adopted a cross-sectional design comprising an online questionnaire administered using the Qualtrics^{TM} online survey platform. The online expert consensus approach has been adopted in previous consensus studies (e.g., Connell et al., 2018), and has several advantages, including efficient, cost-effective means to recruit an appropriate panel of experts; assurance of participant anonymity; promotion of efficient dialogue with participants and means to prompt responses; and streamlined data collection, collation, and analysis (Wright, 2005). We were guided by Waggoner et al.'s (2016) best practice guidelines for online consensus research, which specify that studies should state inclusion criteria, recruit a minimum of 11 experts, provide a predetermined definition of consensus, and conduct comprehensive analysis of consensus data. It should be acknowledged that unanimous agreement in consensus surveys is rare and not expected. Previous research have adopted different criterion values for acceptable agreement in consensus studies (range 51%–80%) (Hasson et al., 2000; Keeney et al., 2006). In the present study we adopted a conservative 80% criterion for acceptable agreement.

In January 2018, eligible participants ($N = 167$) were sent an email invitation to participate in the consensus survey. The invitation provided full information about the study, expectations for participation, and a URL directing them to a welcome page that contained information about the study followed by a consent statement. Participants agreeing with the consent statement were automatically directed to the first page of the survey. Two more email reminders were sent in February 2018. Of the 167 experts contacted, 40 agreed to participate (24% response rate) and 33 completed the whole survey (17% attrition rate). Consequently, we exceeded the recommended number of experts (Waggoner et al., 2016). None of the seven 'non-completers' proceeded beyond the demographic questions, so they were excluded from the analysis. Participant characteristics are presented in Table 2. Data collection was completed by the

Table 2
Expert panel participant characteristics for each stage of the Q-SSP checklist development.

Characteristic	Expert Panel					
	Stage 2		Stage 4a		Stage 4 b	
	n	%	*n*	%	*n*	%
Gender						
Male	19	57.58	9	90.00	1	10.00
Female	13	39.40	1	10.00	9	90.00
Unspecified	1	3.03	0	0.00	0	0.00
Occupation						
Assistant professor	3	9.10	0	0.00	0	0.00
Associate professor/Reader	2	6.10	2	20.00	0	0.00
Full professor	4	12.10	0	0.00	0	0.00
Research professor	1	3.03	0	0.00	0	0.00
Professor	2	6.10	0	0.00	0	0.00
Lecturer	1	3.03	1	10.00	2	20.00
Senior lecturer	1	3.03	2	20.00	0	0.00
Research fellow/associate	1	3.03	4	40.00	2	20.00
Research assistant/PhD student	0	0.00	1	10.00	6	60.00
Unspecified	18	54.50	0	0.00	0	0.00
Region and country of residence						
Europe	19	57.60	6	60.00	5	50.00
Finland	2	6.10	1	10.00	0	0.00
France	0	0.00	1	10.00	0	0.00
Greece	1	3.03	0	0.00	0	0.00
Ireland	2	6.10	0	0.00	0	0.00
Italy	3	9.10	0	0.00	0	0.00
The Netherlands	3	9.10	0	0.00	1	10.00
Spain	0	0.00	0	0.00	1	10.00
Switzerland	1	3.03	0	0.00	0	0.00
UK	7	21.20	4	40.00	3	30.00
North America	3	9.10	0	0.00	3	30.00
Canada	1	3.03	0	0.00	0	0.00
USA	2	6.10	0	0.00	3	30.00
Asia-Pacific	7	21.20	2	20.00	1	10.00
Australia	4	12.10	1	10.00	1	10.00
China and Hong Kong	2	6.10	1	10.00	0	0.00
Singapore	1	3.03	0	0.00	0	0.00
Africa and Middle East	1	3.03	2	20.00	1	10.00
South Africa	1	3.03	0	0.00	0	0.00
Oman	0	0.00	0	0.00	1	10.00
Unspecified	0	0.00	2	20.00	0	0.00
Area of expertise[a]						
Psychology	31	93.9	5	50.00	10	100.00
Health psychology	3	9.10	1	10.00	0	0.00
Sport/exercise psychology	3	9.10	3	30.00	0	0.00
Health science	7	21.20	1	10.00	0	0.00
Social science	2	6.10	0	0.00	0	0.00
Evidence synthesis	6	18.20	1	10.00	0	0.00
Quality appraisal	2	6.10	0	0.00	0	0.00
Unspecified	0	0.00	3	30.00	0	0.00

Note. Q-SSP = Quality of survey studies in psychology.

[a] Participants could check any of the available options for this characteristic, so categories are not mutually exclusive. The pool of experts was unique at each stage.

end of March 2018. Upon finishing the survey, participants received a closing message, thanking them for their participation and encouraging them to contact the authors if they had any further questions or comments. Participants' responses were recorded by the Qualtrics^{TM} software and downloaded into data spreadsheets for analysis. The study was approved by the Research Ethics Committee of the Psychology Department, [Institution name redacted for masked review] (reference number PSY2017-057).

The online consensus survey was divided into five sections. The first section included questions that described participants in terms of age, gender, geographical location, area of expertise, job title, and place of employment. The second section included the candidate set of items each accompanied by scales for participants to rate their agreement for inclusion of the item in the survey and its importance to evaluating study quality. Participants were also prompted to provide further comments and suggestions for modifications, via an accompanying open-ended free-response text box. Specifically, for each item participants were prompted to rate (1) their agreement on whether the item should be included in the checklist on a binary agree-disagree scale; and (2) their evaluation of the importance of the item to study quality on a continuous four-point scale (1 = *not important* and 4 = *very important*). The third section included questions prompting participants to rate their agreement with the proposed scoring system for the tool on a binary agree-disagree scale, and comment on the scoring system via an open-ended free-response box. A guide accompanied the tool items, which included definitions of each item, terms used, and details of the scoring system. The guide was downloadable, available throughout the online questionnaire, and participants received periodic reminders to consult it when responding to the items. In the fourth section, participants were asked to state whether or not they had used the guide during the survey and whether they had found it useful. A final question prompted participants to rate their agreement with the proposed title and acronym of the tool. The initial candidate tool items and the guide are available in the online supplement (see Appendices A and B).

Stage 3 – Refining the Q-SSP Checklist. The initial pool of checklist items and descriptions were refined based on results from the Stage 2 expert consensus study. We considered 80% agreement our minimum criterion for consensus on participants' ratings for inclusion and importance of each item, based on previous recommendations (Hasson et al., 2000; Keeney et al., 2006). Items falling short of the 80% cut-point for consensus were considered candidates for revision or elimination. We computed goodness-of-fit of participants' responses to the agreement and importance scales with our a priori 80% criterion using chi-square analyses. For the purposes of the goodness-of-fit test, participants' importance ratings were dichotomized. Specifically, "important" and "very important" responses were classified as "high importance" category, and "not important" and "somewhat important" responses were classified as "low importance".

We also content analyzed participants' responses to the open-ended questions on the survey for each item. Content analysis provides new knowledge, insights, conceptual models and practical guides to action (Krippendorff, 1980). We followed Elo and Kyngäs' (2008) approach, who describe content analysis as a research method for making replicable and valid inferences from data, through a systematic classification process of coding and identifying patterns or themes. Our approach was *inductive*, i.e., moving from the *specific* (participants' written comments on quality items) to the *general* (creating categories and themes describing participants' expectations and requirements about the tool). The first step of the analysis was an *open coding* procedure, in which entailed multiple readings of participants' responses with extensive notes taken. During open coding, initial categories were generated based on participants' responses that were semantically similar, and by checking the prominence of responses through its co-occurrence. After open coding, the initial categories were grouped under higher-level, broader, *abstract* categories or themes. The final stage involved applying *labels* to the extracted themes. Labels were descriptors capturing the essence of each theme. The first author carried out the content analysis and the second author reviewed the results and offered suggestions for revision and refinement. The content analysis is available online (https://osf.io/xgy69). A further step in the refinement stage involved establishing whether independent raters could produce reliable study quality ratings using the tool. The tool was used to assess the quality of 30 survey studies, extracted from three meta-analyses: Hagger et al. (2017), Hoffmann et al. (2017), and Young et al. (2014). These meta-analyses were chosen because (1) they included psychological survey studies (i.e., the study design that the Q-SSP checklist aims to assess); (2) study quality was assessed in the included studies; and (3) the research included studies representing different psychology fields (e.g., health psychology, social psychology, environmental psychology, traffic/transport psychology, sport psychology and social cognition). Furthermore, these meta-analyses provided their complete methods and procedures in online supplements. The authors of the present study, both with expertise in conducting research syntheses (systematic reviews, meta-analyses) in psychology, assessed all 30 studies independently. Inter-rater reliability was computed to evaluate agreement on each of the tool items using Gwet's (2008) AC_1 coefficient. The AC_1 is an alternative to the kappa statistic, used in situations when the extent of agreement between two raters is high but kappa does not appropriately reflect the extent of the agreement (Cicchetti and Feinstein, 1990; Gwet, 2008). Values equal to or greater than 0.50, 0.70, and 0.80 on the AC_1 coefficient denote moderate, good, and very good levels of agreement, respectively (Gwet, 2008).

Stage 4 – Criterion Validity of Q-SSP Checklist Scores. An important criterion any study quality assessment checklist is that it can be used to effectively and reliably distinguish between studies that vary in quality. We therefore aimed to evaluate the effectiveness of the Q-SSP checklist prototype in distinguishing between "acceptable" and "questionable" studies from a 'criterion set' of studies with established quality scores. However, establishing the quality of a criterion set of studies against which the tool is to be assessed, presented considerable challenges. In order to do this, a first step in this process (Stage 4a) was to randomly select a set of studies with known differences in quality based on two criteria: quality assessments using bespoke quality assessment tools from previous studies and expert consensus. The random selection of studies was done with the use of the online random number generator https://www.random.org/. We selected 20 studies from three previous meta-analyses (Hagger et al., 2017; Hoffmann et al., 2017; Young et al., 2014), 10 that were rated as having good ("acceptable") quality and 10 that were rated as having poor ("questionable") quality based on the assessment tools used in the individual studies. We then asked a panel of judges ($N = 10$), independent from the panels from the previous stages, with expertise in evidence synthesis and/or study quality appraisal to provide an assessment of the quality of each of the 20 studies based on their expert opinion.

The subsequent step (Stage 4 b) aimed to examine whether Q-SSP scores were able to distinguish between studies identified as acceptable and questionable in quality in Stage 4a. Eligible experts ($N = 33$) were identified through their previous publication track record in the field (also refer to our Participants section) and invited to participate in the study by email. Ten agreed to participate (response rate = 33.33%), and all who agreed subsequently completed their assessments (completion rate = 100%). As a guide, judges were provided with a brief narrative identifying the typical expected criteria used to evaluate study quality; a summary of criteria derived from previous quality assessment tools. However, judges were asked to use their own judgment and bring to bear their experience and expertise in making their evaluations. The judges' appraisals were pooled and consistency evaluated using ICC. It was expected that this would provide a set of studies on which there was general consensus from the judges, along with the evaluations from the bespoke study quality assessment tools used in the original meta-analyses from which the set of studies was drawn, on their quality. Then, a final panel of experts ($N = 10$), also independent of the previous panels in previous stages, with similar experience in evidence synthesis and/or quality assessment was then asked to use the Q-SSP checklist provide quality assessment scores for each of the studies. This last panel was also identified through their publications and research track record in the field. Twenty judges were invited, by email, to participate; 10 agreed to participate (response rate = 50%) and all carried out assessments to completion (completion rate = 100%). Table 2 presents participants' characteristics in Stages 4a and 4 b. We then compared overall study quality scores ("acceptable" vs. "questionable") derived from the Q-SSP checklist with the consensus quality judgments of the experts and the assessments from the tools used in the meta-analyses from which the set of studies was drawn using percentage agreement and Gwet's (2008) AC$_1$ coefficient. We also evaluated the goodness-of-fit of the quality score for each study using the Q-SSP checklist with scores from the expert judges. High agreement and close fit for the Q-SSP checklist and expert judgement scores would provide preliminary support for the criterion validity

Table 3
Inter-rater agreement and consensus survey agreement statistics for the Q-SSP checklist.

Item#	Domain and item description[a]	Inter-rater reliability			Consensus								
		Agreement	AC$_1$	*P*	Included			Importance					
					Agreement	χ^2	*p*	*M*	*SD*	*Mdn.*	Agreement	χ^2	*P*
1.	Introduction 1. Were hypotheses or aims explicitly stated?	96.67%	.963	<.001	100.00%	–	–	3.606	0.659	4	93.33%	2.455	.117
2.	Introduction 2. Were operational definitions of predictor (independent) and outcome (dependent) variables provided?	93.33%	.880	<.001	87.87%	1.280	.258	3.273	0.801	3	66.67%	0.030	.862
3.	Introduction 3. Were participant eligibility criteria (inclusion and exclusion) explicitly stated?	86.67%	.781	<.001	100.00%	–	–	3.394	0.747	4	83.33%	0.485	.486
4.	Introduction 4. Were participants recruited using an acceptable recruitment strategy?	93.33%	.876	<.001	87.87%	1.280	.258	3.091	0.843	3	63.3%	0.371	.542
5.	Participants 1. Were participants selected by a random/probability sampling strategy?	90.00%	.817	<.001	75.76%	0.371	.542	2.818	0.983	3	70.00%	3.667	.056
6.	Participants 2. Was the sample size appropriate?	90.00%	.849	<.001	96.97%	5.939	.015	3.576	0.614	4	63.3%	4.008	.045
7.	Participants 3. Were participants randomly assigned into groups/ conditions?	96.67%	.958	<.001	81.82%	0.068	.794	3.121	1.023	3	83.33%	1.091	.296
8.	Data 1. Was the response/ participation/recruitment rate provided?	83.33%	.719	<.001	87.87%	1.280	.258	3.121	0.857	3	83.33%	1.280	.258
9.	Data 2. Was the attrition rate acceptable?	73.33%	.856	<.001	84.84%	0.485	.486	3.091	0.765	3	80.0%	0.068	.794
10.	Data 3. Was the attrition rate treated appropriately in data analyses?	86.67%	.815	.004	87.87%	1.280	.258	3.242	0.708	3	66.67%	0.485	.486
11.	Data 4. Were the chosen statistical tests appropriate to address hypotheses or research questions?	100.00%	1.000	<.001	100.00%	–	–	3.727	0.517	4	93.33%	5.939	.015
12.	Data 5. Did the study include a formative research or pilot phase?	83.33%	.719	<.001	69.70%	2.189	.139	2.303	0.810	2	73.3%	34.008	<.001
13.	Data 6. Were the measures provided in the report (or in a supplement) in full?	80.00%	.723	<.001	84.84%	0.485	.486	2.909	0.980	3	80.0%	1.091	.296
14.	Data 7. Were all measures of established validity, or was a validation procedure undertaken by the authors?	96.67%	.944	<.001	90.91%	2.455	.117	3.212	0.857	3	46.6%	0.030	.862
15.	Data 8. Was the study sample described in terms of key demographic characteristics?	90.00%	.817	<.001	96.97%	5.939	.015	3.303	0.684	3	86.6%	1.280	.258
16.	Data 9. Was the data collection process described with sufficient detail for it to be replicated?	80.00%	.706	<.001	96.97%	5.939	.015	3.424	0.830	4	76.6%	0.485	.486
17.	Data 10. Were generalizations of findings restricted to the population from which the sample was drawn?	90.00%	.805	<.001	78.87%	0.030	.862	2.909	0.914	3	56.6%	3.667	.056
18.	Ethics 1. Was the study approved by a relevant institutional review board or research ethics committee?	100.00%	1.000	<.001	90.91%	2.455	.117	3.364	0.822	4	96.67%	0.030	.862
19.	Ethics 2. Did participants provide informed consent (or assent, where relevant)?	96.67%	.933	<.001	84.84%	0.485	.486	3.061	0.966	3	96.67%	2.189	.139
20.	Ethics 3. Were funding sources or conflicts of interest disclosed?	93.33%	.880	<.001	90.91%	2.455	.117	3.242	0.902	4	83.33%	0.371	.542

Note. Q-SSP = Quality of survey studies in psychology; AC1 = Gwet (2008) AC$_1$ agreement coefficient; χ^2 = Goodness of fit chi-square; *t* = Independent samples *t*-test of difference from scale mid-point.

[a] Items listed in the table are those presented to participants in the expert consensus study prior to revision.

of scores produced by the Q-SSP checklist.

Results

Stage 1 – initial item pool

The review of the literature and existing study quality tools produced the initial list of candidate items for the subsequent development stage of the study using expert consensus. The initial list is available online (https://osf.io/xgy69).

Stage 2 – expert consensus

Participants in the expert panels for the second stage of the Q-SSP checklist development were university faculty and researchers ($N = 33$; age $M = 45.30$, $SD = 8.31$) from fourteen countries. Participant characteristics including region and country of origin, academic rank, areas of expertise, and gender are presented in Table 2. Agreement among raters on the inclusion and importance ratings for each quality assessment item are presented in Table 3, and the data files and analysis scripts are available online (https://osf.io/xgy69). Participants demonstrated very high agreement on inclusion and importance ratings for the majority of the items. Specifically, agreement ratings on whether the item should be included in the checklist was above our 80% criterion for 18 out of the 20 items. Exceptions were items 5 ("Were participants selected by a random/probability sampling strategy?") and 12 ("Did the study include a formative research or pilot phase?"), which had 75.8% and 69.7% agreement, respectively. These agreement proportions fell short of our 80% criterion, although a 70% agreement criterion is often considered acceptable (Hasson et al., 2000; Keeney et al., 2006). Goodness-of-fit chi-square analysis revealed statistically non-significant values for all but three of the items. Specifically, agreement was significantly lower than the 80% criterion for items 5 ("Were participants selected by a random/probability sampling strategy?"; χ^2 (1) = 5.939, $p < .015$), 15 ("Was the study sample described in terms of key demographic characteristics"; χ^2 (1) = 5.939, $p < .015$), and 16 ("Was the data collection process described with sufficient detail for it to be replicated"; χ^2 (1) = 5.939, $p < .015$).

Regarding the agreement ratings for the importance of each study quality item, results indicated that 16 of the 20 items were rated 3 or above on the 4-point scale. Mean scores for items 5 ("Were participants selected by a random/probability sampling strategy?", 12 "Did the study include a formative research or pilot phase?", 13 "Were the measures provided in the report (or in a supplement) in full?", and 7 "Were generalizations of findings restricted to the population from which the sample was drawn?" ranged between 2.82 and 2.91. Chi-square tests indicated that agreement was high, with non-significant chi-square values indicating no difference from the 80% criterion for all but two items: item 6 ("Was the sample size appropriate?"; χ^2 (1) = 4.008, $p < .045$), and 11 ("Were the chosen statistical tests appropriate to address hypotheses or research questions"; χ^2 (1) = 5.939, $p < .015$). The lower rates of agreement for these items suggested that further scrutiny of participants' responses to the open-ended comments for these items was warranted.

Finally, twenty-seven participants (82%) agreed with the proposed scoring system and the majority of participants ($n = 25$, 76%) consulted the guide when making their assessments.

Stage 3 – - Q-SSP checklist refinement based on content-analysis and inter-rater agreement analysis

Content analysis. Four themes emerged from the content analysis of participants' written responses to the open-ended questions: clarity, generalizability, transparency, and scoring flexibility. The themes summarize participants' comments, suggestions, and expectations concerning the content of tool and guide. Details of the content analysis including participants' comments on each item, the co-occurrence of comments, suggestions for improvement, emerging themes, and the steps undertaken to meet participants' expectations are presented online (https://osf.io/xgy69).

Table 4
Final Q-SSP checklist items.

Item #	Domain	Item
1	Introduction	Was the problem or phenomenon under investigation defined, described, and justified?
2	Introduction	Was the population under investigation defined, described, and justified?
3	Introduction	Were specific research questions or hypotheses stated?
4	Introduction	Were operational definitions of all study variables provided?
5	Participants	Were participant inclusion criteria stated?
6	Participants	Was the participant recruitment strategy described?
7	Participants	Was a justification/rationale for the sample size provided?
8	Data	Was the attrition rate provided? (applies to cross-sectional and prospective studies)
9	Data	Was a method of treating attrition provided? (applies to cross-sectional and prospective studies)
10	Data	Were the data analysis techniques justified (i.e., was the link between hypotheses/aims/research questions and data analyses explained)?
11	Data	Were the measures provided in the report (or in a supplement) in full?
12	Data	Was evidence provided for the validity of all the measures (or instrument) used?
13	Data	Was information provided about the person(s) who collected the data (e.g., training, expertise, other demographic characteristics)?
14	Data	Was information provided about the context (e.g., place) of data collection?
15	Data	Was information provided about the duration (or start and end date) of data collection?
16	Data	Was the study sample described in terms of key demographic characteristics?
17	Data	Was discussion of findings confined to the population from which the sample was drawn?
18	Ethics	Were participants asked to provide (informed) consent or assent?
19	Ethics	Were participants debriefed at the end of data collection?
20	Ethics	Were funding sources or conflicts of interest disclosed?

Note. Q-SSP = Quality of survey studies in psychology.

The most prominent theme emerging from the content analysis was the need for *clarity* in the terminology and wording of the quality items and the guide. Participants identified ambiguity and lack of clarity in some of the terms used. In particular, the terms "appropriate", "sufficient", and "acceptable" were flagged as problematic, due to their potential to confuse users of the tool, the provision of relevant definitions in the accompanying guide notwithstanding. The expectation for the tool to be *generalizable* across survey designs (e.g., pencil-and-paper, online, quantitative, qualitative), research questions, and regulations of academic institutions, was a consistent theme. For example, some participants indicated that not all universities have ethics committees or IRBs, and that in some countries survey studies are exempt from committee or IRB approval. It was also prominently indicated that including random assignment and probability sampling as quality criteria would not be relevant to surveys employing other sampling and assignment methods. The imperative of *transparency* in reporting emerged as a theme, with suggestions to rephrase items to prioritize transparency as a study quality criterion. Some participants suggested that published studies in psychology tend not to report crucial information (e.g., attrition rates, a priori sample estimation), and such non-reporting diminishes study quality. It was therefore suggested that identifying whether particular quality criteria are reported "at all" may be a more appropriate for some of the items. The expectation that the tool allows for a degree of *flexibility in scoring* was suggested. For example, it was recommended that the scoring of the quality domains could be non-numerical, and even optional.

Refinement of Checklist Items. Based on the inter-rater agreement analysis and the written responses of the expert panel, items from the initial version of the Q-SSP checklist were revised. The revised items are presented in Table 4.[3] Revisions primarily involved some re-phrasing of checklist items and the guide, with the goal of improving clarity, generalizability, transparency in reporting, and flexibility in scoring. Most revisions were made on the basis of participants' responses to the open-ended comments for each item checked against responses to the inclusion and importance ratings.

In addition, items 5 ("Were participants selected by a random/probability sampling strategy?"), 7 ("Were participants randomly assigned into groups/conditions?"), and 12 ("Did the study include a formative research or pilot phase?") were removed in response to specific feedback provided by participants. Our experts pointed out that random assignment and random sampling, as well as the inclusion of formative research elements, do not often apply to studies adopting survey designs, and the absence of these elements may not necessarily impact study quality.

While items 16 ("Was the data collection process described with sufficient detail for it to be replicated?") and 18 ("Was the study approved by a relevant institutional review board or research ethics committee?") were considered important items, they were substituted for other items. Specifically, item 16 was considered to encompass more than one quality dimension and was therefore replaced with items assessing separate criteria deemed essential to study replication. The criteria were based on recommendations and guidelines from reviews and commentaries on replication (Asendorpf et al., 2013; Norris et al., 2016; Schroter et al., 2012). The item was divided into four separate items: "Was information provided about the person(s) who collected the data (e.g., training, expertise, other demographic characteristics)?"; "Was information provided about the context (e.g., place) of data collection?"; "Was information provided about the duration (or start and end date) of data collection?"; and "Was the participant recruitment strategy described?"

Similarly, item 18 was replaced by two items reflecting ethical conduct: "Were participants asked to provide (informed) consent or assent?" and "Were participants debriefed at the end of data collection?" Participants had alerted us to the fact that ethics committees do not exist in all countries or academic departments, and that survey studies are sometimes exempt from ethical or IRB approval. The substitute items gauge ethical procedures considered essential in human survey research, even in the absence of formal ethical approval, based on published guidelines (Appelbaum et al., 2018). According to these guidelines, informed consent and debriefing procedures are sufficient to cover issues surrounding distress, deception, lack of confidentiality and participant rights.

Finally, we made minor changes based on inter-rater agreement results from the previous stage (see Table 3). Gwet (2008) AC_1 coefficients indicated acceptable agreement ($AC_1 > 0.70$) across the rated studies for all items, with overall agreement levels >80% for all items. Disagreements were resolved through discussion. Without exception, disagreements stemmed from minor variations in the interpretation of quality criteria. Resolution of disagreements resulted in minor revisions to the Q-SSP checklist guide to clarify issues that led to the disagreements. For example, the term 'context' used in the item "Was information provided about the context (e.g., place) of data collection?" was sometimes misinterpreted in studies that collected data via phone and internet. This led to adding text in the guide, further explaining the meaning of data collection "context" and "place". The final version of the QSSP and its accompanying guide are presented in Appendix A (supplemental materials).

Scoring System Development. Quality items in the QSSP are scored with the options: "yes", or "no", "not stated clearly", or "not applicable", based on the information provided in the research report (e.g., article, poster, protocol, thesis) and supplemental material, if available. Quality appraisal and scoring is expected to be based solely on information provided in the published study and any accompanying supplemental materials, instead of raters' interpretation of study elements that may be absent or missing from the report. The quality criteria are grouped into four domains: introduction (study rationale and variables), participants (sampling and recruitment), data (data collection, analyses, results and discussion), and ethics (consent, debrief, and funding/conflicts of interest). The domains represent groups of items designed to assess conceptually-similar aspects of study quality. For example, items gauging procedures of consent, assent, and debriefing, as well as the disclosure of funding sources and conflicts of interest, are grouped into the ethics domain, and items gauging participant inclusion criteria, recruitment strategies and sample size rationale, are grouped into the participants domain. These domains are colour coded on the scoring sheet to facilitate scoring.

An overall quality numerical score is a percentage calculated by dividing the "yes" answers to the quality items by the total number of applicable items. Based on this scoring system, studies are categorized as having "questionable" quality if they do not receive "yes" responses for five or more checklist items, otherwise studies are classified as having "acceptable" quality. Depending on the number of applicable items, a study should receive a "yes" response to between 70% and 75% of items to receive an overall "acceptable" quality score. This criterion corresponds well with recommended cut-offs offered by other general study quality assessment tools (e.g., Glynn, 2006; Husebø et al., 2012). However, it must be stressed that such cut-off values are arbitrary, and other less stringent cut-off values have been proposed. Cut-off values should also be viewed in light of the concerns regarding the use of overall quality scores rather than domain or individual item scores.

Domain-specific scores are simple ratios calculated by dividing "yes" scores by the number of applicable items in each domain. Non-applicable choices are shaded on the scoring sheet to ensure that only appropriate options are selected during scoring. As some items are considered essential to all studies, the "not stated clearly" or "not applicable" options are not considered appropriate (e.g., "Was the problem or phenomenon under investigation defined, described, and justified?" and "Was the population under investigation defined, described, and justified?"). Numerical scoring is at the discretion of users of the Q-SSP checklist. The Q-SSP checklist comes with a guide providing definitions and examples of the terms used in the checklist, and guidance on scoring (see checklist in Appendix A).

Stage 4 – criterion validity of Q-SSP checklist scores

The final stage examined the effectiveness of the Q-SSP checklist in distinguishing between studies of known difference in quality. Ten experts (age $M = 33.70$, $SD = 4.19$) decided whether a set of 10 studies with known differences in quality were of acceptable or questionable quality, based on their extant knowledge and experience (participant characteristics are presented in Table 2). Quality assessment ratings for each study based on the published ratings from the source meta-analysis and ratings of each panel member based on their expertise and the data and analysis scripts are available online (https://osf.io/xgy69). Averaged inter-rater agreement for each study across the experts was good ($ICC = 0.75$, $p < .001$), and final consensus-based ratings are also available online (https://osf.io/xgy69). Overall, eleven studies were judged to be of 'questionable quality' by a majority of the experts (≥60% agreement), while only four studies were judged to be of 'acceptable' quality based on the same criterion. The judges were split on their evaluation of the remaining five studies.

Next, a different panel of experts ($N = 10$; age $M = 33.33$; $SD = 6.04$) used the Q-SSP checklist to assess the quality of each of the final studies

[3] Original and revised versions of the checklist items and guide are provided online (https://osf.io/xgy69). The finalized version of the checklist and guide is also provided in Appendix A (supplemental materials).

(participant characteristics are presented in Table 2).[4] Final ratings for each study and for each panel member are presented online (https://osf.io/xgy69). Inter-rater reliability for the overall classification of the studies based on the Q-SSP checklist items was good (ICC = 0.77). We computed agreement in overall study quality scores ("acceptable" vs. "questionable") across all studies scores computed by the Q-SSP checklist with quality scores derived from expert judgments, and quality scores derived from the other quality assessment tools used in the meta-analyses from which the studies were drawn. Based on a 60% cut-off value for an "acceptable" study, results revealed a moderate level of agreement (75% agreement; Gwet $AC_1 = 0.501$, $p = .018$) between the Q-SSP checklist ratings and the ratings other quality assessment tools used in the meta-analyses from which the studies were drawn. There was slightly lower agreement (65% agreement; Gwet $AC_1 = 0.302$, $p = .174$) for the comparison of the Q-SSP checklist ratings and expert's quality judgments. Although we report results for a 60% cut-off value for an "acceptable" study, it is important to note that we conducted our test of the criterion validity of Q-SSP checklist scores tests using a range of cut-off values. Since the proposed scoring system for the Q-SSP checklist recommends a cut-off value of 75%, we also produced agreement scores for a 75% cut-off value, as well as 65% and 70% cut-off values. Results are available in the supplemental materials: https://osf.io/xgy69. Results of the analyses for the more stringent cut-off values revealed lower agreement than the use of the 60% value.

Finally, we tested the goodness-of-fit of the proportion of "acceptable" or "questionable" scores produced using the Q-SSP checklist with the scores produced by the experts for each study. - Results demonstrated statistically significant differences, or an indeterminant (infinite) chi-square value,[5] in the proportion of "acceptable" and "questionable" ratings for a majority (12/20) of the studies (Table 5). These data suggest a lack of congruence in ratings for more than half of the studies. However, it is important to note that for three of the analyses with indeterminant chi-square values there was a clear trend indicating agreement. For example, Study 20 was rated of "acceptable" quality by all ten judges based on prior knowledge and expertise, while seven raters using the Q-SSP checklist rated the same study "acceptable" and only three rated it "questionable". Therefore, while the difference was significant, these data indicate a clear trend toward agreement for this study. Taking these observations into account, an overall trend toward agreement on quality ratings for most of the studies was evident. However, given the variability in ratings, current results provide only limited evidence for the criterion validity of Q-SSP checklist ratings.

Discussion

The present research describes the development of the quality of survey studies in psychology (Q-SSP) checklist. The checklist comprises 20 items specifically designed to evaluate the quality of survey studies in four domains: introduction (study rationale and variables), participants (sampling and recruitment), data (data collection, analyses, results and discussion), and ethical review (consent, debrief, and funding/conflicts of interest). The checklist was developed using a systematic and rigorous four-stage process based on existing study quality guidelines (e.g., APA, 2010; Asendorpf et al., 2013; Finkel et al., 2017), and recommendations from reviews of previous study quality tools (e.g., Crowe and Sheppard, 2011; Katrak et al., 2004; Khan et al., 2011; Moyer and Finney, 2005). Stage 1 of the development procedure involved an initial search and review of the literature on extant quality appraisal tools in the social sciences. Stage 2 was an online expert consensus study in which experts provided appraisal of the candidate quality items, scoring system, and guide. In Stage 3, a prototype of the tool was used in a pilot test to assess the quality of a pool of survey studies by the authors. The assessments were used to compute inter-rater reliability scores for the checklist items. The Q-SSP checklist was revised and updated based on the results of the expert consensus survey and the pilot-test. Finally, in Stage 4, we conducted a preliminary test of the criterion validity of Q-SSP checklist scores on a selection of survey studies from previous meta-analyses. Specifically, we compared quality scores from the Q-SSP checklist with scores derived from other quality assessment tools and subjective quality judgments from a group of experts for the selection of studies. Results were indicated trends in agreement across the ratings from the different sources, but did not overall provide strong evidence for the criterion validity of scores produced on the Q-SSP checklist.

The development of the current study quality assessment tool is timely given issues raised on the replicability of research findings in psychology. A key factor that may hinder replicability is lack of precision and transparency in the reporting of study methods and findings. These reporting issues impede scientific progress as they hinder replicability, and present problems for the synthesis and comparison of findings across studies. Consequently, researchers have advocated the assessment of study quality, and the need for comprehensive, transparent reporting of findings, as means to resolve the endemic problems with reporting and transparency (Asendorpf et al., 2013). By providing a discipline- and method-specific means to assess study quality, the Q-SSP checklist, therefore, forms part of the solution to move the field toward greater transparency in research reports and better methodological quality.

A key issue arising from the development of the current tool relates to scoring, specifically whether researchers should use summary scores for study quality based on scores for each item, or to treat each item as a 'standalone' measure of a specific quality criterion. Items from the Q-SSP checklist are designed to function as standalone criteria, but we also provide a scoring system to enable researchers to provide overall and domain-specific study quality scores. There is no clear consensus in the literature on study quality on the use of summary or overall scores. A number of previously-developed instruments have advocated the use of overall scores, and have provided guidelines on how to compute the scores (Glynn, 2006). However, others have advocated the use of standalone quality criteria, particularly when assessing study quality in systematic reviews and meta-analyses (Johnson et al., 2014). This is based on the premise that summary scores give equal weighting to each quality item, when such weighting may not be justifiable – some criteria may be considered more important, and impact study quality more than others. One alternative is to weight scores for each quality item by its importance when computing summary scores. However, there is also no clear consensus on the relative importance of the separate quality items in study quality checklists, which presents difficulties in assigning weights. The alternative is to use 'standalone' items to evaluate studies in a systematic review or meta-analysis, in which the effect of each quality criterion from the checklist on the outcomes of interest in the studies assessed. However, this approach can also be problematic due to the sheer number of analyses required to assess the separate effect of each of the 20 study criteria separately.

There has also been debate in the broader quality appraisal literature on whether to compute overall or 'summary' quality scores. Some tools do not advocate the use of summary scores at all, and instead focus on 'vote counts' of "yes" or "no" responses on checklist items (e.g., Jarde et al., 2013). This approach is advocated because decisions on whether studies are of acceptable (or high) and questionable (or low) quality are usually based on arbitrary cut-off values of tool summary scores. While the cut-off points using in the Q-SSP checklist are based on those adopted

[4] Half of the participants in this sample of experts were PhD students, albeit those with considerable experience and training in the evaluation of study quality using checklists. However, given their overall experience is likely to be less extensive than researchers and tenured academic panel members, we conducted an analysis to check whether evaluations differed according to participating member status (PhD student vs, other researcher). Chi-square analyses demonstrated that evaluation of studies as 'acceptable' or 'questionable' did not differ significantly according to status ($ps > .287$).

[5] Indeterminant chi-square values were attributed to the presence of zero values in one of the analysis cells but was, nonetheless, suggestive of a difference.

Table 5
Study quality ratings for the sample of studies ($N = 10$) from stage 4 based on other instruments, expert judgments, and the Q-SSP checklist ratings.

Study	Method of quality assessment[a]							Goodness-of-fit[c]	
	Other instruments	Expert judgments			Q-SSP Checklist ratings (60%) cutoff			χ^2	p
		"Acceptable"	"Questionable"	Overall quality[b]	"Acceptable"	"Questionable"	Overall quality[b]		
1	2	7	3	1	3	7	2	7.619	.006
2	2	1	9	2	5	5	–	6.400	.011
3	1	5	5	–	7	3	1	1.600	.206
4	1	7	3	1	6	4	1	0.476	.490
5	1	8	2	1	7	3	1	0.625	.429
6	2	5	5	–	0	10	2	Inf.	<.001
7	1	4	6	2	0	10	2	Inf.	<.001
8	1	5	5	–	6	4	1	0.417	.519
9	2	8	2	1	6	4	1	2.500	.114
10	2	7	3	1	0	10	2	Inf.	<.001
11	1	9	1	1	7	3	1	4.444	.035
12	1	3	7	2	0	10	2	Inf.	<.001
13	1	6	4	1	3	7	2	3.750	.053
14	1	8	2	1	5	5	–	3.600	.058
15	2	7	3	1	5	5	–	1.905	.168
16	2	3	7	2	1	9	2	4.444	.035
17	1	10	0	1	2	8	2	Inf.	<.001
18	1	10	0	1	6	4	1	Inf.	<.001
19	2	4	6	2	0	10	2	Inf.	<.001
20	1	10	0	1	7	3	1	Inf.	<.001

Note. Q-SSP = Quality of survey studies in psychology.
χ^2 = Goodness of fit chi-square.

[a] 1 = Study judged to be of "acceptable" quality; 2 = Study judged to be of 'questionable' quality.

[b] Overall study quality based on majority of study quality scores.

[c] Goodness-of-fit chi-square analysis of overall quality scores for Q-SSP checklist with expected values set at overall quality scores from experts' judgments.

by previously published quality appraisal tools (e.g., Glynn, 2006), the decision on whether to use summary scores and cut-off values for quality assessment should be guided by the researcher's aims. For example, if the aim of quality assessment is sensitivity or subgroup analyses in the context of a meta-analysis, then summary scores and cut-off values may be appropriate. However, if the aim is to describe the overall quality of a body of evidence in a narrative review, or if achieving acceptable quality on individual or sets of quality items is a criterion for inclusion (e.g., Higgins and Green, 2008), quality may be ascertained using a 'vote count' procedure or general description based on the profile of "yes" and "no" responses to items.

A related issue is whether summary scores should be calculated in separate domains. Criticism of the use of domain-specific scores is similar to those levelled at the use of overall summary scores, that is, summary scores assume equal weight to each criterion included when they should be considered as standalone. However, others (e.g., Moyer and Finney, 2005) argue for domain-specific scores comprising groups of quality components based on conceptual similarity. For example, in the Q-SSP checklist, the 'data' domain comprises the 'measures', 'collection', 'analyses', 'results' and 'discussion' components, which are grouped together because they all address dimensions relating to 'data'. However, given calls for 'flexibility' in scoring expressed by the expert raters in our consensus study, decisions on scoring, including computing overall and domain-specific summary scores, are left to the discretion of the checklist user based on the purpose of their study and quality assessment needs. The tool categorizes studies as questionable if they fail to meet five or more quality items, out of the twenty. However, not all quality items can be applicable to all studies. Specifically, two items relating to ethics procedures ("were participants asked to provide (informed) consent or assent" and "were participants debriefed at the end of data collection), and one item relating to data analyses ("was a method of treating attrition provided"), may be excluded from some studies. Therefore, the overall cut-off score used by the Q-SSP checklist to categorize studies as having acceptable or questionable quality, ranges between 70% and 75%, depending on the number of applicable quality items considered. The cut-off values follow common practices in the field (Glynn, 2006; Hagger et al., 2017; Protogerou et al., 2018), but we note that these practices should be treated as rules-of-thumb, given that no single cut-off value is universally accepted in the literature (Holly et al., 2017).

A related issue is the extent to which researchers should incorporate their own discretion in interpreting the claims of study authors when it comes to judging study quality criteria. The Q-SSP checklist has been developed with the goal of standardizing responses such that quality assessment is uniform across researchers. However, there are instances where a Q-SSP user may have decided that a quality criterion has been fulfilled in a suboptimal way, or deviates from the expected quality requirements. In such instances, Q-SSP users are encouraged to make additional notes on the quality of the arguments provided by the authors of study reports for particular study criteria, or on the ways quality criteria are fulfilled. Such notes can be used to obtain nuanced, in-depth understanding of quality aspects in a particular study or a group of studies. More broadly, we would recommend, for example, conducting a content or thematic analysis on the ways quality criteria were fulfilled in groups of conceptually-related studies, identifying overarching themes on the strong and weak quality domains. Consequently, the Q-SSP checklist may be used to guide qualitative assessments of study quality based on the criteria specified in the checklist and on raters' notes on the ways its items were fulfilled.

It is also important to note that evaluation of study quality using the Q-SSP checklist is heavily dependent on reporting quality of the study. Numerous researchers have emphasized the link between reporting quality and study quality (Asendorpf et al., 2013; Buccheri and Sharifi, 2017; Higgins and Green, 2008). Although transparent and comprehensive reporting alone does not equate to high quality in terms of study conduct, appropriate evaluation of study quality is only possible in the context of transparent and appropriate reporting. Furthermore, lax reporting (e.g., lack of transparency, omission of critical methodological details, reporting only the bare minimum of details) tends to be associated with poorer study quality and researchers across multiple disciplines tend to agree that non-transparent reporting is strongly associated with biased findings (Buccheri and Sharifi, 2017; Mullins et al., 2014). Researchers using checklists like the Q-SSP to evaluate study quality almost always have only the study write-up itself and any supplemental materials as the basis on which to make judgements on quality, coupled with their own interpretation of the reporting. In cases where a study write-up does not provide sufficient information to provide a clear judgement on whether a particular quality criterion has been fulfilled, the Q-SSP checklist default position is to adjudge that criterion as unfulfilled. In some cases, this may lead to a 'false negative' as that quality criterion may, in reality, have been fulfilled, but given the interrelatedness between reporting quality and study quality, the adoption of this position is not unfounded and, in the absence of other information, the researcher has no alternative. Finally, given the study write-up and any available supplements is the only information available, the researcher also has no leeway afford the study any benefit of the doubt when making a determination on the study quality.

Another key issue that arose in our study related to specific terminology used in the preliminary candidate items of the Q-SSP checklist. Specifically, we had initially adopted evaluative terms, such as "appropriate", "sufficient", "acceptable", "adequate", and "clear" in checklist items, following wording norms of published quality appraisal tools (e.g., Durant, 1994; Glynn, 2006; Jadad et al., 1996), experts in our consensus survey identified the subjectivity of these terms, with the likelihood that their use may lead to ambiguity and uncertainty when users interpreting the items. We removed all instances of these adjectives from the revised version. We argue against incorporating evaluative terms in future quality appraisal tools and advise caution when applying tools that incorporate such terms, as the resulting appraisals may lead to ambiguity in interpretation which may affect consistency of responding and the reliability and validity of the scores produced.

The issue of criterion validity requires future attention. We aimed to provide initial support for criterion validity of Q-SSP checklist scores by testing whether researchers could use the scores on the tool to distinguish between sets of studies with 'known differences' in quality. Our analysis showed some trends toward agreement based on ratings from the bespoke quality assessment tools used in the meta-analyses from which the studies were drawn and the ratings of experts based on their subjective judgment of the study quality. However, there was considerable variability in the ratings such that current data provided only limited support for the criterion validity of Q-SSP checklist scores. Further research with larger numbers of participants and studies is required to confirm the validity of the Q-SSP checklist scores as a means to distinguish between studies of known difference in quality.

Findings on the criterion validity of Q-SSP checklist scores should also be viewed in light of the inherent problems in identifying clear criteria against which to judge the validity of study quality assessments. Our original purpose was to identify a set of studies with known differences in study quality against which to evaluate the Q-SSP checklist, and we reasoned that a converging evidence approach would be fit-for-purpose to do so, that is, using previous assessments from other instruments and expert judgments to arrive at the set of criterion studies. However, it is clear from our current analysis that even this step presented considerable challenges; there was also substantive variability in the quality scores provided by the experts and those from previous tools. This step in the procedure was, therefore, subject to the same problems inherent in study quality assessment that we had aimed to resolve by developing the Q-SSP checklist in the first place. For example, the quality scores from the bespoke set of quality assessment tools in the meta-analyses from which the sample of criterion studies was drawn were likely subject to multiple biases including variability in the tool criteria, scoring systems, and purpose. Similarly, although expert consensus may be a potentially valid means to produce a set of studies with 'known differences' in quality, the experts' judgments were not based on a specific set of criteria. Specifically, experts' based their judgments on their experience rather than specific criteria, so it was possible that their judgments were based on a narrow set of criteria or idiosyncratic features of the study rather than across a comprehensive set of criteria. The lack of clear, valid means to assess the quality of the selected studies highlights the difficulty in identifying an appropriate set of studies with 'known differences' in quality to be used in a criterion validity test of scores from the Q-SSP checklist. Current evidence for the criterion validity of Q-SSP checklist scores based on the current study should, therefore, be interpreted in light of these limitations, and more adequate tests of the criterion validity of scores produced by the Q-SSP checklist are required going forward.

Applications of the Q-SSP checklist

The value of the Q-SSP checklist is that it provides researchers, referees, editors, science writers, and stakeholders with a means to assess the quality of survey studies in psychology. The checklist has a number of important applications going forward. A primary purpose is for researchers conducting evidence syntheses. The tool can be used by researchers to assess the quality of studies as an inclusion criterion in systematic reviews and meta-analyses, and to test the effects of study quality on effects in these analyses using sub-group and sensitivity analyses. The tool may also serve to guide the planning and the writing-up of the research by providing guidance on important quality criteria. In addition, the tool may be used by professionals (e.g., clinicians, physicians, practitioners) wishing to evaluate the quality of psychological evidence that may inform their practice. Stakeholders in organizations in the fields of healthcare and education can use the tool to evaluate the quality of evidence based on psychological surveys used to inform policy on evidence-based practice, either in conjunction with a formal evidence synthesis or in groups of studies collected on a particular topic. It may also be useful for educators to illustrate issues relating to study quality in research methods courses.

Strengths, limitations, and recommendations for further research

Our approach to developing the Q-SSP checklist has a number of strengths including the development of an initial pool of quality items based on a review of the literature of previous quality assessment tools, the use of a rigorous consensus approach using panels of experts with appropriate backgrounds in evidence synthesis and study quality to further develop item inclusion, content and descriptions, pilot testing the tool, and computing inter-rater agreement on a random set of survey-based studies from previous meta-analyses. The approach has provided evidence to support the face and content validity of the checklist. The high degree of consensus among experts on the inclusion and importance of each quality domain of the checklist with supporting comments and positive feedback provided converging evidence for the face and content validity of the Q-SSP items. In addition, the tool also demonstrated good inter-rater reliability in a pilot test of its application.

However, a number of limitations should be acknowledged. First, the response rate to the expert consensus survey was relatively low. Low response rates have commonly been reported in expert consensus research, particularly research using online methods, and has been attributed to time constraints of experts and difficulty in reaching participants using email alone (e.g., Burnett et al., 2005; Hutchings et al., 2006). While we were able to recruit an appropriately-sized sample of experts, our findings should be interpreted in light of the potential for systematic bias due to the low response rates. Another limitation of the Q-SSP checklist, which can also be levelled at all quality appraisal tools, is that precision of the assessment is highly dependent on the available information on each quality criterion available in the study report. While trends toward 'open science' have led to increased use of repositories and online journal supplements by researchers to provide comprehensive detail on the design and conduct of their research (Hagger, 2019; Nosek et al., 2015), the use of such supplements is a relatively recent development and not universally available. Users of the Q-SSP checklist are, therefore, only able to base their appraisal on what is reported in research reports and supplements. Increased awareness of the importance of quality appraisal, and greater frequency of use of tools like the Q-SSP checklist to evaluate study quality, will raise the profile of reporting standards and drive greater precision in the reporting of survey study methods. Over time, this will enable more accurate and comprehensive assessment of study quality.

An additional limitation is the inherent problems in identifying studies of known quality, which present considerable challenges to providing strong support for the criterion validity of scores from quality assessment tools like the Q-SSP checklist. One of the key limitations of the current study used to assess criterion validity was the small number of studies and sample of experts, which likely contributed to the high variability. A possible solution would be to conduct a scaled-up version of the criterion validity study. Specifically, a large-scale trial is needed in which the checklist is used to evaluate a large sample of survey studies in psychology with known differences in study quality, verified through the consensus from a large number of experts. Such as study might involve hundreds of experts and studies. A costly but worthy endeavour to develop a tool that produces scores with adequate criterion validity. Such a study would enable researchers to test the sensitivity (the ability to correctly identify studies of “acceptable” quality) and specificity (the ability to correctly identity studies that are not of “acceptable” quality) of the tool, and examine the trade-off between specificity (false positives) and sensitivity (false negatives) by plotting a receiver operating characteristic (ROC) curve (Streiner and Cairney, 2007). Such a study will provide definitive evidence supporting the criterion validity of scores from the Q-SSP checklist.

Finally, we acknowledge that the same sets of quality criteria identified in the Q-SSP checklist may also be appropriate to judge the quality of studies using the same, or similar, methods in other social or behavioral science disciplines (e.g., sociology, social work, cognitive sciences, education, communication science). However, without the same set of rigorous development and expert-consensus rounds adopted in the current study, such a cross-application would be speculative and contra-indicated. However, the current study may form a template to inform the development of study quality tools in other disciplines that share similar methodological approaches to psychology, just as existing tools formed the initial basis for the development of the Q-SSP. But we would argue that the same development stages and expert consensus rounds would be necessary to ensure that the tool content was fit-for-purpose for evaluating study quality that discipline.

Conclusion

The present study describes the development of a checklist to assess the quality of studies adopting survey designs in psychology. We adopted a rigorous expert-consensus approach. We initially developed a candidate set of checklist items, based on currently available quality assessment tools, published recommendations on developing such tools and discipline-specific survey research guidelines. The candidate items were evaluated using an expert consensus study that informed the development of the final set of checklist items and descriptions. We also pilot-tested the checklist to evaluate study quality in a set of survey studies and conducted inter-rater reliability checks. The final 20-item Q-SSP checklist has good agreement among experts as a means to appraise the quality of survey studies in psychology. A preliminary test of the criterion validity of Q-SSP checklist scores provided some indicative trends toward agreement across studies varying in quality verified by experts but, overall, the support for the criterion validity of checklist scores was limited. We advocate further application of the tool to evaluate sets of studies with known differences in quality and using multiple raters to provide additional evidence for the criterion validity of scores produced by the tool.

Declaration of Competing Interest

The authors declare no conflicts of interest.

CRediT authorship contribution statement

Cleo Protogerou: Conceptualization, methodology development, data collection and management, data curation and analysis, and article draft, review, and editing. **Martin S. Hagger**: Conceptualization, methodology development, data curation and analysis, and article draft, review, and editing.

Funding

Martin Hagger’s contribution was supported by a Finland Distinguished Professor (FiDiPro) award (1801/31/2105) from Business Finland.

Acknowledgements

We thank the following researchers for their contribution to the expert consensus stages of the study: Mark Allen, Kwaku Oppong Asante, Matthias Aulbach, Derwin K. C. Chan, Maria Chassandra, Guillaume Chevance, Richard Cooke, Hein de Vries, Oliver Desrichard, Martina deWitte, Lucy Finkelstein-Fox, Emma Gibson, Laura Girelli, Kyra Hamilton, Christin Hoffmann, Jennifer Howell, Blair T. Johnson, Navin Kaushal, Chris Keyworth, George Koulierakis, Taru Lintunen, Fabio Lucidi, Luca Mallia, Kaylyn McAnally, Desmond McEwan, Lauren McKinley, Anne Moyer, Suzette Moyers, Nikos Ntoumanis, Joanne Parsons, Gerlinde Pilkington, Francis Quinn., Panagiotis Rentzelas, Ryan Rhodes, Ainize Sarrionandia, Cecilie Thogersen-Ntoumani, Cathy Ward, Thomas Webb, Ellie Whittaker, Taylor Willmott, Chun-Qing Zhang, and Guangyu Zhou.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.metip.2020.100031.

References

Addington, D.E., McKenzie, E., Norman, R., Wang, J., Bond, G.R., 2013. Essential evidence-based components of first-episode psychosis services. Psychiatr. Serv. 64, 452–457. https://doi.org/10.1176/appi.ps.201200156.

Alderson, P., Green, S., Higgins, J.P.T., 2003. Cochrane Reviewers' Handbook. http://www.cochrane.org/resources/handbook/hbook.htm. (Accessed 1 October 2018).

APA, 2010. American Psychological Association Publication Manual, sixth ed. American Psychological Society, Washington, DC.

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

Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.J.A., Fiedler, K., Wicherts, J.M., 2013. Recommendations for increasing replicability in psychology. Eur. J. Pers. 27, 108–119. https://doi.org/10.1002/per.1919.

Buccheri, R.K., Sharifi, C., 2017. Critical appraisal tools and reporting guidelines for evidence-based practice. Worldviews Evidence-Based Nurs. 14, 463–472. https://doi.org/10.1111/wvn.12258.

Burnett, J., Grimmer, K., Kumar, S., 2005. Development of a generic critical appraisal tool by consensus: presentation of first round Delphi survey results. Internet J. Allied Health Sci. Pract. 3, 7.

Check, J., Schutt, R.K., 2012. Survey research. In: Check, J., Schutt, R.K. (Eds.), Research Methods in Education. Sage, Thousand Oaks, CA, pp. 159–185.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 551–558. https://doi.org/10.1016/0895-4356(90)90159-M.

Ciliska, D.T.H., Buffet, C., 2008. An Introduction to Evidence-Informed Public Health and a Compendium of Critical Appraisal Tools for Public Health Practice. http://www.nccmt.ca/pubs/2008_07_IntroEIPH_compendiumENG.pdf. (Accessed 1 November 2018).

Connell, L.E., Carey, R.N., de Bruin, M., Rothman, A.J., Johnston, M., Kelly, M.P., Michie, S., 2018. Links between behavior change techniques and mechanisms of action: an expert consensus study. Ann. Behav. Med. 53, 708–720. https://doi.org/10.1093/abm/kay082.

Crowe, M., Sheppard, L., 2011. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. J. Clin. Epidemiol. 64, 79–89. https://doi.org/10.1016/j.jclinepi.2010.02.008.

Deeks, J.J., Dinnes, J., D’amico, R., Sowden, A., Sakarovitch, C., Song, F., Altman, D., 2003. Evaluating non-randomised intervention studies. Health Technol. Assess. 7, 1–179. https://doi.org/10.3310/hta7270.

Delbecq, A.L., Van de Ven, A.H., Gustafson, D.H., 1986. Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes. Green Briar Press, Middleton, WI.

Durant, R.H., 1994. Checklist for the evaluation of research articles. J. Adolesc. Health 15, 4–8. https://doi.org/10.1016/1054-139X(94)90381-6.

Elo, S., Kyngäs, H., 2008. The qualitative content analysis process. J. Adv. Nurs. 62, 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x.

Faragher, E.B., Cass, M., Cooper, C.L., 2005. The relationship between job satisfaction and health: a meta-analysis. Occup. Environ. Med. 62, 105–112. https://doi.org/10.1136/oem.2002.006734.

Fink, A., Kosecoff, J., Chassin, M., Brook, R.H., 1984. Consensus methods: characteristics and guidelines for use. Am. J. Publ. Health 74, 979–983. https://doi.org/10.2105/AJPH.74.9.979.

Fink, A., Kosecoff, J., Chassin, M., Brook, R.H., 1991. Consensus Methods: Characteristics and Guidelines for Use. RAND, Santa Monica, CA.

Finkel, E.J., Eastwick, P.W., Reis, H.T., 2017. Replicability and other features of a high-quality science: toward a balanced and empirical approach. J. Pers. Soc. Psychol. 113, 244–253. https://doi.org/10.1037/pspi0000075.

Glynn, L., 2006. A critical appraisal tool for library and information research. Libr. Hi Technol. 24, 387–399. https://doi.org/10.1108/07378830610692154.

Greenhalgh, J., Brown, T., 2017. Quality assessment: where do I begin? In: Boland, A., Cherry, M.G., Dickson, R. (Eds.), Doing a Systematic Review: A Student's Guide. Sage, London, UK, pp. 61–83.

Greenhalgh, T., 2014. How to Read a Paper: the Basics of Evidence-Based Medicine. Wiley, London, UK.

Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61, 29–48. https://doi.org/10.1348/000711006X126600.

Hagger, M.S., 2019. Embracing open science and transparency in health psychology. Health Psychol. Rev. 13, 131–136. https://doi.org/10.1080/17437199.2019.1605614.

Hagger, M.S., Koch, S., Chatzisarantis, N.L.D., Orbell, S., 2017. The common-sense model of self-regulation: meta-analysis and test of a process model. Psychol. Bull. 143, 1117–1154. https://doi.org/10.1037/bul0000118.

Hasson, F., Keeney, S., McKenna, H., 2000. Research guidelines for the Delphi survey technique. J. Adv. Nurs. 32, 1008–1015. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x.

Herdman, M., Rajmil, L., Ravens-Sieberer, U., Bullinger, M., Power, M., Alonso, J., 2002. Expert consensus in the development of a European health-related quality of life measure for children and adolescents: a Delphi study. Acta Paediatr. 91, 1385–1390. https://doi.org/10.1111/j.1651-2227.2002.tb02838.x.

Higgins, J.P.T., Altman, D.G., 2008. Assessing risk of bias in included studies. In: Higgins, J.P.T., Green, S. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions. Wiley, Chichester, UK.

Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., Sterne, J.A.C., 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 343, d5928. https://doi.org/10.1136/bmj.d5928.

Higgins, J.P.T., Green, S., 2008. Cochrane Handbook for Systematic Reviews of Interventions. Wiley, Chichester, UK.

Hoffmann, C., Abraham, C., White, M.P., Ball, S., Skippon, S.M., 2017. What cognitive mechanisms predict travel mode choice? A systematic review with meta-analysis. Transport Rev. 37, 631–652. https://doi.org/10.1080/01441647.2017.1285819.

Holly, C., Salmond, S., Saimbert, M., 2017. Comprehensive Systematic Review for Advanced Practice Nursing, second ed. Springer, New York, NY.

Hsu, C.C., Sandford, B.A., 2007. The Delphi technique: making sense of consensus. Practical Assess. Res. Eval. 12, 108.

Hunter, J.E., Schmidt, F.L., 2004. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, second ed. Sage, Newbury Park, CA.

Husebø, A.M.L., Dyrstad, S.M., Søreide, J.A., Bru, E., 2012. Predicting exercise adherence in cancer patients and survivors: a systematic review and meta-analysis of motivational and behavioural factors. J. Clin. Nurs. 22, 4–21. https://doi.org/10.1111/j.1365-2702.2012.04322.x.

Hutchings, A., Raine, R., Sanderson, C., Black, N., 2006. A comparison of formal consensus methods used for developing clinical guidelines. J. Health Serv. Res. Pol. 11, 218–224. https://doi.org/10.1258/135581906778476553.

Jadad, A.R., Moore, R.A., Carroll, D., Jenkinson, C., Reynolds, D.J.M., Gavaghan, D.J., McQuay, H.J., 1996. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Contr. Clin. Trials 17, 1–12. https://doi.org/10.1016/01972456(95)00134-4.

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. Int. J. Clin. Health Psychol. 13, 138–146. https://doi.org/10.1016/S1697-2600(13)70017-6.

Johnson, B.T., Low, R.E., MacDonald, H.V., 2014. Panning for the gold in health research: incorporating studies' methodological quality in meta-analysis. Psychol. Health 30, 135–152. https://doi.org/10.1080/08870446.2014.953533.

Jones, J., Hunter, D., 1995. Qualitative Research: consensus methods for medical and health services research. BMJ 311, 376–380. https://doi.org/10.1136/bmj.311.7001.376.

Jorm, A.F., 2015. Using the Delphi expert consensus method in mental health research. Aust. N. Z. J. Psychiatr. 49, 887–897. https://doi.org/10.1177/0004867415600891.

Katrak, P., Bialocerkowski, A.E., Massy-Westropp, N., Kumar, V.S., Grimmer, K.A., 2004. A systematic review of the content of critical appraisal tools. BMC Med. Res. Methodol. 4, 22. https://doi.org/10.1186/1471-2288-4-22.

Keeney, S., Hasson, F., McKenna, H., 2006. Consulting the oracle: ten lessons from using the Delphi technique in nursing research. J. Adv. Nurs. 53, 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x.

Khan, K., Kunz, R., Kleijnen, J., Antes, G., 2011. Systematic Reviews to Support Evidence-Based Medicine, second ed. Hodder Arnold, London, UK.

Krippendorff, K., 1980. Content Analysis: an Introduction to its Methodology. Sage, Newbury Park, CA.

Lipsey, M.W., Wilson, D.B., 2001. Practical Meta-Analysis. Sage, Thousand Oaks, CA.

Michie, S., Carey, R.N., Johnston, M., Rothman, A.J., de Bruin, M., Kelly, M.P., Connell, L.E., 2017. From theory-inspired to theory-based interventions: a protocol for developing and testing a methodology for linking behaviour change techniques to theoretical mechanisms of action. Ann. Behav. Med. 52, 501–512. https://doi.org/10.1007/s12160-016-9816-6.

Michie, S., Johnston, M., Abraham, C., Lawton, R., Parker, D., Walker, A., 2005. Making psychological theory useful for implementing evidence based practice: a consensus approach. Qual. Saf. Health Care 14, 26–33. https://doi.org/10.1136/qshc.2004.011155.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Wood, C.E., 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann. Behav. Med. 46, 81–95. https://doi.org/10.1007/s12160-013-9486-6.

Minas, H., Jorm, A.F., 2010. Where there is no evidence: use of expert consensus methods to fill the evidence gap in low-income countries and cultural minorities. Int. J. Ment. Health Syst. 4, 33. https://doi.org/10.1186/1752-4458-4-33.

Moyer, A., Finney, J.W., 2005. Rating methodological quality: toward improved assessment and investigation. Account. Res. 12, 299–313. https://doi.org/10.1080/08989620500440287.

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

National Institutes of Health, 2014. Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies. https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools/cohort. (Accessed 14 November 2016).

Norris, J.M., Plonsky, L., Ross, S.J., Schoonen, R., 2016. Guidelines for reporting quantitative methods and results in primary research. Lang. Learn. 65, 470–476. https://doi.org/10.1111/lang.12104.

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

Nulty, D.D., 2008. The adequacy of response rates to online and paper surveys: what can be done? Assess Eval. High Educ. 33, 301–314. https://doi.org/10.1080/02602930701293231.

Oxman, A.D., Guyatt, G.H., 1988. Guidelines for reading literature reviews. Can. Med. Assoc. J. 138, 697.

Oxman, A.D., Guyatt, G.H., 1991. Validation of an index of the quality of review articles. J. Clin. Epidemiol. 44, 1271–1278. https://doi.org/10.1016/0895-4356(91)90160-B.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A.C., Salsberg, J., Jagosh, J., Seller, R., 2012. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int. J. Nurs. Stud. 49, 47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002.

Ponto, J., 2015. Understanding and evaluating survey research. J. Adv. Pract. Oncol. 6, 168–171.

Protogerou, C., Hagger, M.S., 2019. A case for a study quality appraisal in survey studies in psychology. Front. Psychol. 9, 2788. https://doi.org/10.3389/fpsyg.2018.02788.

Protogerou, C., Johnson, B.T., Hagger, M.S., 2018. An integrated model of condom use in sub-Saharan African youth: a meta-analysis. Health Psychol. 37, 586–602. https://doi.org/10.1037/hea0000604.

Schroter, S., Glasziou, P., Heneghan, C., 2012. Quality of descriptions of treatments: a review of published randomised controlled trials. BMJ Open 2. https://doi.org/10.1136/bmjopen-2012-001978.

Singleton, R.A., Straits, B.C., 2009. Approaches to Social Research. Oxford University Press, New York, NY.

Stephens, A., Bohanna, I., Graham, D., 2017. Expert consensus to examine the cross-cultural utility of substance use and mental health assessment instruments for use with indigenous clients. Eval. J. Australas. 17, 14–22. https://doi.org/10.1177/1035719x1701700303.

Streiner, D.L., Cairney, J., 2007. What's under the ROC? An introduction to receiver operating characteristics curves. Can. J. Psychiatr. 52, 121–128. https://doi.org/10.1177/070674370705200210.

Velligan, D.I., Weiden, P.J., Sajatovic, M., Scott, J., Carpenter, D., Ross, R., Docherty, J.P., 2010. Strategies for addressing adherence problems in patients with serious and persistent mental illness: recommendations from the expert consensus guidelines. J. Psychiatr. Pract. 16, 306–324. https://doi.org/10.1097/01.pra.0000388626.98662.a0.

Waggoner, J., Carline, J.D., Durning, S.J., 2016. Is there a consensus on consensus methodology? Descriptions and recommendations for future consensus research. Acad. Med. 91, 663–668. https://doi.org/10.1097/acm.0000000000001092.

Wright, K.B., 2005. Researching internet-based populations: advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. J. Computer-Mediated Commun. 10, JCMC1034. https://doi.org/10.1111/j.1083-6101.2005.tb00259.x.

Yap, M.B.H., Pilkington, P.D., Ryan, S.M., Kelly, C.M., Jorm, A.F., 2014. Parenting strategies for reducing the risk of adolescent depression and anxiety disorders: a Delphi consensus study. J. Affect. Disord. 156, 67–75. https://doi.org/10.1016/j.jad.2013.11.017.

Young, M.D., Plotnikoff, R.C., Collins, C.E., Callister, R., Morgan, P.J., 2014. Social cognitive theory and physical activity: a systematic review and meta-analysis. Obes. Rev. 15, 983–995. https://doi.org/10.1111/obr.12225.

Zeng, X., Zhang, Y., Kwong, J.S.W., Zhang, C., Li, S., Sun, F., Du, L., 2015. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J. Evid. Base Med. 8, 2–10. https://doi.org/10.1111/jebm.12141.

Zhang, C.Q., Wong, M.C.-Y., Zhang, R., Hamilton, K., Hagger, M.S., 2019. Adolescent sugar-sweetened beverage consumption: an extended health action process approach. Appetite 141, 104332. https://doi.org/10.1016/j.appet.2019.104332.

Cleo Protogerou, Psychological Sciences and Health Sciences Research Institute (HSRI), University of California, Merced, Merced, USA, and Department of Psychology, University of Cape Town, Cape Town, South Africa; Martin S. Hagger, SHARPP Lab, Psychological Sciences, University of California, Merced, Merced, USA and Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland.

Martin S. Hagger's contribution was supported by a Finland Distinguished Professor (FiDiPro) Award from Business Finland (1801/31/2105).