An Umbrella Review of Reporting Quality in CHI Systematic Reviews: Guiding Questions and Best Practices for HCI

KATJA ROGERS, University of Amsterdam, Amsterdam, Netherlands
TERESA HIRZLE, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
SUKRAN KARAOSMANOGLU, Universität Hamburg, Hamburg, Germany
PAULA TOLEDO PALOMINO, Sao Paulo State College of Technology (FATEC) - Matão, São Paulo, Brazil
EKATERINA DURMANOVA, University of Waterloo, Waterloo, Canada
SEIJI ISOTANI, Harvard University, Cambridge, Massachusetts, USA
LENNART E. NACKE, University of Waterloo, Waterloo, Canada

Systematic reviews (SRs) are vital to gathering and structuring knowledge, yet descriptions of their procedures are often inadequate. In human–computer interaction (HCI), SRs are still uncommon but gaining momentum, which prompted us to explore how SRs are reported at CHI—the flagship HCI conference venue. To assess the reporting quality of CHI reviews that aim for a systematic approach, we conducted an umbrella review and applied reporting guidelines for SRs (PRISMA and ENTREQ) to our corpus. We contribute the first exploration of how well SRs at CHI meet guidelines for reporting quality, showcasing strategies for improvement in reporting and conducting SRs especially in the domains of appraisal, synthesis, and documentation (i.e., protocol development). Finally, we present guiding questions for HCI researchers and practitioners for reporting SRs, as well as suggestions for best practices.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Human-centered computing → Human computer interaction (HCI)**;

Additional Key Words and Phrases: systematic review, systematic literature review, umbrella review, reporting quality, best practices, methodology, synthesis

The research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (#RGPIN-2022-03353 and #RGPIN-2023-03705), the Social Sciences and Humanities Research Council of Canada (SSHRC) Insight Grant (#435-2022-0476), the Canada Foundation for Innovation (CFI) JELF Grant (#41844).

Authors' Contact Information: Katja Rogers (corresponding author), University of Amsterdam, Amsterdam, Netherlands; e-mail: k.s.rogers@uva.nl; Teresa Hirzle, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark; e-mail: tehi@di.ku.dk; Sukran Karaosmanoglu, Universität Hamburg, Hamburg, Germany; e-mail: sukran.karaosmanoglu@uni-hamburg.de; Paula Toledo Palomino, Sao Paulo State College of Technology (FATEC) - Matão, São Paulo, Brazil; e-mail: paula.palomino@fatec.sp.gov.br; Ekaterina Durmanova, University of Waterloo, Waterloo, Canada; e-mail: edurmanova@uwaterloo.ca; Seiji Isotani, Harvard University, Cambridge, Massachusetts, USA; e-mail: seiji_isotani@gse.harvard.edu; Lennart E. Nacke, University of Waterloo, Waterloo, Canada; e-mail: lennart.nacke@acm.org.

ACM 1557-7325/2024/11-ART57
https://doi.org/10.1145/3685266

ACM Reference format:
Katja Rogers, Teresa Hirzle, Sukran Karaosmanoglu, Paula Toledo Palomino, Ekaterina Durmanova, Seiji Isotani, and Lennart E. Nacke. 2024. An Umbrella Review of Reporting Quality in CHI Systematic Reviews: Guiding Questions and Best Practices for HCI. *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 57 (November 2024), 55 pages.
https://doi.org/10.1145/3685266

1 Introduction

More articles are published at the ACM Conference on Human Factors in Computing Systems (CHI) every year. Keeping track of the state of research within **Human–Computer Interaction (HCI)** is continuously becoming more difficult—even when focusing only on HCI subfields. The value of papers that provide a survey, overview, or review of existing literature on a topic thus increases every year. These papers gather knowledge and evidence. They structure, map, or chart the findings into synthesized summaries of the field (also called *research synthesis* [28]), ideally systematically: i.e., **Systematic Reviews (SRs)** [23]. Furthermore, in some cases such papers can derive new insights from the field, often by identifying research gaps, or—more rarely—by developing new intermediate-level knowledge [61] in the form of guidelines, taxonomies, frameworks, or design spaces.

In these SRs, rigorous and comprehensive reporting reduces bias to aid in producing trustworthy findings for the field: it both "*distinguish[es] SRs from traditional reviews [and is] a necessity and a hallmark of any well conducted SR*" [6]. However, SRs often follow informal methodologies, disregard basic reporting standards, or fail to provide a rationale for methodological choices, as reported in several different disciplines [49, 63, 101]. While recent work by Stefanidi et al. [135] gives an overview of contribution types, general topic areas, and databases in general reviews, we are not aware of an in-depth exploration of SR reporting in HCI to date. Yet in our experience as researchers and reviewers, there seems to be a lot of confusion about reviews as a systematic methodology. This is partly driven by problems with terminology (e.g., what counts as "systematic"?), and partly due to the at first glance bewildering array of potentially relevant methodologies, guidelines, and resources available. These cater to reviews as a whole, individual stages within reviews, and specific review types developed for particular fields. This makes it hard to discern which steps to conduct and report—and how—when undertaking an SR of the literature.

Yet research on methods and the wide variety of guidelines and suggestions for systematizing research synthesis have galvanized evidence-based and policy-driven research as a gold standard in other fields [88]. There is thus a lot of useful information on how to conduct and report SRs available in other fields like health sciences [6, 53, 108], environmental science [52, 151], social sciences [88, 115], and software engineering [71, 72]. By exploring how these different guidelines might apply within HCI, we will be able to draw upon a lot of valuable and substantial resources to improve the conduct and reporting of SRs in our field. To do so, we need to first learn how our field currently conducts and reports SRs and how well common guidelines fit the kinds of review found in our field. The first inclination may be to explore quality of conduct, i.e., whether the methodological choices make sense for that particular review question. However, that estimation can only be made if the reporting quality is high enough, and thus *reporting quality* is our focus for this work.

This research is primarily driven by the following questions:

- We aim to assess how reviews are reported in HCI when they come with the label of "systematic"—using CHI as a flagship HCI conference as our sampling site: one that has "*shaped*

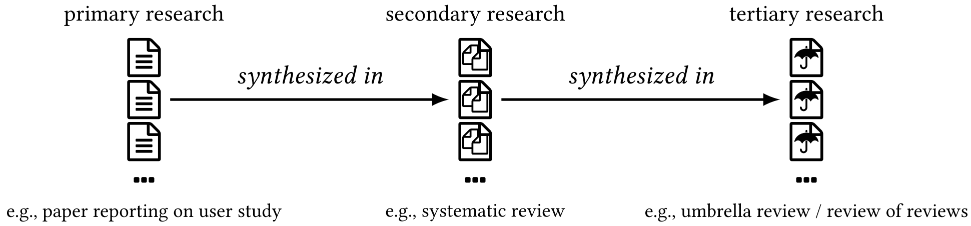


Fig. 1. Degrees of separation from the primary research: papers reporting on empirical data directly are considered primary research. These are synthesized in secondary research like SRs. These, in turn, are synthesized in tertiary research, like umbrella reviews/reviews of reviews.

and defined the field of HCI" [90]. In particular, do SRs at CHI provide enough information for future researchers to be able to replicate them? This question targets reporting quality of CHI reviews, which we assess using two checklists: "*Preferred Reporting Items for SRs and Meta-Analyses,*" i.e., PRISMA [108]; and "*Enhancing transparency in reporting the synthesis of qualitative research,*" i.e., ENTREQ [142].

- Complementary to the previous question: how well do common checklists fit CHI reviews? As there are many different types of SRs, and existing checklists originate from different methodological niches and scientific fields, we want to explore how well the two checklists we use to assess reporting quality suit the kind of research done at CHI.

Additionally, we are especially interested in the synthesis methods used, as synthesis is a key step of review methodology that depends strongly on the **Research Questions (RQs)**, and may thus vary in HCI compared to other fields. We hope that by exploring these questions, this paper will serve as a primer for researchers new to and interested in conducting SRs in HCI.

To answer the above questions, we conducted an *umbrella review*[1] of reviews stating a systematic approach published at CHI. With this, we provide in-depth insight into reporting quality in such papers, as a complementary work to the broader overview of literature reviews generally in HCI that already exists [135]. Umbrella reviews are a type of SR that synthesize findings not from primary research (e.g., findings from papers reporting on user studies), but from secondary research (e.g., findings from SRs of primary research) [5]. That is, the unit of analysis for an SR is most commonly a research paper reporting directly on some kind of data. For an umbrella review, the units of analysis consist of review papers that synthesize research papers—see Figure 1. We followed guidelines for SRs, and umbrella reviews in particular [6, Ch. 10], and in doing so, we developed, pre-registered, and followed a formal protocol for our review.

In our analysis, we applied PRISMA and ENTREQ as reporting standards from other fields to CHI SRs. Our findings show that most of the reviews were poorly compliant with these standards, particularly in terms of in the review stages concerning appraisal,[2] the **Risk of Bias in Synthesis (ROBIS)** tool,[3] synthesis of results itself[4], and general documentation (e.g., protocol[5] usage).

[1] We address our choice of terminology in more detail in the paper's limitations.

[2] Critical appraisal or quality assessmemt of the selected papers, for additional insight into the field and for identifying low-quality or poor-fit papers for subsequent exclusion or separate weighing / handling in the analysis [118].

[3] Attempts to identify various kinds of bias that may be part of the synthesis due to characteristics of the selected papers, for example, when identified studies only explore specific demographics [108, 150]. As indicated by the term "risk of bias" this reflects the positivist leanings of the origins of SRs that positions bias as a negative influence—but in other epistemic perspectives this could also involve reflecting on the role of the researchers in gathering the data or selecting the papers.

[4] The stage in which findings are developed based on the gathered data from selected papers [108].

[5] In the context of SRs, we must distinguish between general guidance—which is sometimes named protocol (e.g., the PRISMA protocol [108])—and (usually *a priori*, and sometimes pregistered) documentation specific to a single SR [47].

Additionally, there were numerous difficulties in applying the PRISMA and ENTREQ items. This suggests that SRs at CHI–or more generally in the field of HCI–may need a different, specialized set of reporting standards. We therefore conclude our work with a set of guiding questions and suggestions for best practices with the purpose of prompting reflection on the SR process. The guiding questions incorporate the findings of this umbrella review as well as our combined experience in conducting and reviewing research synthesis to aid researchers in the reporting of future SRs in the field of HCI. Finally, based on existing literature on research synthesis methodology, the corpus of our umbrella review, and our experience of applying the PRISMA and ENTREQ items to the corpus papers, we present suggestions for best practices as a first set of recommendations for improving the quality of both reporting and conduct in HCI reviews.

With this paper we hope to bring methodological clarity and rigor in research synthesis to the forefront of discourse in the CHI and HCI research communities. By raising awareness and summarizing best practices of reporting quality criteria in this context, we aid researchers in the field who are planning, conducting, documenting, and reviewing SRs, and thereby improve the future of research synthesis in HCI.

2 Background

Research synthesis has a history that reaches as far back as the 18th century, and rapidly became more systematic and established over the 20th century [23]. This was driven in particular by uptake by the medical field in the 1970s and 1980s, resulting in the establishment of international research organizations,[6] the **Joanna Briggs Institute (JBI**; https://jbi.global**)**, the Campbell Collaboration (https://campbellcollaboration.org/), and the **Collaboration for Environmental Evidence (CEE**; https://environmentalevidence.org/**)**, dedicated to supporting, as well as gathering and developing resources for SRs [6].

SRs gather information found in primary research (e.g., papers reporting directly on new results), and attempt to synthesize the information found in multiple papers into an overview, or new insight. This type of research can identify and synthesize different evidence for a specific RQ (e.g., based on multiple studies of similar or different type), reliably summarize the state of art of a defined area of the field, highlight research trends or gaps, and even develop new forms of intermediate-level knowledge like models or frameworks [88, 132]. This is particularly important given the drastic increase in academic publications over the past years and decades: Cooper [28] reports researchers in 1971 already speaking of being overwhelmed by the number of publications. In words of our current times, "*We are in the grip of a pandemic of evidence*" [36]. Like other fields, HCI produces far more publications than can reasonably be read each year.[7] This makes it continuously harder to gain or maintain an overview of even a subset of the field. SRs help other researchers by systematically summarizing and synthesizing the knowledge found in a broad sample of thematically related papers, among other functions already noted.

Oulasvirta and Hornbæk [107] have pointed out that "*ignorance of previous results decreases the problem-solving capacity we possess as a field*" and that "*we rarely identify and address anomalies in research.*" SRs can work against these issues, and directly contribute to the criteria for problem-solving capacity they propose (building on Laudan's philosophy of science [82]): SRs can identify stakeholders in specific contexts and use cases (e.g., primary, secondary and tertiary users, or individuals with specific occupations) and their key points for improvement (*significance*), chart relevant problem and evaluation characteristics (*effectiveness*), identify and raise awareness of resources and information (*efficiency*), assess how solutions have been applied and identify new

[6] For example, Cochrane (https://www.cochrane.org/).

[7] A search for "HCI" in the ACM Full-Text Collection yields 6,652 publications for 2023 alone.

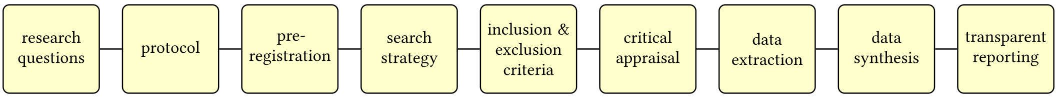


Fig. 2. A generic SR may consist of these components. It should be noted that some SR types will not require a protocol or pre-registration; these two steps can vary depending on the methodology and conventions of the field. Finally, we emphasize that this reflects the components common to other fields' understanding of SRs. Which components should be included in SRs in HCI is even more unclear than in other fields.

contexts for future application (*transfer*), and directly address issues of empirical validity and robustness of solutions (*confidence*). For example, an SR could determine who is interested in new solutions in a given context and which solutions might be most valuable to them, to what extent current solutions do or do not work effectively for them, and which new solutions exist and how they can be implemented efficiently. Further, it can explore a broader, more holistic view of the contexts in which these new solutions might work beyond individual studies (how do the solutions transfer), and to what extent we can trust the solutions in the first place given the results in the literature so far.

2.1 Procedural and Structural Components

While there is extensive debate about what counts as an SR [11, 44, 50, 94], we here present the steps involved in creating a generic SR based on multiple sources: Aromataris and Munn's JBI manual [6, Ch. 1], the PRISMA Statement article by Page et al. [108] and the protocol extension PRISMA-P [98], as well as Siddaway et al. [132], and Littell [88]. We chose these because the JBI manual offers guidance for the protocol development, conduct, and reporting of many different types of SRs. We also included the PRISMA guidelines because it is the guideline we have most often seen referenced in HCI reviews. Finally, we looked at Siddaway et al. [132]'s more general positioning of narrative vs. SRs, and Littell [88] for guidance more focused on meta-analyses. However, we note that of course there are many other overlapping references influencing our choice for this background information more broadly.

While this is of course only one perspective on what counts as "*systematic,*" we use this as a starting point. As illustrated in Figure 2, an SR is likely to contains the following components:

(1) RQs: SRs should formulate and aim to address a specific RQ (or fulfill a specific objective [108]). This component shapes the review's other components, and is thus essential information for reporting and conduct quality.
(2) Protocol: By developing an *a priori* detailed documentation of the planned review, researchers can minimize selective reporting and transparently document changes to the review methods [5, 98].
(3) Pre-registration: Similar to pre-registrations of experimental studies [103], many guides to SRs recommend that protocols of reviews be pre-registered (e.g., with the OSF,[8]). In some fields, protocols even undergo peer review [88, 124].
(4) Search strategy: This refers to the procedure by which relevant studies or papers are identified, and again aims to increase transparency and replicability [108]. It covers information like the exact query for each database, any filters or limits that were applied, and validation strategies to check for known relevant studies. It should generally also detail when each search was last conducted, and whether additional procedures such as snowballing [58] were applied.

[8] OSF, https://osf.io/

(5) **Inclusion Criteria (IC)** and **Exclusion Criteria (EC)**: These criteria formalize the process by which studies or papers are included in a review through a screening process [88]. The goal of these criteria is to minimize sampling bias, and support both transparency and replicability [88, 132].
(6) Critical appraisal: Also termed quality assessment, this stage explores included studies' methodological quality—although what is understood as that concept varies [88, 132]. For example, PRISMA 2020 [108] refers primarily to risk of bias, for which they distinguish between the kind resulting from individual studies (e.g., sampling bias [75]), and the kind that results from synthesis of multiple studies (e.g., publication bias [51]). Appraisal results can then be used to weigh information from particular studies during synthesis [88, 132], or exclude low-quality studies from the corpus papers [6].
(7) Data extraction: Standardized forms or tools should be employed to gather information from the included papers, again to minimize risk of bias [5]. Many guidelines strongly recommend the use of double extraction, i.e., having two researchers conduct extraction on all papers separately to safeguard against errors.
(8) Data synthesis: Having extracted information from the identified relevant studies or papers, the resulting data needs to be analyzed to discern findings: "*assembling the jigsaw of evidence*" [115]. **Analysis and Synthesis (AS)** methods can be quantitative, qualitative, or mixed methods. Guidelines often focus on only or primarily one specific kind of synthesis, for example, meta-analysis (e.g., PRISMA [108]), or thematic synthesis (e.g., ENTREQ [142]).
(9) Transparent reporting: Finally, the above methods and their results should be reported comprehensively, again to enable transparency and reproducibility [51].

Finally, we note that the terminology above and many of the components themselves reflect the (post-)positivist epistemic origins of SRs, hence the frequent goal to reduce bias of various kinds. We describe our own epistemic perspective in more detail in the discussion, but at this point note that we do not believe that bias is always a negative thing, and that certain types of bias can also be a strength if critically and transparently reflected on.

2.2 Types and Families of SRs

There are many different types of reviews. The most prominent distinctions exist between SRs and non-SRs (also called *traditional* or *narrative reviews* [115]). Unfortunately, there are many different opinions on which review types should count as systematic, i.e., where the line should be drawn between systematic and non-SRs. For some, an SR requires the upholding of strict criteria, e.g., requiring a critical appraisal stage, extensive search strategies that incorporate multiple databases and grey literature, and formalized synthesis methods [50]. Others construe the term more flexibly, often focusing requirements for systematic procedure on the search strategy step only, or loosening requirements for some factors (e.g., not advocating searches for grey literature).

Further, some researchers distinguish between *conventional* SRs and *mapping studies* (or *scoping studies* [113]), with mapping studies featuring "*coarser-grained RQs*" and correspondingly different synthesis methods [73]. As noted by Kitchenham et al. [73]: "*This distinction [...] can be somewhat fuzzy.*" We note that Kitchenham et al. [73] classified both of these two types as systematic, although this is not an opinion shared by all researchers (e.g., Arksey and O'Malley [4]). However, even those with a strict interpretation of "systematic" increasingly accept qualitative forms of research synthesis, which are not always called SRs. Thus, confusingly, there are SR types that are not termed "SRs." The confusion about terminology and methodological classification has been compounded by the fact that the foothold of SRs in academia has increasingly prompted traditional reviews to

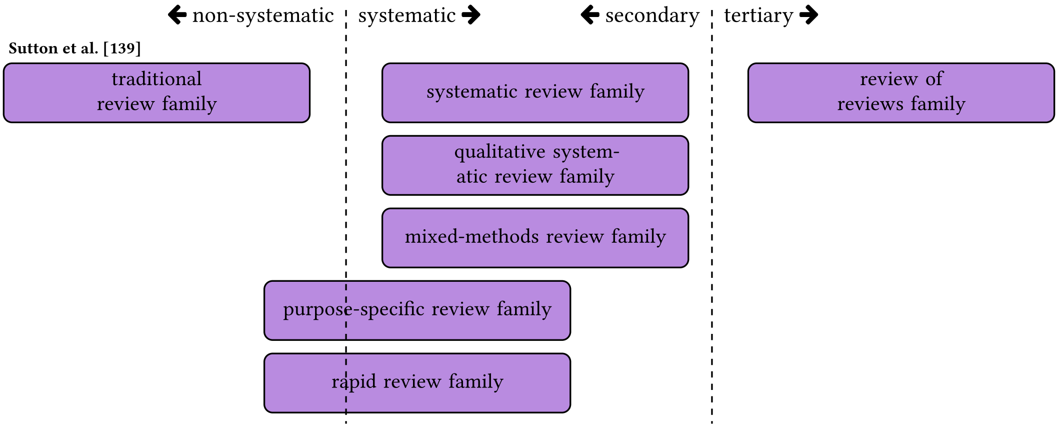


Fig. 3. Review classification as systematic/non-systematic and secondary/tertiary, based on our understanding of Sutton et al. [139]'s broader review families classification. The purpose-specific and rapid review families were drawn to overlap into the non-systematic area, as this classification depends on one's definition of systematic (e.g., does the term require a critical appraisal? Or does the term simply require that a critical appraisal is done *if one is necessary* for that particular RQ?).

also adopt "*greater systematicity*" [139], e.g., through greater effort in transparent reporting and/or reproducibility.

In 2019, Sutton et al. [139] presented an overview of 48 distinct types of reviews, within a categorization of seven review families based on review purpose and characteristics (see Figure 9 in the appendix). This consisted of: the *traditional review family* (non-systematic but also increasingly using systematic aspects), the *SR family* (including the quantitative "SR"), the *review of review family* (including umbrella reviews), the *rapid review family* (which loosen requirements for systematicness), the *qualitative review family* (including the "qualitative SR"), the *mixed methods review family*, and the *purpose specific review family* (which includes scoping studies and systematic mapping reviews). Other review categorizations exist, but often focus predominantly on specific types of reviews, and/or can be mapped within Sutton et al. [139]'s categorization of review families[9]. Progress in review methodologies in different fields has advanced extensively in the past decade. In particular, methods of qualitative research synthesis have boomed [53, 139]. Increasingly, traditionally more quantitative-focused disciplines like medicine are starting to adapt their SR processes to include qualitative data [53]. Given the proliferation of review types across different disciplines, and the inclusivity of Sutton et al. toward review types of different methodologies (e.g., not primarily focusing on meta-analyses or quantitative, conventional SRs), we consider this a very useful categorization. Thus for the purposes of this paper, we subscribe to this categorization, and show our understanding of how the review families are classified as systematic/non-systematic or secondary/tertiary in Figure 3. Examples of each are shown in Figure 9 in the appendix.

Umbrella Reviews. With the increasing number of SRs in many fields, researchers are now seeking ways to synthesize evidence across secondary research. This can be done through umbrella reviews [139] (or review of reviews [133] or tertiary study [73]), which can therefore be classified as tertiary research (in contrast to primary and secondary research, see Figure 1). They are commonly used to assess and synthesize the results of multiple SRs that explore the same question. We note that

[9] For example, JBI's review types [6] in the medical field extend Sutton et al. [139]'s to include review types prevalent in the medical fields, e.g., diagnostic test accuracy reviews.

our umbrella review differs from this, as we conducted a systematic umbrella review of SRs that explore *different* questions.[10] Similar explorations have been conducted in HCI (e.g., a review of literature reviews [135]) and other fields (e.g., a scoping review of SRs in software engineering [30]. More conventional examples focusing on reviews of the same topic include Jemioło et al. [67]'s umbrella review of affect recognition studies using emotion elicitation stimuli, and Kari [69]'s "SR of SRs" on exergaming benefits.

2.3 SRs in HCI

As mentioned, researchers' interest in and acceptance of SRs has spiked in various disciplines throughout the past decades [23]. The medical field in particular has driven the evolution of this methodology, through a variety of guidelines aimed in particular at synthesizing results of randomized controlled trials and interventional study designs (e.g., [108]). As the methodology has wandered from health sciences to other disciplines, researchers in those fields have developed their own guidelines based on the types of research and epistemology inherent to their discipline (e.g., the CEE's guidelines [37] for environmental science, and Kitchenham [71]'s guidelines for software engineering). As our findings (later) and a recent review of literature reviews in HCI [135] suggest, it seems that SRs as a research contribution may be growing in popularity in HCI as well. In their review, Stefanidi et al. [135] summarize the contribution types, review topic areas, methods used (i.e., which databases were searched and whether inter-rater reliability was calculated for screening) as well as publication venues of HCI literature reviews. However, reporting quality is not the core interest of their review. They noted how many papers either refer to reporting standards like PRISMA or QUOROM or provide a flowchart to depict the selection process. But—understandably, given the extensive breadth of their analysis—the paper refrains from taking a critical perspective or providing a more than surface-level summary when it comes to reporting quality. Yet this leaves an evident gap that this paper aims to fill: an in-depth account of reporting quality, specifically in reviews labeled as "systematic."

Viewing HCI as a field with its own particular types of research problems [107], the question emerges how well the reporting guidelines of other disciplines transfer to SRs in HCI. For instance, HCI welcomes a broad array of approaches under the banner of empirical problem-solving alone [107] (e.g., consider research through design [153]), and features a strong focus on user studies as well as indirect studies of user data (e.g., content analysis of tweets [66]), which affects how data are reported [19]. Additionally, it tackles conceptual and constructive research problems [107]. By virtue of these traits, SRs in HCI would benefit from a detailed analysis of reporting quality and may need adapted reporting guidelines of their own.

Finally, we note that researchers like Hornbæk et al. [62] have lamented that HCI journals and conferences focus on originality and novelty over the consolidation of existing work. We believe that rigorously conducted and transparently reported SRs can play a key role in a similar spirit of consolidating knowledge of the field. Nevertheless, HCI still grapples with prejudice against literature reviews as an academic contribution: in our experience as researchers and reviewers, we have seen literature reviews be judged harshly for not going deep enough even when the focus was on breadth and overview, and not going broad enough when the focus was on depth and specificity. We have even encountered reviewers who see no benefit to SRs at all, a notion we strongly disagree with. While this variety in perspectives on SRs is present in many fields, we suspect it is more pronounced within HCI due to our field's interdisciplinarity [123]. Whether a review aims to showcase and establish evidence within prior work and/or point out research gaps for future research directions and/or develop new theories and models from prior work: each of these

[10] We nevertheless apply the term umbrella review because our unit of analysis consists of SRs.

purposes can benefit our academic community by introducing scholars to a subfield, communicating and consolidating knowledge, inspiring future work, and/or presenting new knowledge in its own right. HCI as a research community needs to initiate a discussion about what we expect from SRs. To inform this discourse, however, we first need to understand more about the current state of SRs in HCI.

3 Method

We conducted an umbrella review to identify and assess existing reviews at CHI that claim a systematic approach. Preliminary searches of Google Scholar,[11] the ACM digital library,[12] and Scopus,[13] yielded no prior umbrella review of SRs at CHI (nor in HCI in general).

3.1 Protocol, RQs, and Rationale

Protocol Development and Registration. We began with the development of a protocol, following the JBIs recommendations [6] for the development of an umbrella review protocol. We chose this particular manual as it considers both quantitative and qualitative syntheses in umbrella reviews. Due to its origins in the medical field, the guide adheres to the **Population-Intervention-Comparison-Outcome (PICO)** framework, e.g., in developing the RQ. We strayed from the JBI's recommendations for these aspects, as our review does not cover an interventional RQ. Our protocol included our formalized RQs, IC and EC, search strategy and initial search results, and planned procedures for study selection, critical appraisal, data extraction (using a modified version of the JBI data extraction form), and synthesis. The protocol was registered with the **Open Science Foundation (OSF)**; it is available for review.[14]

RQs. We defined the following RQs, with the first two as the main ones:

RQ1a: How well are common existing guidelines for systematic literature reviews adopted by SRs in CHI publications?

For this purpose, we employed PRISMA and ENTREQ as such guidelines for reporting literature reviews. We introduce these guidelines and rationalize our choice below. However, we note that this application of guidelines for this work should not be understood as us endorsing them. While we use the guidelines to structure our investigation of reporting in CHI SRs, "failing" to meet a specific item should not inherently be understood as a negative thing. Rather, the application of any guideline must also suit the research being done, we thus also aimed to investigate:

RQ1b: How well do these guidelines work for the kinds of SRs present in CHI publications?

Finally, based on our own experience with existing SRs, and guidelines for conducting SRs, we noted that the synthesis stage of reviews is often neglected in both guidelines and reporting. Yet this depends strongly on the type of RQ being asked, and so may be key in displaying differences to SRs in other fields. Our final and secondary RQ thus focused in particular on this stage:

RQ2: What kind of syntheses methods are used in SRs at CHI?

CHI: Flagship Venue and Representative Sample. There is precedence of limiting the search of an SR to a specific conference, including to only CHI publications, e.g., [19, 87, 90]. CHI is a highly

[11] https://scholar.google.com/
[12] Specifically, the Guide to Computing Literature; https://dl.acm.org/
[13] https://www.scopus.com/
[14] Anonymized protocol: https://osf.io/hjgrs/?view_only=644d3eecd31b4acf89f33e86b2b00232

influential HCI venue [90], with researchers applying a similar CHI focus to their review referring to both its influence in terms of citations and impact factor [87]. Like in previous reviews, it can thus be considered a "*defensible, purposeful limitation*" [19]. As others before us [107], we focus on CHI publications as a representative snapshot of the general field of HCI, although of course future work will have to explore how well the trends we uncover generalize to other HCI conferences. We discuss this further in our limitations section.

Rationale for PRISMA and ENTREQ as Checklists. Our goal in this umbrella review is to gain insight into the quality of SR reporting in CHI papers. For this, the first step needs to focus on whether the reviews in our corpus hold enough information to follow what was done. This means that we refrain at this stage from using checklists that introduce more subjective assessments of quality: we ask questions like "does this review report an RQ and IC?" and leave questions like "do the IC make sense in light of the RQ?" for future work. This allows us to follow a clear-cut answering scheme (yes/no/partially) and perform a reasonably quick and reproducible assessment of CHI reviews to date.

Even with this narrowed scope, there are many checklists available that could be applied to check for review reporting (see Table 13 in the appendix). As HCI conferences and journals rarely set guidelines for reporting [7], we oriented our selection to other fields. We chose PRISMA [108] and ENTREQ [142]: in the following, we briefly introduce PRISMA and ENTREQ and explain why we chose them (for a discussion of alternatives, refer to our limitations section).

PRISMA. The PRISMA protocol [108] is a guide to help researchers report their SRs and meta-analyses. It consists of 27 items (42 including sub-items) to facilitate researchers' transparent reporting process (see PRISMA checklist items in Tables 3–11 in the appendix). While the PRISMA checklist contains items for general reporting purposes (e.g., specifying a RQ), it includes many items that are only applicable to SRs using quantitative synthesis of interventional studies and randomized controlled trials or specifically meta-analyses.

We chose PRISMA as one of our critical appraisal checklists because they are well established as guidelines for meta-analyses in particular, yet also used as guidelines for quantitative SRs more generally. Further, in our experience as HCI researchers and reviewers, when SRs in HCI profess to use a systematic process, they often mention this set of guidelines. However it is not well suited to more qualitative approaches. We thus used PRISMA as a checklist for reviews focusing on quantitative synthesis, and a different checklist for qualitative syntheses.

ENTREQ. The ENTREQ statement [142] is a reporting guide that aims to facilitate transparent reporting in the synthesis of qualitative research. This caters to qualitative synthesis methodologies such as thematic synthesis [141] (see also Figure 9 in the appendix). This guide contains 21 items categorized as relating to a review paper's introduction, methods and methodology, literature search and selection, appraisal, and synthesis of findings [142].

In our review, we chose ENTREQ as a less well-known but well-developed checklist for transparency in qualitative reviews. When review papers in our corpus were geared towards qualitative synthesis methods (e.g., thematic analysis), we applied this checklist.

3.2 Eligibility and Search: IC and Procedure

IC. Based on our RQs, we defined our IC and EC as follows:

(IC1)—The paper contains a review of relevant items of interest (papers, apps, and so forth) that was conducted in a systematic manner.

(EC1)—The paper is not a full paper (e.g., an extended abstract accidentally classified as a research article).

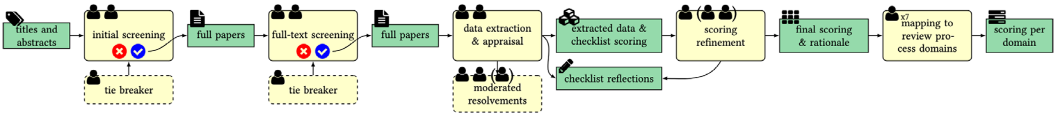


Fig. 4. Our procedure consisted of double screening (initially on titles and abstracts, then on full papers), with a tie-breaking third coder for disagreements. We also conducted double extraction and double appraisal, and resolved disagreements in discussion sessions that were moderated by the first author in the interest of consistent coding. The scoring was refined (in consultation with the respective coders) to improve consistency further. Checklist items were mapped to review process domains (by all authors), after which the first author calculated the scores per domain for the corpus papers.

(EC2)—The paper contains no review of papers or apps (or other units of analysis).

(EC3)—The paper's review method is explicitly proclaimed to be unsystematic or informal.

(EC4)—The paper's review method is not proclaimed to be systematic—either by adjectives such as "systematic" or "methodical" or "systematized," or by claim of specific SR type (e.g., scoping review, rapid review)—at any point in the paper.

Search Procedure. We used the key terms "SR," "systematic" and "review" to build a query for the ACM digital library (The ACM Guide to Computing Literature, not the more limited ACM Full-Text Collection). (We discuss alternatives and our rationale in our limitations section.) In particular, we sought papers that used the term "SR" anywhere, or papers that used both "systematic" and "review" anywhere, using the following exact query:

```
[All: ``systematic review''] OR [[All: ``systematic''] AND [All: ``review'']].
```

We used the ACM web interface's filter options to narrow the results down to those classified as research articles, and published at CHI. Our initial searches in the month prior to protocol registration resulted in data mostly ranging between 690 and 698 records.[15] The final search after protocol registration was conducted on 18 June 2021 and resulted in 695 records. The first author used a custom Python script to identify duplicates based on title and year of publication, and removed these (n = 2) after a manual check. The list of 693 unique records was uploaded as a CSV file to our pre-registration.[16]

3.3 Screening Stages: Finalizing the Corpus

Screening. The remaining 693 unique records were screened based on their title and abstract. The screening process was conducted by all seven authors using Dovetail.[17] Each paper was screened by two authors, while a third author was assigned as tie-breaker in case of disagreements (see Figure 4). This resulted in a pool of 151 records.

Full-Text Eligibility. Full-text screening of the 151 records was again conducted by all seven authors, again using Dovetail to keep track of screening decisions. The papers were again randomly distributed among the authors for full-text eligibility screening, such that the authors largely did not screen the same papers in both phases. As before, each paper was screened by two authors, with a third author assigned as tie-breaker. The authors asynchronously discussed any questions or edge cases. We also introduced several meta / management tags for this stage: "Non-Literature Review" and "Edgecase" were created for this a priori (to classify reviews that did not assess papers as their unit of analysis, and reviews that were difficult to code as include or exclude, respectively).

[15] Barring one day when the results swerved widely between 0, 266, and 692.
[16] Query results: https://osf.io/5mzwr/?view_only=1ab6ef51b01a4ebfb85f7d92281cb33e
[17] https://dovetailapp.com/

Edge cases consisted of papers such as Gathani et al. [42] or Li et al. [86] that followed steps of an SR, but never explicitly claimed to use a systematic process—without an explicit claim, these were excluded. Additionally, during this process we noticed that several papers had been misclassified by the ACM digital library (e.g., classified as part of CHI proceedings in the system while actually part of HCI Korea proceedings [83]); we excluded these as well. As a final step, the first author gathered all papers that were listed for inclusion, and extracted a reason for inclusion from them (e.g., the line in the paper where the authors identify it as an SR or as using a systematic approach). Any edge cases and unclear inclusions in this list were then discussed with the respective coders until final agreement on inclusion was reached. As part of this step, we excluded two papers which self-described as a "comprehensive literature review" [65] and a "meta-review [...conducted] to build up an in-depth understanding" [84]—which had initially been understood by coders as implying systematicness, but ultimately this potential claim was considered not explicit enough.[18] This resulted in a list of 51 papers for inclusion.

Manual Addition. While familiarizing ourselves with the 51 papers found through the screening stages, we came across a reference to a CHI paper that was also an SR, but not in our records. It matched our keywords and IC, and thus should have appeared in our initial dataset. We double-checked that this had not been the case, and thus attribute this to the foibles of the ACM DL (which we address in more detail later). Going forward, the first author instructed all researchers to pay attention to referenced examples of SRs in their assigned papers, and to flag any other potential additions for manual screening. This paper by Bargas-Avila and Hornbæk [8] remained the only paper flagged in this way, and was included in our analysis manually (see Figure 5, rightmost column). Our final corpus proceeding to the appraisal and extraction stage consisted of 52 papers, although for our synthesis we will focus on its 41 papers that focused on reviewing literature (as opposed to apps or other items). All corpus papers are listed in the supplementary materials, alongside additional information relating to the results of the appraisal, data extraction and coding, and analysis phases (see below).

3.4 Critical Appraisal and Data Extraction

Due to the nature of this review, the critical appraisal and data extraction held a degree of overlap that is usually not found in umbrella reviews (as they usually synthesize information from multiple reviews on the same topic, and seek an answer to a RQ shared by all of them). We again followed a similar process for data extraction and critical appraisal as in the screening stages: papers were distributed among the research team so that each paper was assigned to two researchers. The two assigned researchers conducted double extraction (i.e., independent from each other), to reduce errors and implicitness of bias [17].

We again used Dovetail to keep track of the papers (using tags to indicate which coder had done which task, to identify whether the review was a paper's main contribution, and to note review frameworks/examples that were being cited). We conducted this as a two-part process, with each coder first extracting basic data using the modified JBI data extraction form and voting for PRISMA, ENTREQ, or both. Disagreements about which checklist to apply were resolved before coders continued with checklist application.

To facilitate data extraction, we used Airtable,[19] to create web forms that directly input the data into a spreadsheet. As described in the protocol, we used a modified version of the JBI data extraction form (with minor changes to the protocol version to clarify items/improve phrasing).

[18] I.e., their inclusion would have been incompatible with our goal to better understand the CHI community's practices in the reporting of "*systematic*" reviews, specifically.

[19] https://airtable.com/

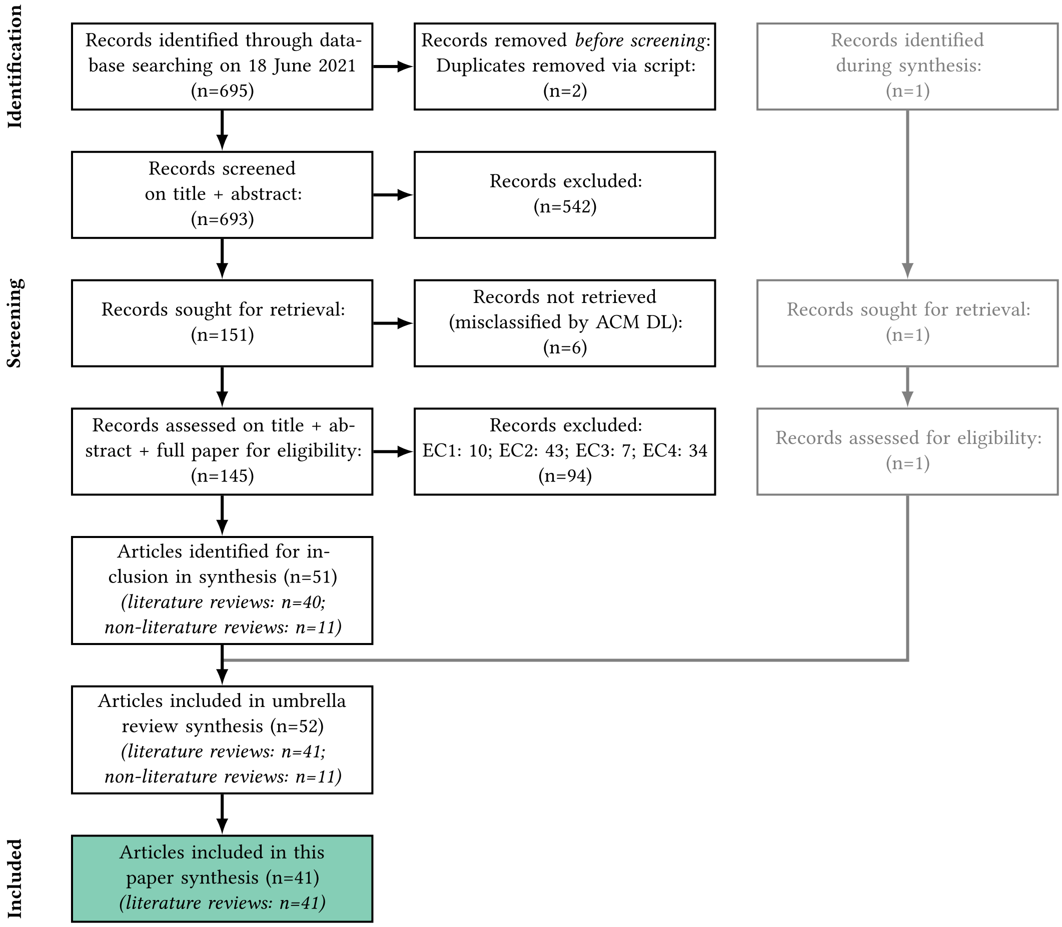

Identification
Screening
Included
Records identified through database searching on 18 June 2021 (n=695)
Records removed *before screening*: Duplicates removed via script: (n=2)
Records identified during synthesis: (n=1)
Records screened on title + abstract: (n=693)
Records excluded: (n=542)
Records sought for retrieval: (n=151)
Records not retrieved (misclassified by ACM DL): (n=6)
Records sought for retrieval: (n=1)
Records assessed on title + abstract + full paper for eligibility: (n=145)
Records excluded: EC1: 10; EC2: 43; EC3: 7; EC4: 34 (n=94)
Records assessed for eligibility: (n=1)
Articles identified for inclusion in synthesis (n=51) *(literature reviews: n=40; non-literature reviews: n=11)*
Articles included in umbrella review synthesis (n=52) *(literature reviews: n=41; non-literature reviews: n=11)*
Articles included in this paper synthesis (n=41) *(literature reviews: n=41)*

Fig. 5. This PRISMA flowchart[20] demonstrates the SR process for the stages from identification to the final sample of included papers. The final corpus papers are listed in the supplementary materials.

This consisted of data items such as the authors, the review objective, the unit of analysis, filters applied to the search, the time range of results, search details, appraisal and analysis details, and reported limitations. As part of our data extraction form, coders also indicated which checklist should be used: PRISMA, ENTREQ, or both (in addition to the corresponding votes in Dovetail). When there was agreement between both assigned coders, they then applied the chosen checklists to each paper, with *yes*, *no* and *partially* answer options. All forms (modified JBI data extraction form and PRISMA/ENTREQ checklist forms) are provided in the supplementary materials.

Choice of Checklist. We aimed to use the PRISMA for review papers using a quantitative method or referencing a guideline for quantitative synthesis (e.g., [35, 87]); ENTREQ for review papers using a qualitative method or referencing a guideline for qualitative synthesis (e.g., [91, 147]), or both if review papers synthesized both quantitative and qualitative research or referenced corresponding guidelines (e.g., [60]). Yet during the process of voting for checklists, we soon discovered that many

[20] Based on the 2020 template: https://www.prisma-statement.org/prisma-2020-flow-diagram.

review papers were not clearly either quantitative or qualitative in nature. For papers where the nature of the synthesis was unclear to us, we also decided to apply both checklists (e.g., [20]).

Double Extraction and Coding: Resolving Conflicts. We used custom R scripts to flag potential cases of disagreement for all forms: data extraction and the applied checklists for appraisal. For these instances, the two researchers discussed and resolved the difference in opinion, as a more engaged alternative to tie breaking. These discussions were moderated by the first author, to ensure that one researcher had a comprehensive overview of all papers, and to facilitate consistent coding. In case of disagreements among the papers assigned to the first author, discussions were held with two researchers, and disagreements were flagged for a third researcher to look over. All coders were instructed to be generous in their application of the checklists: When coders had significant difficulties deciding between two scores, we went with the more positive option (e.g., *partially* instead of *no*).

Our Reflections on Checklist Suitability. During the application of PRISMA and ENTREQ checklists to the corpus, all reviewers were asked to continuously note any issues with understanding checklist items, and their opinions on how well the items suited the papers they were reviewing. The authors noted these thoughts asynchronously in the form of sticky notes on a virtual Miro[21] whiteboard. In three additional sessions throughout the data extraction and appraisal process, the authors convened in online moderated sessions to discuss these. Based on the discussions in these sessions, as well as asynchronous discussions on the project channel, the first author iteratively derived a set of guiding questions and suggestions to help guide HCI authors of SRs in reporting all relevant aspects of their reviews (presented in the Guidelines section).

Criss-Cross Refinement Process for Consistency. Papers coded earlier in our process were initially coded more strictly than later papers (because coders had been exposed to fewer papers in the corpus, i.e., fewer variants of SRs at CHI), and our interpretation of some checklist items changed over time as we discussed them. To improve the quality of the scoring, we applied a final consistent refinement process via a form of axial coding [125]. For this, the first author extracted and documented the rationale via a paper-based coding for each item score (including intermittent discussions between five of the original coders). We followed up on this with item-based coding for the checklists to develop a clearer set of requirements for each answer option and improve consistency. The first author carried this out with occasional check-ins with the second, third, and last authors for input on ambiguous decisions. Our supplementary materials document all final checklist scores and scoring rationales (rows can be read for the paper-based coding perspective, and columns for the item-based coding perspective).

3.5 AS

We are not aware of a similar prior umbrella review with a similar style as ours on which we could have based our methodology. To establish a rigorous process, we chose a three-stranded approach that mixes descriptive summaries, categorization, quantization, and narrative synthesis:

(1) We generated descriptive statistical summaries (e.g., averages) and informally categorized the coded paper characteristics based on our modified JBI data extraction form. For example, we explored which papers cited which frameworks and which review methodologies they used. We derived this informal categorization inductively by iteratively tagging papers to cluster and narrow down potential reporting types. We chose this approach because we are confident it provides the broadest overview of the data.

[21] https://miro.com/

(2) We structured our synthesis of the checklist score data using an approach inspired by Whiting et al. [150]'s ROBIS domains. The ROBIS assessment tool lists questions meant to signal risk of bias and categorizes them based on the review process (domain) to which they refer. We borrow this domain-based approach for our synthesis and structure our assessment of the reporting quality of CHI reviews within these domains:
    (a) RQs and rationale;
    (b) eligibility criteria, identification, and selection;
    (c) data collection and appraisal;
    (d) analysis and synthesis;
    (e) ROBIS, and
    (f) uncategorized.

    With this approach, we can analyze the corpus reviews in a more granular way to better integrate our analysis of reporting quality between quantitative and qualitative reviews. This also allows us to identify where reporting quality in our field can be most improved by allowing us to compare the results of our (subjective) interpretation of checklist items for the different review process domains. For this purpose, we needed to assign and map the PRISMA and ENTREQ items to the review process domains (mapping described in detail below). We decided which checklists to apply to each paper based on the JBI votes as described in the previous section. We then defined and applied a score calculation (also described below) for each review process domain. This allowed us to quantify how well the corpus papers reflect the expected reporting in each checklist based on our interpretation of the items, grouped by each review process domain (e.g., comparing reporting relating to RQs and rationale vs. data collection and appraisal).

(3) We created a narrative synthesis of our notes, observations, and discussions on how well the checklist items could be applied to the CHI reviews in our corpus. We chose this more informal approach to begin to understand how suitable these checklists are for the CHI community.

Mapping PRISMA and ENTREQ to Review Process Domains. To map the PRISMA and ENTREQ items to the ROBIS review process domains, we conducted a card-sorting activity, configured via Optimal Workshop,[22] and independently completed by all seven authors. It asked each researcher to assign each of the 63 checklist items—42 PRISMA (sub-)items and 21 ENTREQ items—to one of the above-mentioned domains. 43 items (84%) were mapped to a domain with high agreement among the research team, a threshold which we set as at least five of the seven researchers. The remaining 20 items (~32%) were discussed asynchronously to resolve conflicts. The final result of this mapping process and the detailed card sorting results are presented in the appendix.

Score Calculation. Based on our classification of PRISMA and ENTREQ items by the review phase they refer to, we calculated an average score per paper for all contributing PRISMA and ENTREQ items, respectively. All scores are based on 2 points for a *yes* rating and 1 point for a *partial* rating. For example, the **Research Question and Rationale (RQR)** domain for PRISMA was mapped to items P3 and P4 (see Table 3). P_{total} for the RQR domain then is the sum of all scores for the P3 and P4 items across all 24 papers that were assessed using PRISMA (with the maximum possible being 4 points per paper—2 for each item), i.e., 96. Further, P_{mean} was then calculated for each domain as the sum total score for PRISMA (P_{total}) divided by the number of papers that had the PRISMA applied (24). ENTREQ's E_{total} and E_{mean} are calculated analogously.

[22] https://www.optimalworkshop.com/

Table 1. We Report the Average Score Per Paper for PRISMA (P_{mean}) and ENTREQ (E_{mean}) in the Context of the Highest Possible Score for Each Domain, Based on the Number of Contributing Items

	● RQR Research Questions & Rationale *relating to why the review was conducted*	● EIS Eligibility Criteria, Identification & Selection *relating to selecting the units of analysis (e.g., papers)*	● DCA Data Collection & Appraisal *relating to extracting information from the selected units of analysis (e.g., papers) and how they were critically appraised*	● AS Analysis & Synthesis *relating to analyzing and synthesizing the extracted information*	● ROB Risk of Bias in Synthesis *relating to how the synthesized information was critically appraised for this review (e.g., limitations in evidence and methodology)*	● MISC Uncategorized *everything else*
Items	P3, P4	P5, P6, P7, P8, P16ab	P9, P10ab, P11, P18	P12, P13abcdef, P17, P19, P20abcd, P23ad	P14, P15, P21, P22, P23bc, P25, P26	P1, P2, P24abc, P27
P_{mean}	3.54/4	7.25/12	3.33/10	13.67/30	4.33/16	1.38/12
P_{total}	85/96	174/288	80/240	328/720	104/384	33/288
Items	E1	E3, E4, E5, E6, E7, E9	E10, E11, E12, E13, E14	E2, E8, E15, E16, E17, E18, E19, E20, E21	n/a	n/a
E_{mean}	1.2/2	6.8/12	1.29/10	7.54/18	0/0	0/0
E_{total}	42/70	238/420	45/350	264/630	0/0	0/0
$P{+}E_{total}$	127/166	412/708	125/590	592/1350	104/384	33/288
	77.78%	58.19%	21.19%	43.56%	27.1%	11.5%

Additionally, we list the sum total score for PRISMA (P_{total}) and ENTREQ (E_{total}) for each domain based on all the review papers in the context of the highest possible score for that domain, based on the number of items. Finally, we list the total scores across both PRISMA and ENTREQ. All scores are based on 2 points for a *yes* rating and 1 point for a *partial* rating. For example, we calculcate P_{mean} as the sum total score for PRISMA (P_{total}) divided by the number papers that had the PRISMA applied (24). The maximum total for P_{mean} and P_{total} derive from the number of items contributing to that domain times under the assumption of *yes* ratings, i.e., times 2 points. For ENTREQ, E_{mean} and E_{total} are calculated analogously. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article. For better distinction, we use the abbreviation 'ROBIS' when referring to the original ROBIS assessment tool and its review process domains as a whole, but 'ROB' when referring to the specific review process domain of the same name within the overall tool.

Based on this, we report all calculated values, as well as the total score for each domain across both PRISMA and ENTREQ in absolute numbers and percentages in Table 1. Finally, we provide a normalized graph of the *yes*, *partial*, and *no* ratings for each domain in Figure 7.

We refer readers to Tables 3–11 in the appendix for a detailed overview of the exact items in each domain and notes on how the items were applied in practice. Further, the supplementary materials provide the comprehensive overview of all scores (including the *specific* scores[23]) for all papers—with rationale.

[23] Papers could receive different variants of the same scores. For example, for the item E5 (search databases), we gave a *partial* score for failing to provide the date of search but also a (different) *partial* score for failing to provide date of search *and* the rationale for the choice of database. Both *partial* scores were then handled equally (i.e., as 1 point) for the score calculation.

Critical Reflection. The application of the checklists and the card-sorting mapping into review domains is ultimatively a subjective process. To acknowledge how our individual backgrounds may have shaped this process, we describe this as a factor more explicitly. The research team for this paper consisted (at the time of coding and analysis) of seven authors at four different universities in three different countries and three different geographic regions, namely, North America, South America and Europe. Our backgrounds vary also in experience with SRs—all of us have been involved in at least one prior SR, but our perspectives while coding covered that of graduate student, PhD student, postdoctoral researcher, and associate and full professor. One researcher has conducted reviews primarily in software engineering, and thus is more aligned with Kitchenham [71]'s guidelines and perspectives. In contrast, the first author's practical experience was based on applying a formal qualitative synthesis method from medical fields to a research topic within HCI [121]. Even among this small team, the definition of SRs was a matter of debate. As well described by Martinic et al. [94]: "*A standard or consensus definition of an SR does not exist.*" Thus, one of the most important tasks in conducting and reporting SRs is to provide a specific definition of the review and its steps. In our case, we were able to shed some light on implicit and ambiguous definitions of SRs upheld in HCI and related fields throughout discussions among the team, and reach some consensus about identifying "*good*" qualities of conducting and reporting SRs in our field. These discussions and their results are reflected in our Guidelines section.

Finally, we note that—as primarily early-career researchers—we may be susceptible to being hesitant about being critical of the field, or about advocating an opinion when disagreeing with a senior researcher in the team. To counter this, the first author explicitly encouraged all researchers prior to resolvements meetings to feel free to voice and argue for their opinions about paper ratings. Further, we are ourselves embedded in the field: we have read other work by many of the authors of our review papers (including prominent researchers in the field), and in some cases have met them or worked with them. One author's own review paper was included in the final corpus. We mitigated bias by ensuring that each researcher did not rate or extract data from a paper that they might be in direct conflict with, yet cannot rule out unconscious biases.

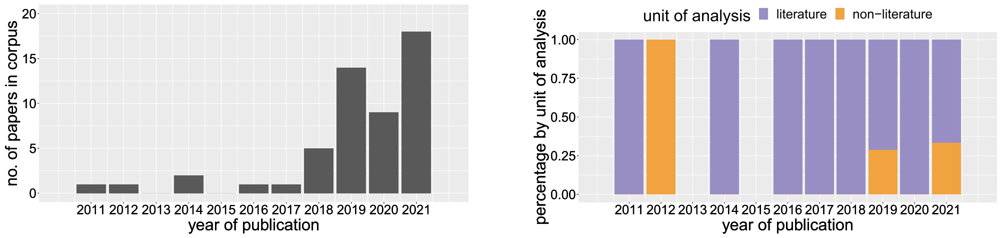
We provide an epistemic reflection in the discussion.

4 Results

Our initial differentiation was based on the papers' unit of analysis: 41 of the papers dealt with literature as their unit of analysis; 11 were non-literature review (see Figure 6). For reasons of scope, our results section focuses on the former: the 41 papers classified as an SR of *research papers* as their primary unit of analysis (rather than an SR of something else, e.g., apps). We structure this section into a general description of the corpus (including a categorization of review "archetypes"), our results of the score calculation (reporting quality grouped by review process domain), and finally a narrative synthesis of the applicability of the checklists. Please note that while the higher-level results are reported in this article, the full breadth and depth of our results is made transparent in the supplementary materials, including a large spreadsheet with all 63 checklist items (42 PRISMA and 21 ENTREQ), our interpretations of those items, and our rationale for the end scores.

4.1 General Summary of the Corpus

The majority of papers focus only on the SR as their main contribution (26 papers), while others frame the review as one key contribution of several (11). Four frame the SR as a minor contribution.



(a) Distribution plot: papers in each year

(b) Stacked percentage bar chart by unit of analysis

Fig. 6. Based on the initial corpus, i.e., before our focus narrowed to the review papers evaluating literature as their unit of analysis, this chronological overview of papers in the initial corpus on the left (a) shows a steep rise in SRs at CHI over recent years. Most of the papers review the literature, as opposed to non-literature, e.g., applications or systems (b).

Many papers cited a framework or existing methodology (15). Among these papers, PRISMA [108][24] was most popular (9; e.g., [54]). Five cited a framework deriving from software engineering literature (e.g., Brulé et al. [16] citing Kitchenham [71]); three review papers cited QUORUM [99] as their framework [8, 97, 114]. Another common approach among the corpus papers was citing prior examples of literature reviews (at CHI—e.g., [97]—and other venues—e.g., [62]; 14 papers), i.e., methodological shortcut citations [134]. Yet a similar number of papers cited neither frameworks nor examples for the purpose of situating their methodology (15; e.g., [19]). (This adds up to more than 41 because five papers followed a combined approach, i.e., citing multiple frameworks like Pater et al. [111], or frameworks and examples like Mekler et al. [97].) Regardless, citing frameworks or examples rarely coincided with details on how they were followed.

We provide an overview of basic metadata (the number of authors per paper, the range of years covered, as well as the size of the corpora and initial searches) in the appendix (Table 12).

4.1.1 *Review Archetypes.*

In our categorization of the SRs, we arrived at archetypes that distinguish how results were presented: categorical reporting, narrative summaries, and meta-analyses:[25]

Categorical Reporting. Most commonly (29), review papers in our corpus used some form of classification of elements found in the papers, and then reported frequencies (as counts or percentages presented in text, tables, or charts) for how often the elements occurred. This type of review methodology—together with rarely including an appraisal step—matches most closely with scoping review methodology [127]. This was often accompanied by narrative summaries and examples of the elements found (thus overlapping with the *narrative summaries* archetype below). Very rarely this review archetype also included statistical tests (e.g., reporting frequencies of online vs. remote study designs, and then comparing sample sizes between the two [19]).

The classification of elements was applied in various ways. Occasionally this was specified as quantitative content analysis [80, 149] (e.g., [57, 87]). In other cases, the classification was developed through qualitative analysis methods such as open coding [125] (e.g., [76]). However, in many cases, categorical reporting was conducted based on categories whose development was only vaguely described, including whether the categories were developed inductively or deductively.

[24] While we cite and use the recently updated 2020 PRISMA statement in this article [108], review papers of course mostly cited its previous version [100].

[25] Our overview in the supplementary materials shows as which archetype each paper was classified.

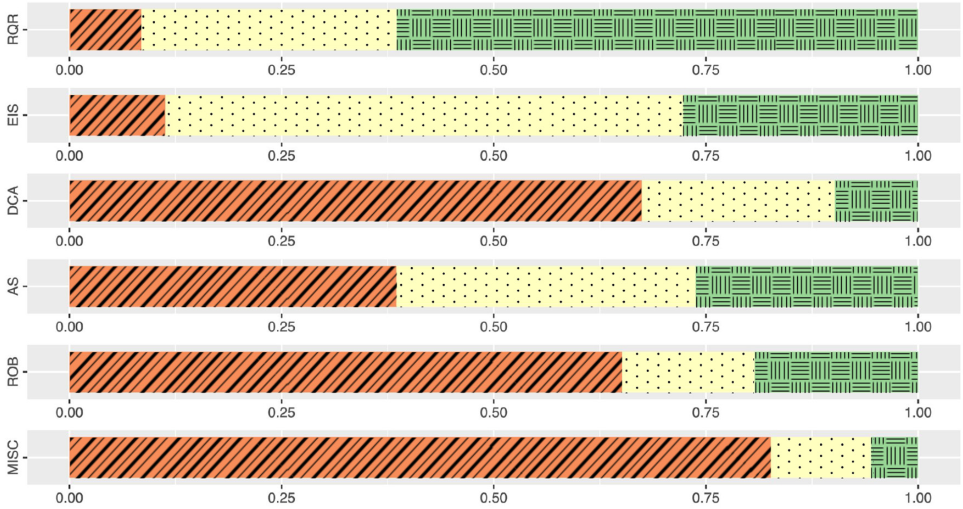


Fig. 7. By domain, percentage stacked bar chart of how many *yes* (green with cross-hatching)/*partial* (yellow with dotted hatching)/*no* (red with slashed hatching) scores the papers got for PRISMA and ENTREQ items categorized within that domain. Note that each row is based on a different number of items and scores (see Figure 8 for a non-normalized visualization), i.e., are based on a differing level of insight.

Narrative Summaries. This archetype (represented by 10 papers) primarily or only used narrative summaries to report their findings. Here, no quantitative measure is given of how often classified elements occurred; rather, reporting focuses on providing detailed descriptions of examples. Sometimes this resulted in the development of a taxonomy (e.g., [15]). The reporting methodology most closely matches conventional or narrative reviews [139].

Meta-Analysis. This review type was the most rare, with only two valid contenders [35, 137]. (Four PRISMA-checked papers showed indications they might contain a meta-analysis, however only two papers actually were considered to have conducted one.) These feature more targeted RQs and use primarily statistical methods to "combine and summarize the results of several primary studies" [27]. One of the corpus examples combined the meta-analysis with a deductive thematic analysis (which was reported in terms of frequencies of occurring codes).

4.2 Reporting Quality by Review Process Domain

In the following, we report the results of the PRISMA and ENTREQ checklists as categorized by review domain based on the mapping process we described in our Method section. We applied the PRISMA checklist to 24 review papers in our corpus, and the ENTREQ checklist to 35 review papers; 18 papers were thus assessed using both PRISMA and ENTREQ items.

We next briefly introduce each domain based on how the items were interpreted and applied—see Tables 3–11 in the appendix for a detailed overview of the original items in each domain, notes on how items were interpreted and applied, and the resulting scores per item. For readers wanting more details, we emphasize that we provide a comprehensive overview of scores with rationale for all papers and all items in the supplementary materials. In the following, we provide a brief summary of our results found in Table 1 (percentage of maximum possible score: $P + E_{total}$) and Figure 7 (normalized graph of score distribution per domain). For better distinction, we use the abbreviation 'ROBIS' when referring to the original ROBIS assessment tool and its review process domains as a whole, but ‘ROB’ when referring to the specific review process domain of the same name within the overall tool.

RQR. For this domain, PRISMA items expected[26] papers to provide a rationale for conducting their review that relates to existing knowledge, and to specify the objective or RQ they explore. ENTREQ only required the RQ.

As listed in Table 1, this domain yielded the highest score (~78% of the maximum possible), with by far the most *yes* scores. *Partial* scores resulted from papers not substantiating the “context of existing knowledge” to describe the rationale for the review, or not providing their RQ explicitly (instead it was inferred from their contribution statement, e.g., [1]).

Eligibility Criteria, Identification and Selection (EIS). In our application of the PRISMA items mapped to this domain, items required to specify the databases searched (and when), the exact search queries, IC/EC and procedures for screening (how many reviewers; how disagreements were resolved), the number of studies that were included vs. excluded, and examples of “*edgecase*” papers at the threshold of inclusion. For ENTREQ, descriptions of edgecases are not required. However, it adds an expectation for a rationale for the databases, numbers of studies with screened with reasons for exclusion, and an indication of whether papers followed “*comprehensive search strategies to seek all available studies*” or rather attempted “*to seek all available concepts until [...] theoretical saturation is achieved*” [142].

This domain was reported in the second most detail in the review papers (~58% of the maximum possible score, see Figure 7 and Table 1). Very commonly, the review papers did not report the date (or the exact date) on which the search was conducted (e.g., only a range: “September 2020” [60] or “Q1/2019” [76]). Often they did not describe the screening procedure in detail. Sometimes, this was because review papers focused their search on a specific time range (e.g., all papers published in specific conference proceedings, e.g., CHI 2014 [19]) and thus did not conduct a screening phase (although this was rarely stated explicitly). In many other cases, authors used passive voice, or phrased screening actions in terms of “*we*” (e.g., [152])—this left it unclear how many were involved in each stage, and how disagreements about paper inclusion were resolved.

In some cases, it was unclear how many papers were found in the initial search (overall, e.g., [126] or for each database, e.g., [137]). Search queries were also only rarely reported explicitly, and instead described only with key terms [18] or without specifics [137], or an example query was given for only one of the databases used [10]. Further, in a surprising number of cases (51.22%; contributing to P17 and E8), it was not possible to clearly determine which papers were included in the final corpus (although sometimes due to broken links to external sources [8]). Only PRISMA required edgecases (P16b); yet of the PRISMA-reviewed papers, only 7.6% discussed such edgecase papers with a specific example and rationale to outline the search space (e.g., [62]). Overall, the information for this domain was hard to locate and missing details, although most papers did contain at least some or most of the information expected.

Data Collection and Appraisal (DCA). For this domain, PRISMA items required details on the extraction/collection procedure (how data were extracted, how many reviewers collected data, whether they worked independently and so forth) as well as a list of all primary and secondary extracted data items. For appraisal, PRISMA items expected papers to clarify which methods were used to assess risk of bias or more generally quality assessment in the included studies. ENTREQ

[26] Note that our analysis is based on the requirements as stated in the PRISMA and ENTREQ item descriptions. Thus, we report what a correct use of the items would imply. We clarify that we do not expect the papers that did not *intend* to follow the items to follow them. Rather, we are interested in understanding how well these items are suited to HCI reviews when used without modification.

items expected papers to report details about appraisal (rationale, items, process, and results); and how data extraction was conducted.

As indicated in Table 1 and visualized in Figure 7, items in this domain were rated with *no* scores at a high rate (reaching only ~21% of the maximum possible score). While the papers in our corpus generally reported at least some information about the data items extracted and the data extraction procedure, appraisal was rare. In one case, the term appraisal was used to refer to eligibility screening [76]. Only three papers reported any kind of appraisal relating to risk of bias or quality assessment [16, 35, 137]. This significantly impacted the scoring results of this domain.

AS. The fifteen PRISMA items were strongly focused on reporting required from meta-analyses. They ask for effect measures for each outcome, information on processes used to decide which studies should be synthesized by which method, methods of preparing data for synthesis, graphical or tabular overviews of results, descriptions of the main method of synthesis along with a rationale, and methods and results of sensitivity analyses and investigations of heterogeneity. Finally, results as well as study characteristics and risk of bias need to be summarized for each synthesis, and also discussed in the context of prior work and what implications this holds. The ENTREQ items, expecting a more qualitative review, ask for characteristics of all included studies, but then focus more on procedural aspects: identification of synthesis methodology with rationale, software used, number of reviewers involved in coding and analysis, how comparisons within and between studies were constructed, whether themes were developed inductively or deductively, whether quotes are provided, and whether the synthesis offers "*rich, compelling and useful results that go beyond a summary*" [142].

While performing better than the DCA domain, overall the AS domain reached only ~45% of the maximum possible score. Expectations for meta-analysis results and methods were unsurprisingly not met by most papers (e.g., sensitivity analysis for P20d). Synthesis methods were rarely statistical: only five of the PRISMA-checked papers (P20b) fully reported statistical results; five more included some form of a statistical test but omitted details like effect sizes. Most commonly, they employed categorical reporting as percentages or absolute counts. Similarly, ENTREQ-scored papers—although containing qualitative syntheses—often did not clearly describe coding methodology and how themes (or theme-like constructs) were conceptualized and developed (E21). The main issue with corpus papers regarding their identification or description of the synthesis methodology was that they did not describe their rationale for the choice (48.7% for P13d; 97.1% for E2).

As mentioned, identification of corpus papers was not as prevalent as expected. Key characteristics of corpus papers (P17/P20a; E8) were similarly more sparsely provided than expected. Instead, characteristics were often discussed in aggregated terms in the corpus papers (e.g., [62]) or while examples are addressed comprehensively, there is no full overview of characteristics of individual corpus papers beyond citation metadata like the references of included papers (e.g., [1]).

The majority of PRISMA-checked papers provided an interpretation of their results that also referenced prior work (87.5% for P23a). This is at least similar to ENTREQ's item E21 (whether the review papers go "*beyond a summary*" [142]). Here, the score was lower, but 60% were rated as offering more than a summary: a framework or taxonomy (e.g., [15]), recommendations (e.g., [92]), or similar). Additionally, we note that several papers developed an interactive, "*living*" [2, 33, 34, 104] review application that can be used to explore the data [3, 18, 120, 129].

Risk of Bias in Synthesis (ROB). This domain was covered only by eight PRISMA items; ENTREQ does not require any estimation of risk of bias in the synthesis process. The PRISMA items cover methods and results for assessing risk of bias due to reporting bias and certainty in the evidence (including general quality assessment), with a focus on the synthesis as a whole, rather than the individual studies. Additionally, it expects statements regarding financial and non-financial support, competing interests of authors, and limitations of the evidence and the review processes.

Overall, this domain reached ~27% of the maximum possible score. Similarly to appraisal of individual studies, only few papers included methods for assessing reporting bias or quality assessment in general in the review syntheses [16, 35, 137] (P15; P22). Most PRISMA-coded papers featured statements of support in their acknowledgements, yet competing interests were very rarely addressed ([87] address their own bias as a limitation), although we note that for 16 papers the corpus contained work by the review authors themselves. Roughly 17% did not discuss limitations of the evidence, and only 37% addressed such limitations explicitly (P23b). Limitations of the review process (P23c) were discussed by 70% in some way—however, this was often very limited in nature: Only 25% addressed limitations of multiple stages of the review (e.g., search/identification and analysis/synthesis [87]). Often this only consisted of repeating the search parameters (e.g., having only reviewed papers in a specific time range or published in specific venues).

Uncategorized (MISC). This domain also only contained PRISMA items. The items require that SRs are identified as such in their title, that abstracts follow the separate PRISMA 2020 Abstracts checklist [106, 108], information on protocol development and registration, and availability of materials (e.g., data extraction forms, extracted data, and analytic code).

Overall, this domain was rated lowest in reporting quality at ~12% of the maximum score. The abstracts item (P2) was scored as "no" for all papers: following the PRISMA 2020 Abstracts checklist is simply not possible[27] within the 150-word limit for abstracts instituted by CHI. About one third of papers identified themselves as an SR in the title, and another 20% did so in the abstract. For 11 papers, neither title nor abstract identified the paper as an SR. None of the papers addressed registration, nor specified whether an *a priori* protocol had been developed. A few papers of course referenced the PRISMA protocol (e.g., [92]) but we distinguish between the use of guidelines like PRISMA and the development of an *a priori* protocol *specific to the review at hand.* Hansson et al. [54] present a section titled "Protocol and registration" yet it only references PRISMA—no *a priori* protocol specific to their review seems to have been created or registered. We note that we do not doubt that many authors will have prepared a review plan of some sort prior to undertaking the review, however it was not documented as such in the papers nor was its documentation provided in external sources. For amendments to the review protocol, we thus scored non-explicit examples with partial, of which there were two examples: MacArthur et al. [92] describe a scoping phase in which they adjusted their categorization scheme; Pettersson et al. [114] mention "sharpen[ing] selection criteria" based on cross-checks of the initial papers.

Finally, none of the papers provided all the data expected by PRISMA (P27). Seven papers provided some of the listed types of materials, two papers showed intent to provide materials, but access did not work ([10, 59]), and one stated they would provide materials on request [62].

4.3 Applicability of the Checklists

We here provide high-level narrative summaries; more detailed explanations of which aspects were debated and how items were interpreted in application is reported in Tables 3–11 in the appendix, with further details provided in the supplementary materials. For notes on how the checklists differ for each domain, we refer readers to the supplementary materials as well.

[27] For reference, the checklist constitutes 12 items that expect abstracts to identify the paper an SR, as well as detail an explicit objective, eligibility criteria, information sources, methods for risk of bias and synthesis, the total number of and key characteristics of included studies, results for main outcomes, limitations of evidence, an interpretation of results with implications, primary funding sources, and registration details.

Applicability of Items Mapped to the RQR Domain. These items generally worked well. We had to decide what counts as "context of existing knowledge" and how to rate implicit RQs/objectives (e.g., a *post hoc* contribution statement, rather than an a priori question/objective). Additionally, ENTREQ requires a clear RQ (not allowing objectives). However none of the items had to be re-interpreted or modified.

Applicability of Items Mapped to the EIS Domain. PRISMA items were modified in terms of secondary requirements (e.g., item P5's "how studies were grouped for the syntheses" is not explicitly addressed by most papers and would distract from the item's primary requirement of IC/EC) and minimum requirements (how many edgecase examples are required in P16b). We debated what counts as a full date of search or search strategy, how to deal with cases without clearly defined screening stages (i.e., targeted proceedings searches), how granularly to report with multiple database sources, and distinction between edgecase examples and exclusion examples; these aspects are not specified by the checklist or the elaboration paper [109].

For ENTREQ, we modified items to clarify interpretation (e.g., requiring a full date of search to match P6), and establish default expectations. We also discussed search strategy classification.

Applicability of Items Mapped to the DCA Domain. All PRISMA items had to be modified or specified further to establish clear scoring criteria (e.g., what counts as an "outcome" variable for P10a). We re-interpreted risk of bias (P11, P18) to also accept quality assessment more generally. We also note that some PRISMA items referencing risk of bias must be considered on a study level (affecting study findings) and others on a synthesis level (affecting findings across studies). The application was further complicated by the many different ways authors describe data collection/extraction: this process overlapped with coding, analysis, screening, and categorization.

ENTREQ items for this domain largely worked well although we had to specify what qualifies as describing an appraisal "approach," and establish default expectations. Here also, data extraction/collection often overlapped with coding, annotation, generation and analysis of data.

Applicability of Items Mapped to the AS Domain. This domain required a lot of modification and re-interpretation, as for example, CHI review papers do not tend to clearly specify what counts as a synthesis (P13a), or describe "methods used to tabulate or visually display results" (P13c). Further we had to decide what qualifies as "characteristics" of a study (P17), when "results of statistical syntheses" are complete (P20b), and whether to also look in supplementary materials. We largely considered AS as synonymous for the application of these items, although we note that these are not necessarily the same thing. We extensively discussed, for example, what level of detail should be required for describing synthesis methods and the rationale behind their choice.

ENTREQ items similarly required a lot of discussion and interpretation. We again treated AS as synonymous, and matched PRISMA in several aspects (e.g., in our interpretation of "characteristics" of studies for E8). There were extensive discussions of what to require for many items, e.g., what qualifies as a full description of coding (E17) or theme derivation (E19). The largest point of contention was when results "*go beyond a summary*" [142]. This very subjective criteria is difficult to apply with consistent scoring, and further complicated by vague descriptions of how taxonomies or frameworks were developed based on review results.

Applicability of Items Mapped to the ROB Domain. This domain consisted of only PRISMA items, as ENTREQ does not cover it. Almost all PRISMA items were modified or re-interpreted, for example to specify what qualifies as discussing limitations (P23b and c). Risk of bias items were interpreted broadly to include quality assessment as well, unless the item descriptions specified a type of risk of bias.

Applicability of Items Mapped to the MISC Domain. This domain again was covered only by PRISMA items, and largely referenced expectations for the title and abstract, as well as protocol registration and development and availability of data. We note that the abstract criteria (P2) cannot be fulfilled with CHI abstract length restrictions. Additionally, some reviews appear to have a misconception about what protocols mean in the context of SRs (referencing guidelines like PRISMA as opposed to the development of an a priori documentation specific to the review at hand). We did not adjust interpretation for this.

5 Discussion

Our approach applied checklists rooted in largely (post-)positivist epistemology to our corpus of SRs at CHI. While we did so in a pragmatic/critical realist fashion, our first point of discussion must thus be what we can say about the reporting quality *based on those checklists*, i.e., relating to reporting standards with post-positivist origins.

We first summarize the "shape" of SRs at CHI, and then discuss key opportunities to improve reporting quality across the review process domains based on (our interpretation of) the checklists we applied, and what this implies about the term "systematic." We then address limitations of the evidence, our methodology—including issues of generalizability for HCI more broadly—, and the underlying database infrastructure. Finally, we discuss the use of these checklists for tertiary research, and consider epistemological tensions.

5.1 The Shape of SRs at CHI

Overall, the majority of papers featured characteristics most closely aligned with scoping reviews or mapping reviews [4, 116]. Reviews at CHI do well at presenting their RQ(s), but they tend to be phrased broadly rather than focusing on specific effects. Search process details were generally reported but often missing details expected by the checklists (e.g., the exact queries used in each database). There often was a screening/selection process from initial results to the final corpus, but this process was heterogeneous in nature and not always described in enough detail to be able to repeat it. Further, there seems to be a subcategory of "targeted proceedings searches": reviews that do *not* involve a search and screening process, but instead directly select specific proceedings as the corpus for analysis. Further, search and selection results—i.e., the papers included in both initial and final corpus—were often not fully reported. While we do not believe that omitting some relevant papers will necessarily render an SR useless, we do believe that it matters *which* papers are included or not—emphasizing the importance of stating which papers were selected. Data extraction was most often not described as a process; instead, papers offered the resulting spreadsheet, or parts of the process were implied by the resulting reporting. The corpus reviews only rarely feature anything resembling an appraisal stage. Analysis was primarily based on some form of classification, and categorical reporting of results, which does not match the expectations laid out by PRISMA or ENTREQ. Whether this is something our community should aim for, however, is a different matter.

5.2 Where We Can Improve Reporting Quality

Leaving aside whether the characteristics described in the previous section match the term "systematic" for now, our findings reveal significant areas of concern in the reporting quality in CHI review papers based on the PRISMA and ENTREQ reporting guidelines. Particularly, the domains of *data collection and appraisal*, *ROBIS*, and *miscellaneous* had high ratings of insufficient reporting. To be more precise, as a research community, we seem to substantially under-report appraisal, risk of bias and limitations in synthesis, and documentation in the form of protocol development and registration. In fact, based on the reporting, we barely seem to be conducting appraisal and are not developing *a priori* protocols at all. Further, while *AS* and *eligibility, identification, and search* had fewer "no" ratings than the previously mentioned domains, here too, many scores indicated at least incomplete information (~72% with *partial* or *no* scores for EIS, and ~74% for AS).

We assume that in many cases, authors did *conduct* the steps expected by the checklists but do not report every detail. Certainly, we do not mean to imply that these review papers are of poor quality. However, incomplete reporting crucially hampers our field: it means that many of our reviews cannot be replicated and updated, sets an imperfect standard for future reviews, and reveals a high potential for overlooking nuances and key findings, making it more difficult to judge and rely on these reviews as a reader and researcher. This lack of transparency is not unique to our corpus; it is a common issue in SRs also in other fields [49]. Haddaway and Macura [51] have presented a summary of major types of bias that can occur in SRs. Their list is extensive and covers each of the different phases of a review: describing limitation through subject drift, and biases relating to selection, databases, publication, and interpretation, among others. Many of these biases are clearly at least potentially present in CHI SRs, but usually not acknowledged or discussed as a limitation. We acknowledge that for some types of research and knowledge, bias is not a concern, but rather viewed as a strength. However, to our understanding, these types of research greatly value critical in-depth engagement and reflexivity, and so would also benefit from documenting and explicitly addressing aspects viewed as "bias" by other epistemic perspectives.

In the following we highlight key aspects that should be improved in CHI SRs if we aim to better meet the reporting standards expected from the perspective of the PRISMA and ENTREQ checklists:

Single vs. Double Screening/Extraction. Only very few review papers reported double screening or double extraction. In many cases, it was unclear how many researchers were involved in these steps. When it was made clear, the research team often split the data among each other, but each record was only assigned to a singular researcher to assess and extract. Doubling this procedure—and resolving disagreements—is time consuming, but reduces error and implicit bias [49].

Lack of Appraisal. The corpus papers barely ever conducted an appraisal of the included papers and studies. While of course many papers do critically engage with the papers in the discussion, there is for the most part no systematic and transparent critical engagement with the included work. This is concerning as Haddaway and Macura [51] list five different types of bias as potentially resulting from this stage alone: outcome reporting bias, full publication bias, multiple publication bias, funding bias, and lack of consistency regarding validity. We again note that this is a common issue in other fields [49]. From our own discussions, we believe this may stem from a lack of awareness about which guidance to use for quality appraisal, a hesitance to call out other researchers' work for holding potential issues or implicit bias, as well as the fraught nature of the term "bias" within different epistemologies among HCI researchers. We believe that the HCI community needs to grapple with its understanding of terms like bias, credibility, validity and trustworthiness to be able to perform critical appraisal in knowledge synthesis, echoing similar thoughts in other fields on evaluating and integrating research from different epistemic backgrounds [46, 64, 79, 136].

Lack of Formality in Synthesis. In many review papers, the synthesis method (items P13d, E2) was often only roughly described or identified, and rarely rationalized. We believe that researchers may be struggling with the lack of applicability of conventional statistical methods for SRs (e.g., meta-analysis) in HCI and as-of-yet unclear guidelines on how to synthesize mixed-methods results. It is worth emphasizing that there are a lot of formal synthesis methods developed specifically for SRs as well as scoping reviews, that authors may be interested in exploring. We provide references as a starting point for such an exploration in the next section.

Lack of Limitations. Limitations were often addressed only briefly and omitted large parts of the review process (e.g., focusing only how papers were selected for the review, but not how they were analyzed). Perhaps this is an issue stemming from page limits in earlier review papers, and ongoing expectations to keep papers concise, which can lead authors to skirt best practices. However, we were surprised when papers with a qualitative synthesis did not provide any reflections on the limitations of their method, as with this kind of methodology, reflexivity is generally considered important [96]. ENTREQ itself does not require anything like this either.

Documentation: A Priori Protocol and Registration. Not a single review paper reported developing a protocol for their review before conducting it, or provided it for review. We of course assume that researchers did plan their approach, and likely documented it in some form, as also suggested by the two mentions of changes to planned approaches [92, 114]. However, without providing clear documentation of the initial plan, the potential for subject drift in the corpus is high [51], i.e., that reviews' RQs could have changed over time, even after researchers began their analysis phase. This risks drifting to a RQ or method that the search phase and the corpus is no longer suited for.

5.3 The Term "Systematic"

With the rarity of appraisal and protocol usage, most corpus papers would probably be described by many as scoping reviews—with their eligibility for the term "systematic" in question depending on one's definition. We do not wish to prescribe an interpretation of when a review is systematic to the field of HCI. Our field contains far too many research disciplines and methodologies to be able to decide that based on this single exploration of a flagship conference. We also emphasize that non-SRs can certainly also be useful and impactful. However, we believe that, as a community, we need to be more clear about why we use this term when we apply it. We also believe that our field would benefit from more attention to appraisal and protocol development. We should be more careful when we use the term systematic in the context of reviews—in particular, when we mean that only *specific stages* of the review follow a systematic approach, or that specific stages were conducted methodically rather than *ad hoc* without relating to SR methodology. This again emphasizes the importance of transparent reporting. Inversely, when we do use systematic approaches, we should denote this clearly—so that it can be found by future tertiary research, and early-stage researchers and students can discern and learn from best practices and rationales.

We strongly encourage authors of reviews to provide a methodological positioning of their type of review, e.g., based on Sutton et al. [139]'s review family classification, and to consider clearly specifying what they understand to be "systematic." There is no "true" agreed-upon definition of what makes an SR "systematic." However, we wish to position our own understanding of the term after conducting this project: our personal working definition views reviews as systematic when they follow all steps outlined in the background section (Figure 2) along with the suggested best practices in the upcoming section—*as long as*, in each deviation from this, they provide a suitable rationale for it. Time will tell whether this suggested understanding is adopted by the HCI field or not, or whether it will require stricter or looser interpretation in the long term. In particular, it remains to be seen how this view will be interpreted by other epistemic approaches or used as a reference point for departure to better reflect different epistemological perspectives.

5.4 Limitations of the Umbrella Review

Limitations of the Evidence. For limitations of the evidence, we largely refer to our results and discussion, as appraisal was central to our RQ. We focus here instead on aspects beyond those already addressed. We note that our appraisal was conducted based on authors' self-reports and our interpretation thereof. It was not always clear what exactly had been done. While we did try to give review papers the benefit of doubt when aspects were ambiguously reported, and conducted double extraction and refinement stages, it is possible that some review papers received low scores for specific steps despite having conducted them well, simply because they did not report this or we misunderstood or could not find specific aspects of the description. Further, our sample spans across years in which CHI upheld a page limit and years in which this was not the case. This shift likely affected review papers' capacity for rigorous reporting.

Our corpus provides us with a good overview of SRs and reviews that claim a systematic process at CHI pre-2021. However, we note that with our search keywords and IC, the evidence we collected excludes papers that arguably also contain SRs in content when they did not use the term "systematic." In our review process, we noticed at least two such examples [90, 117]. Additionally, as demonstrated by our manual-addition paper [8], the ACM DL is imperfect, and so we may still have missed other examples of SRs that *do* match our IC. As all reviews, our replicability is limited by the reliability of the search database, yet our comprehensive reporting should enable updates of this review in the future.

Further, while we are using this corpus to explore SRs, not all of the papers in our corpus are necessarily actually SRs, as this depends on one's definition of the term. Our corpus in large parts consists of papers that declare themselves to be SRs, but also contains review papers that merely claimed (sometimes as an offhand sentence) to review papers systematically without or with only little description of the system they used or what makes it systematic. They may have used the term colloquially, and so—as also the case in prior reviews, such as [135]—the work may be compared to criteria it did not aim to meet. We provide a full list of each paper's reason for inclusion in the supplementary materials to showcase this transparently. We would like to emphasize that the inclusion of papers with ambiguous goals towards "systematicness" helps to better pinpoint what makes reviews "systematic" in HCI, and to form an understanding of the concept based on which we develop guiding questions for authors of future SRs in the next section.

Limitations of the Methodology. Our AS are based on CHI papers—which, as mentioned, was a purposeful yet also pragmatic limitation. As the flagship conference of HCI, CHI fills a prescriptive role of sorts, as also suggested by the number of SRs in our corpus that referenced other SRs at CHI in their rationale or as the basis for their method. We thus suspect that authors targeting other HCI venues with their SRs may also use CHI SRs as examples to follow—in addition to examples from those specific venues. Thus, SRs published at other HCI venues may well show similar features and limitations. We believe that the strong interdisciplinarity of HCI [123] leads to a similar variety in SR methodology in other venues as at CHI. Further, the formal requirements and expectations may be similar in some ways: other HCI conferences also tend to have abstract limitations and so cannot fulfill PRISMA's expectations for an extended abstract [106]. However, other aspects may well differ between subfields and venues: the greater tolerance for longer manuscripts at HCI journals may lead to more detailed reporting. Further, the emphasis on presenting a contribution statement—that was sometimes used in lieu of an explicit RQ in corpus papers—may be less pronounced in other HCI venues.

Answering definitively how our results generalize to HCI is beyond the scope of this paper, and will of course have to be explored in more detail in future research. Nevertheless, our method of quantizing scores based on our interpretation of PRISMA and ENTREQ to compare review process domains based on ROBIS [150] is a clearly documented first approach towards allowing future research to compare SR reporting across different venues and subfields. By switching out the checklists for other checklists or for alternative means of investigating reporting quality (or related concepts like trustworthiness), authors of future umbrella reviews can use this paper as either a template or a jumping-off point to reflect on how a similar investigation could be carried out based on a different epistemology.

One key limitation of our search methodology is our use of a single database, namely the ACM DL. Manual screening of proceedings was also considered, but discarded because it would have not caught papers that do not make clear they contain an SR in the title or abstract. We considered using Scopus as an additional database that covers CHI proceedings, however this would not have allowed us to filter results from CHI by their research article status. The Scopus search thus yielded a number of results too high for us to screen at the time while upholding our double screening with tie-breakers procedure. Further, we restricted our search to a single conference, and used this as a representative sample of HCI publications at large. A broader search was not possible with our available resources, however, we note that future work will have to explore how generalizable findings truly are for SRs at other HCI publication venues, e.g., journals like ACM CSUR or TOCHI, but also smaller, more specialized conferences like ACM UIST or IEEE VR.

Further, we note with the contested nature of the term "SR" come many potential synonyms and related terms like "survey" or "meta review." We deliberately did not include these, as these are not established terms for SRs.[28] Nevertheless, it is certainly possible that some papers using these terms are also in fact systematic (despite not using the term).

Our synthesis methods—categorization of review archetypes, our domain-based mapping and checklist application to generate a quantitative score for reporting quality, and narrative summaries of checklist applicability—are informal. This is partially due to the unusual nature of our umbrella review in exploring SRs on different topics, rather than the same topic. The application of PRISMA and ENTREQ items and our mapping of these PRISMA and ENTREQ items to Whiting et al. [150]'s ROBIS domains is of course subjective to the interpretation, ratings, and discussions of this review team. The two checklists also differ (e.g., ENTREQ does not require discussion of limitations but does expect a rationale for the database choice; PRISMA requires edgecase examples). Moreover, the final refinement process in which the items were interpreted and sometimes modified—although also involving discussions between multiple authors—was carried out primarily by the first author, which introduces further subjectivity.

Further, as noted above, both checklists are intended for SRs (and two specific types of SRs at that). Yet our corpus could—depending on one's definition of systematic—be described as primarily scoping reviews and/or non-systematic, and containing only a few reviews that match the specific types of reviews these checklists were intended for. Thus the scores we report must be assessed in light of that. Additionally, neither checklist was intended as an evaluation tool, so issues with ambiguous phrasing complicated the scoring process. Alternatives do exist, for example, we considered the AMSTAR 2 [131], which was developed for critical appraisal of SRs. However, this strictly targets reviews of interventional studies, relies on the PICO framework, and does not cater to qualitative synthesis. We attempted to mitigate issues relating to checklist phrasing with our score discussion and refinement processes, as well as our extensive appendix and supplementary materials, however, this is of course also subjective.

Finally, we acknowledge that some researchers may disagree with our subscription to the term "umbrella review" over the term "SR of SRs." Depending on the definition of these terms, we could be one or the other, both, or neither. However, our unit of analysis is SRs, and we synthesize disparate, heterogeneous SRs under the umbrella of analyzing reporting quality; the lens by which we analyze the papers is the same. Thus, the term umbrella review seemed the most applicable to us, despite it being an unusual case.

[28] For example, "meta review" is used as a synonym for SR [111], but also for umbrella reviews [24], as well as the summarizing review of conference reviews [12].

Limitations of the Database Infrastructure. Underlying all reviews is the database and search engine infrastructure that we use to locate potentially eligible research. Yet there are significant issues with these databases. In our own experience with Scopus and the ACM DL, search results can fluctuate wildly over time—and not just with the inclusion of new publications. Issues of this kind were also reported by review papers in our corpus [54, 92], as was uncertainty over the specification of ACM DL search fields (like "full-text" vs. "any field" [76]). We can report that for our own pilot searches, some of the fluctuations resulted from duplications of entries. In contrast, the paper that we had to add manually was missing despite its clear eligibility based on the keywords we employed. Despite multiple emails to the ACM DL team, we have been unable to get a clear answer as to why the search results vary so much over time. Further, when asked for documentation on the specific search-within options, they only referred to very rudimentary tutorials,[29] that do not specify necessary details. This would be useful, however, to explain why the Abstract option does not always only match keywords on the abstract, and to ensure comparable searches across different databases (e.g., to match Scopus's TITLE-ABS-KEY).

In exploring this further, we found that academic databases are simply not nearly as reliable as we had assumed. For example, there are many reports of documents not being indexed by a specific database, despite the database claiming to index the corresponding journal [13, 78]: "*even if two databases index the same journal the coverage of articles might differ*" [13]. Additionally, meta data and bibliometric data can be incorrect [38, 77]. Assessing Scopus and Web of Science as indexing databases in comparison to information reported by a specific journal's publisher, Liu et al. [89] identified many discrepancies. They trace these to differences in policies relating to publication date versions, missing documents, duplicate documents, and metadata errors. Given the importance laid on replicability of reviews (for post-positivist or pragmatic approaches), this is a concern that needs to be addressed by the research community in the future. More immediately, papers should not be rejected just because the query no longer results in the same number of results: we interpret replicability to be fulfilled when the reporting enables readers and reviewers to perform the search confidently—rather than when it results in the exact same number of papers.

5.5 PRISMA and ENTREQ as Reporting Checklists for Tertiary Research

With categorical reporting and narrative summaries as the most common style of reporting, it is unsurprising that the corpus scored poorly on many items of PRISMA (which is purely quantitative and focuses on meta-analyses), as well as ENTREQ (which expects clear descriptions of coding and theme development). As shown in Tables 3–11, almost all checklist items required some form of discussion and/or reinterpretation before they could be applied to the review papers. We discussed what this means for our own results in the Limitations section. Nevertheless, this leads us to two questions that will need to be answered by the HCI community going forward: (1) does this indicate a need to develop our own reporting guidelines for the types of papers published in our field as primary research? This could then better embrace and reflect reporting expectations depending on specific research approaches in HCI. Or (2) do we need to adjust how we report SRs to better fit existing guidelines?

The latter could imply a methodological move towards more clearly qualitative or clearly quantitative synthesis using formal methods. Reflecting on HCI as an interdisciplinary field that contains a multitude of methodologies and epistemologic backgrounds, we do not think that this is advisable as a general rule. Further, other fields have similarly seen an evolution of quantitative SRs to integrate qualitative methods [53], suggesting that a full exploration of a RQ in many cases will require mixing of quantitative and qualitative synthesis. As mentioned, our own approach lies within the

[29] https://libraries.acm.org/training-resources/new-dl-features

backdrop of pragmatism and critical realism; we view the mixing of quantitative and qualitative methods as having the potential for triangulation but also deeper understanding and richer insight [95]. Yet of course when handling research suitable to PRISMA or ENTREQ, these checklists may still be useful and applicable. As such, we think the answer is “yes, both”—but with an emphasis on the former: the PRISMA and ENTREQ checklists are only fully suitable when the reviews are either clearly quantitative with statistical methods and interventional studies or qualitative with theme development, respectively. In the long term, we may indeed need to develop new checklists or other forms of guidance that are better suited to HCI and its plurality. However, we also need to adjust how we conduct and report SRs to improve transparency and trustworthiness and either reduce or better capture and document bias (depending on one’s goal and epistemology). In tertiary research, existing checklists can still point out such issues even when they do not fully apply. Thus using these and other checklists or guidance instead in the meantime (perhaps those for scoping reviews, e.g., [29, 112] or [6, Ch.11]) can be useful for secondary research reporting, and analysis of reporting quality through tertiary research like this umbrella review—if, as we aim to do here, they are applied with flexibility and reflection instead of as a rigid ideal.

5.6 Epistemological Tensions

Most corpus papers were situated somewhere between clearly quantitative and clearly qualitative approaches, and unearth long-standing [39] epistemological tension in the field. For some epistemic persuasions, a review replication (cf. [70, 102]) should be able to yield the same results—albeit still potentially differing in interpretation. For others, any replication will necessarily be contingent on and thoroughly shaped by the researchers’ construction of the analysis (e.g., qualitative themes). For the latter, practices of reflexivity [96, 140]—critically reflecting on the researchers’ role in the synthesis, i.e., their “footprint” on the findings—can be exceedingly helpful. We argue that this could improve transparency in the interpretative steps of more post-positivist-leaning reviews, as well. Yet such reflections did not occur in the corpus papers. Additionally, clear communication of the authors’ goals and epistemological understanding should help to alleviate misunderstandings. We advocate a degree of tolerance towards other epistemological viewpoints to ease these tensions further. As part of this, we provide an epistemic reflection on our approach in the following.

Epistemic Reflection. Our approach applied checklists rooted in largely (post-)positivist epistemology [46] to the corpus of SRs at CHI in a fashion best described as pragmatic [68] (valuing the role of context and ethical considerations; a prioritization of flexibility in adapting methods for problem-solving) or critical realist [95] (valuing epistemic relativism, contextual embeddedness, a critical stance, and reflexivity). We do not see them as checklists in the sense that we aim to prescribe or even recommend their usage. Rather, we use them as a starting point and scaffolding component for our investigation into the types of reviews at CHI claiming a systematic approach. In line with our non-positivist approach, we transparently and comprehensively reflect on how we applied and understood the checklist items as part of our reporting. Of course this does mean that the SRs in our corpus are analyzed in light of their compliance with guidelines that the papers potentially do not profess to follow, and may not even match their specific variant of SR. However, this is a necessary approach when the definition of SRs is not clearly established in our field, and one with precedence [135]. Following a pragmatist/critical realist approach, this still yields (and structures) valuable insight into the kind of information being reported about SRs. On top of that, our process of using the checklists allowed us to also reflect on the checklists themselves: whether the items are clear, applicable, and most of all useful for the context of HCI.

Further, in line with our pragmatic/critical realist stance, we made use of a mix of steps depending on what we considered most reasonable at that point: our steps can be considered well aligned with (post-)positivism just as much as pragmatism in the initial stages of the review (e.g., search and data collection). We think this can lead to a more comprehensive initial corpus in reviews, and do not see this as incompatible with more constructivist-leaning approaches to the analysis, as it allows a broader and more comparable starting point for AS of any epistemic background. In later stages (after corpus selection), we favored steps that are in our view clearly non-positivist, for example in the way that we applied mixed methods for AS, and our commitment to reflection and critical discussion among the authors to capture our own role in shaping knowledge. Yet a fully constructivist approach would likely not use checklists in combination with scoring; this again reflects tensions entrenched in HCI.

6 Guiding Questions and Suggestions for Best Practice

We now move away from the direct scope of our findings to attempt to develop guidance for the reporting of SRs within HCI. Our approach here aims for the pragmatic goal of creating something useful for the HCI community as a whole, or barring that, for large parts of it. As such, we purposefully do not create a new checklist, but rather guiding questions and suggestions for what we view as best practice in reporting. The guiding questions are intended as prompts for reflection, and so can be applied to all SRs–no matter one's epistemic persuasion—as they are not meant prescriptively. There may well be good reasons to not follow the developed set of guiding questions, however these reasons should be clearly articulated (e.g., by stating that a certain step is incompatible with the authors' epistemology). The guiding questions should not be viewed as the be-all-end-all of quality review reporting. They reflect how much detail is reported, because without reporting, quality of *content* cannot be assessed. While we believe that checklists may be limited in use when assessing quality of SR *conduct* (unless very carefully formulated, and applied both very flexibly and in a targeted manner), we are confident that checklists are an effective means of assessing reporting quality of literature reviews—as long as they are applied flexibly to allow careful, reflective interpretation of items.

The suggestions for best practices again return to a pragmatic/critical realist focus, and as such need to be understood as suggestions for only specific kinds of SRs, namely those at least compatible with this perspective.

6.1 Domain-Based Guiding Questions for Reporting Quality in SRs

We developed a set of guiding questions for improving reporting quality in SRs. This was based primarily on our experience with applying the PRISMA and ENTREQ checklists, and specifically our note-taking and synthesis check-in meetings. However, it was also informed by our combined experience with writing and reviewing review papers of various kinds and knowledge of the literature on review methodologies. Using the ENTREQ as a starting point (due to its fewer items and comparative applicability), we used our card sorting by review domain and re-visited each item to then either include and adapt, or exclude. The first author developed a first draft and then presented it for discussion and iteration among the rest of the research team.

The resulting version of the guiding questions for reporting is presented in Table 2. For example, it prompts researchers to clearly report the driving RQ, without which readers may not be able to judge whether the search and screening methods were appropriate; a mismatch between them can impact validity and utility of the SR. The guiding questions also prompt researchers to report the initial and final corpus papers, which makes it easier to follow what the analysis was conducted on, and whether (or why) key papers may have been omitted.

Although the questions are meant as prompts for SRs of all kinds, our development of them was based primarily on the types of SRs found most commonly in our corpus: ones that develop and/or

Table 2. This List of Guiding Questions Separated by Review Domain Can Help Authors Improve Their Reporting Quality

Domain and Subdomain		Guiding Questions for Reporting: Does the paper (or appendix or supplementary material) ...
RQR	Research Question	... state the research question? ... describe a rationale for conducting the review? ... consider the context of existing knowledge in that rationale?
EIS	Search	... specify all databases that were searched? ... provide a rationale for the choice of databases? ... specify when the search was conducted? ... present the exact search query used for each of the databases? ... report all the search filters and limits that were used for each search query / database? ... identify how many and which papers were identified through the search as the initial corpus to undergo the screening process? ... report how many papers were excluded and for what reasons?
	Screening	... report a screening stage, or a rationale if not? ... state the inclusion / exclusion criteria used for screening? ... report the process of screening in terms of which parts of the paper were considered (e.g., title and abstract)? ... report the process of screening in terms of how many reviewers were involved? ... report the process of screening in terms of how disagreements between reviewers were resolved? ... report which papers were left eligible after screening?
DCA	Appraisal	... report an appraisal stage, or a rationale if not? ... describe what was appraised (e.g., assessment of study validity vs. reporting)? ... provide a rationale for the appraisal? ... identify which tools or criteria were used for the appraisal? ... describe the appraisal process in terms of how many reviewers were involved? ... describe the appraisal process in terms of how disagreements between reviewers were involved? ... report the results of the appraisal for each paper? ... report the results of the appraisal in summarized form? ... describe how the appraisal results impacted the review (e.g., exclusion or weighing of evidence in synthesis)? ... give a rationale?
	Data Collection	... report which papers were in the final corpus for data collection/extraction? ... indicate which section(s) of the paper was considered for data extraction? ... report which data was extracted from the paper? ... describe the extraction process in terms of how many reviewers were involved? ... indicate in what form extracted data was stored?
AS	Method	... identify and describe a formal synthesis methodology that was used, or describe the methodology, if not? ... describe a rationale for the choice of synthesis method? ... describe which software was used to conduct the analysis? ... identify how many researchers were involved in the coding of data, if applicable? ... describe how disagreements in coding were resolved? ... identify how many researchers were involved in the synthesis? ... describe how data was coded in terms of how the coding scheme was derived? ... specify whether this coding scheme was inductive or deductive? ... describe how data was coded in terms of how the coding scheme was applied? ... describe how researchers dealt with newly arising concepts outside of the initial coding scheme? ... specify whether themes were constructed from the coding and analysis of data, and if so, whether these themes were inductive or deductive?
	Results	... provide a full list of included papers? ... include basic characteristics such as year of publication, and extracted data items of interest (e.g., study design) in that list?
	Discussion	... not just present results as a summary, but also interpret results in the context of existing evidence?
ROB	Risk of Implicit Bias in Overall Synthesis	... report any methods used to assess risk of hidden or implicit bias or certainty in the evidence? ... address reflexivity: how did personal, methodological, contextual or other factors [64] impact the synthesis (e.g., if low-quality studies were weighed improperly in the synthesis or if researchers' background affected interpretation of results)? ... discuss limitations of the papers included in the review (e.g., due to sample size)? ... discuss limitations of the review process in terms of the search and screening process? ... discuss limitations of the review process in terms of the appraisal and extraction process? ... discuss limitations of the review process in terms of the synthesis process?
MISC	Documentation: Protocol and Registration	... identify the paper as an SR in the title or abstract? ... report whether a review protocol was developed beforehand? ... make that protocol available? ... clarify any adjustments made to the protocol, if applicable? ... make available any data like data extraction forms and the extracted data?

For better overview, we split some review domains into sub-domains, e.g., DCA separately considers appraisal and data collection.

apply a coding scheme or classification, and then develop and report themes or frequencies of category occurrence. From a pragmatic/critical realist perspective, we would thus use the guiding questions in addition to existing checklists for reporting, if they apply. For example, for reviews that report a meta-analysis, we recommend the additional use of PRISMA [108], or the ROSES alternative [52] (which improves on a number of issues of PRISMA, but which is longer, and was developed for the environmental science field). Similarly, for review papers using thematic analysis as their synthesis methodology, we strongly recommend that authors use checklists for thematic analysis reporting or discuss concepts like trustworthiness in addition to these guiding questions.

The questions are meant primarily to aid researchers who are conducting an SR in HCI and wish to ensure high reporting quality. Following these guiding questions will not ensure that a review is systematic, or even necessarily good quality overall. This is partially because there is currently no agreed-upon definition of what "systematic" means for reviews in HCI. Answering the questions (or stating why they are not applicable) instead should guide authors towards providing enough information in their review papers that their readers have a chance to make an *informed choice* about whether the review could count as systematic (depending on their understanding of the term) or could count as high quality. Which practices and steps are necessary for a review to count as systematic remains a big question for the field to decide on. Which practices and steps should constitute high quality remains a bigger one.

6.2 Suggestions for (Pragmatic/Critical Realist) Best Practices

Based on the existing literature on SR methodology, and our reflection on the analysis of our corpus with PRISMA and ENTREQ, and our own experience with conducting, reporting and reviewing reviews, we offer the following suggestions for best practices. As stated before, these should be understood as suggestions for a pragmatic/critical realist approach and may not be applicable to others (e.g., constructive/interpretivist ones).

(1) Develop a comprehensive *a priori* protocol (e.g., [121] in supplementary materials), and transparently report any changes to the procedure that are decided with their rationale. In our appendix (Table 13), we recommend resources as a starting point for how to develop a protocol. Ideally, the review's protocol should be registered in a public registry like OSF so researchers can avoid starting a review that is already being conducted by others.

(2) For the search process, carefully consider your choice of databases and report your rationale for this choice. We refer to Gusenbauer [48]'s overview of common databases' absolute and relative coverage as a starting point to help researchers make an informed decision. Ideally, use multiple databases to reduce indexing bias/infrastructure issues [146];[30] if so, the search queries for all databases should be provided. Make clear whether your search was focused on breadth/scope vs. a narrow target, and provide a rationale for this choice. Report your exact search string for each database used, along with the (full) date of search, and the exact filters and limits. Pilot the search multiple times to get an idea of how the number of results fluctuates and mitigate the risk of missing relevant publications due to database search issues; the pilot searching should also be documented. Formulate and report the exact IC and EC used for screening papers.

(3) In reporting and conduct: clearly distinguish the review phases by domain. Reviews require a lot of focused organization and documentation over an extended period of time, and in our own experience with other reviews, running multiple phases at the same time (e.g., because a

[30] Depending on research question and field, it can also make sense to search for relevant unpublished research, i.e., grey literature. Whether this is a worthwhile approach for HCI remains to be seen and is out of scope of this paper, however we refer to Garousi et al. [40] for an answer in software engineering as a starting point.

coder is delayed) can easily lead to a paper being forgotten, leading to moments of panic down the line. In reporting, try to keep methodological information for the review in one section overall to help readers understand what was done, instead of spreading the information throughout the paper. Additionally, the ROSES [52] guideline offers a template for review metadata which could be useful for reviewers if supplied as supplementary materials.

(4) Use existing resources like our guiding questions in Table 2, and more specialized guidelines for your specific review type or synthesis method for reporting. For quantitative SRs/meta-analyses, consider using something like PRISMA [108] or ROSES [52]. For qualitative- style reviews (e.g., using thematic analysis), we suggest the use of ENTREQ [142] in addition to our guiding questions because ENTREQ does not cover discussions of limitations.

(5) If possible, employ double screening, double appraisal, double extraction, double coding to reduce errors. In reporting, make bias explicit to then transparently and critically acknowledge and reflect on it. If you do not have the resources for multi-approaches to these stages, consider a percentage-based split (e.g., 30% doubling as a calibration phase) before subsequently screening/appraising/extracting/coding the remaining data individually. However, the reporting should be clear about how many reviewers were involved in each phase and how disagreements were resolved or addressed.

(6) Conduct and report on an appraisal phase. We again suggest resources in our appendix (Table 13) as a starting point for potential quality assessment. If you do not have an appraisal phase, provide a rationale. We do not aim to prescribe that all SRs *need* an appraisal phase (although some would not classify the resulting review as systematic [50])—for some RQs, the quality of studies in the papers will not be immediately relevant. For example, reviews that aim to explore conceptualizations of specific terms or constructs across the literature (e.g., game enjoyment [97] or realism [121]) might view study validity (commonly the focus of appraisal) as out of scope.

(7) Explicitly identify and carefully describe your synthesis methodology. Consider using a formal synthesis methodology—for overviews of potential methodologies, cf. [9, 32, 43, 110, 141].

(8) If you aim to use a less formal synthesis method (e.g., narrative summary or categorical reporting), identify it and describe clearly how you develop and interpret occurrences of categories or patterns. Further, more formalized methods like framework synthesis [21, 22, 31] or content analysis [80, 149] may not require many additional steps, yet could benefit the work by reducing implicit bias, improving clarity, and adding structure to the approach.

(9) If you categorize elements found in the papers that you review, be very clear whether your categorization is based on an inductive or deductive approach, and describe how you developed it or your rationale for choosing an existing one. Critically reflect on your role in developing and applying the categorization—if this review was conducted by someone else, would they arrive at the same findings? An explicit statement of reflexivity [140] might help to position your review in the context of existing knowledge. Although more commonly found in qualitative research, we suspect that this prompt could also be useful for the interpretation of quantitative results.

(10) Try to go beyond just a summary of evidence to synthesize intermediate-level knowledge [61], as for example Brudy et al. [15]'s extensive cross-device taxonomy. Further, it should be made clear how the taxonomy or framework that is developed is connected to and was derived from the review (e.g., taxonomy development of [55]). For this purpose, consider using formal methodology for taxonomy-building [105] or type-building analysis [81], or

using best-fit framework synthesis [21, 22] to apply and refine a conceptual framework (e.g., [122]). Finally, it may be worthwhile to develop an interactive visualization as a "*living*" [2, 33, 34, 104] interactive SR, as for example done by Altarriba Bertran et al. [3], Butler et al. [18], Seifi et al. [129] and Qamar et al. [120]. Such interactive tools can draw on strengths and expertise well established in the HCI community.

(11) Clearly address limitations of both evidence (i.e., the included papers) and methodology. For the latter, go beyond merely repeating your search parameters; consider limitations of each of the review phases.

(12) Provide supplementary materials. Ideally, make sure that these materials are hosted with the paper (e.g., via the ACM DL) to avoid broken links in the future. For example, consider providing the full list of all included papers, the list of all initial search results, the data extraction form, appraisal results, and so forth. Especially when academic conferences create tension between rigorous reporting and concise writing wherein the page length matches the size of the contribution, the supplementary materials can be used to off-load information. A good example in our corpus is provided by Linxen et al. [87].

A Note to Future Reviewers. Our best practices are targeted at researchers conducting and reporting SRs, but we expect that future reviewers may also take note of these. As such, we want to make clear that limitations relating to the above should not *inherently* cause a reviewer to reject the paper. There are many factors that can necessitate a diversion from the "ideal" SR. For example, characteristics of the specific RQ may make an appraisal of studies unnecessary. (However, its absence should be rationalized.) Further, available resources play a key role—e.g., in terms of research team size, time, and access to databases—yet are not equitably distributed among our research community. Our stance is that an "ideal" SR is exceedingly rare, if not impossible and in any case depends on one's epistemic position: in lieu of this achievement, we instead emphasize the importance of transparent handling of limitations and how they may have affected the synthesis. In the expectation that this paper will spark discussions about what counts as "systematic"—we emphasize that there is space in the academic landscape for non-SRs as well [25, 45].

6.3 Limitations of the Guiding Questions and Best Practices

Although the guiding questions for reporting—as non-prescriptive prompts for reflection—target reviews from potentially any epistemic background, they were developed by our team of researchers, and the resulting guiding questions certainly reflect our own pragmatic/critical realist perspective. To what extent they are useful for strongly constructivist or strictly post-positivist approaches, for example, will have to be determined by researchers from those backgrounds. In particular, we see a distinction between reporting a review's search and selection process transparently (resulting in the same corpus papers, if we can assume that the databases are reliable), and reporting and conducting the synthesis, where different epistemic approaches and subjectivity may lead to different results. The suggestions for best practices, on the other hand, much more directly target reviews by researchers whose research paradigm or "way of knowing" is compatible with pragmatism and/or critical realism. We believe that subjective expertise is necessary when combining scientific rigor with critical reflection, and that it can be invaluable to the review process as a whole. Our own collective expertise both informed and was in turn shaped by the reflective and interpretive process of conducting the review, and the data we collected—this is also reflected in our guiding questions and suggestions for best practices.

7 Conclusion

SRs are a research contribution of great potential importance to the field of HCI, yet their reporting quality is often insufficient to confidently follow or replicate. Our umbrella review of SRs at CHI as the flagship HCI conference venue applied two existing checklists for reporting quality (PRISMA and ENTREQ) as a method of critical appraisal. Using a domain-based mapping of the checklists, we extensively explored reporting quality of review papers stating a systematic approach at CHI, and assessed how well the checklists applied to the review papers in our corpus.

With our results, we showcase and describe the domains in which SRs in our field could improve reporting quality, particularly with regards to appraisal, synthesis, and documentation. Based on this research and our combined experience with SRs in HCI, we present guiding questions for HCI researchers to aid in the reporting of future SRs, and compile a set of suggestions for best practices for SRs compatible with a pragmatic or critical realist perspective. We hope that this paper can serve as a primer for researchers interested in research synthesis, and improve methodological clarity and rigor in our community.

Appendix

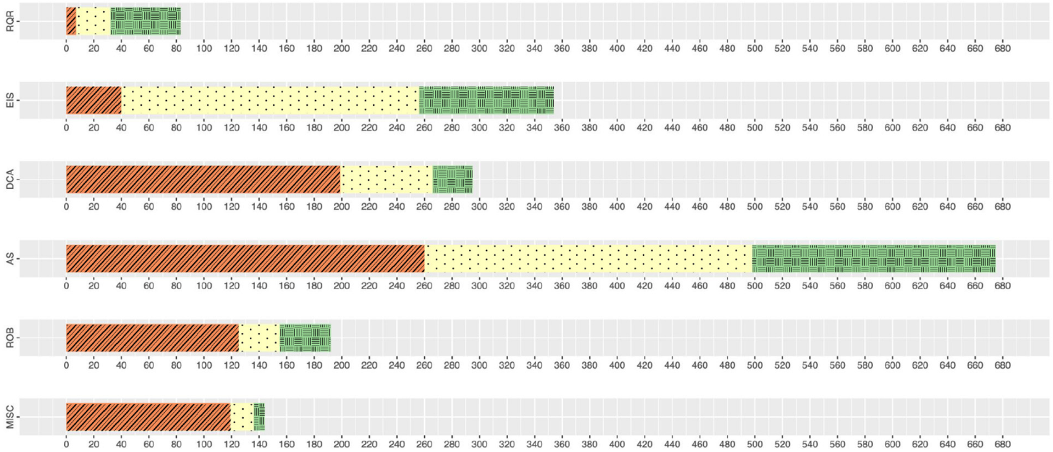


Fig. 8. By domain, stacked bar chart of how many *yes* (green with cross-hatching)/*partial* (yellow with dotted hatching)/*no* (red with slashed hatching) scores the papers got for each item (PRISMA and ENTREQ) categorized within that domain. Note that each row is thus based on a different number of items and scores, hence the varying width - the PRISMA and ENTREQ essentially provide different foci for each domain.

Table 3. Categorization of PRISMA & ENTREQ Items Based on the RQR Review Process Domain

● RQR: Research Questions and Rationale

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P3. Rationale	Describe the rationale for the review in the context of existing knowledge.	💬 Meaning of and location "Context of existing knowledge"; decided on 3 + references and any location. Fewer references as "partial". ❯ 21× yes (of these, 1 elsewhere than intro), 2× partial, 1× no
P4. Objectives	Provide an explicit statement of the objective(s) or question(s) the review addresses.	💬 Any location accepted. For post-hoc contribution statements, we decided on "partial" scoring. ❯ 18× yes (13× in intro, 5× elsewhere), 5× partial, 1× no
E1. Aim	State the research question the synthesis addresses.	💬 Equivalent to P4 but requires question. Objectives and implied research questions were coded as "partial". ❯ 12× yes, 17× partial, 4× no

The final column describes what was discussed in the application of each item (💬) and the resulting scores (❯). The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 4. Categorization of PRISMA Items Based on the EIS Review Process Domain

● EIS: Eligibility Criteria, Identification and Selection, PRISMA

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P5. Eligibility criteria	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	✎ Omitted the second half (synthesis grouping). 💬 Subjective criteria (e.g., "papers addressing X") and cases without inclusion/exclusion screening due to targeted proceedings selection were coded as "partial". ❯ 14× yes, 9× partial, and 1× no
P6. Information sources	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	💬 What counts as a date; partial dates coded as "partial" due to PRISMA documentation. Discussed whether to add rationale requirement to match E5 but did not. Google Scholar accepted as a database. ❯ 2× yes, 21× partial, and 1× no
P7. Search strategy	Present the full search strategies for all databases, reg- isters and websites, including any filters and limits used.	💬 What counts as "full" search strategy. For "yes": keywords as long as it is clear how they were combined and which fields the search was based on; exact (full) query; description of manual search procedure in case of a targeted proceedings selection/search. ❯ 6× yes, 17× partial, and 1× no
P8. Selection process	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	💬 Cases without clearly defined screening stages; targeted proceedings cases were coded as "partial". ❯ 9× yes, 14× partial, and 1× no
P16a. Study selection (a)	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	💬 Reporting of records for multiple sources (required for "yes"); cases with multiple hazily delineated stages. Flow diagram (PRISMA or other) not required for "yes". ❯ 17× yes, 6× partial, and 1× no
P16b. Study selection (b)	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	✎ PRISMA expects a list of all studies close to meeting inclusion criteria; instead we required at least one edgecase/contentious exclusion (with specific reference and rationale) for "yes". 💬 Edgecase *in*clusions, exclusion examples, and edgecase exclusions without specifics or rationale were coded as "partial". ❯ 3× yes, 5× partial, and 16× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The ENTREQ items for this domain are presented in Table 5. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 5. Categorization of ENTREQ Items Based on the EIS Review Process Domain

● EIS: Eligibility Criteria, Identification and Selection, ENTREQ

Item	Description (from [108, 142])	Modifications / Discussions & Scores
E3. Approach to searching	Indicate whether the search was pre-planned (comprehensive search strategies to seek all available studies) or iterative (to seek all available concepts until they theoretical saturation is achieved).	✎ Terms vs. description mismatch; based interpretation on description, not terms. Required clear statements for “yes”, implied descriptions coded as “partial”. 💬 Mixed approaches difficult to infer. ❯ 0× yes, 29× partial, and 4× no
E4. Inclusion criteria	Specify the inclusion/exclusion criteria (e.g. in terms of population, language, year limits, type of publication, study type).	💬 Inclusion and/or exclusion criteria accepted; cases with no screening and unclear phrasing rated “partial” . ❯ 16× yes, 14× partial, and 3× no
E5. Data sources	Describe the information sources used (e.g. electronic databases (MEDLINE, EMBASE, CINAHL, psycINFO, Econlit), grey literature databases (digital thesis, policy reports), relevant organisational websites, experts, information specialists, generic web searches (Google Scholar) hand searching, reference lists) and when the searches conducted; provide the rationale for using the data sources.	✎ Required full date of search to match P6. 💬 Differs from P6 in added rationale requirement; did not adjust P6 but do note this as a potential improvement. “Partial” score if no rationale, no date, or only some databases listed. ❯ 1× yes, 32× partial (2× no rationale, 16× no date, 11× no rationale or date, 3× only some databases), and 0× no
E6. Electronic search strategy	Describe the literature search (e.g. provide electronic search strategies with population terms, clinical or health topic terms, experiential or social phenomena related terms, filters for qualitative research, and search limits).	✎ Required “full” search strategy to match P7. “Partial” if search process was unclear incl. paraphrased query or targeted proceedings search without explicit search description. ❯ 6× yes, 27× partial (3× query but unclear how filtered, 6× paraphrased but clear filters, 13× paraphrased and unclear how filtered, 5× targeted proceedings search without description), and 0× no
E7. Study screening methods	Describe the process of study screening and sifting (e.g. title, abstract and full text review, number of independent reviewers who screened studies).	✎ “Full paper”-review assumed as default if not stated. “Partial” if multi-screening without statement regarding how disagreements were handled; also for targeted proceedings searches as long as it was implied that there was no screening. ❯ 10× yes, 21× partial, and 2× no
E9. Study selection results	Identify the number of studies screened and provide reasons for study exclusion (e.g. for comprehensive searching, provide numbers of studies screened and reasons for exclusion indicated in a figure/flowchart; for iterative searching describe reasons for study exclusion and inclusion based on modifications to the research question and/or contribution to theory development).	✎ Tied to exclusion criteria more than P16a; require number of papers excluded for each exclusion criteria for “yes” (can be 0 if targeted proceedings search). ❯ 12× yes, 16× partial, and 5× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The PRISMA items for this domain are presented in Table 4. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 6. Categorization of PRISMA & ENTREQ Items Based on the DCA Review Process Domain

● DCA: Data Collection and Appraisal

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P9. Data collection process	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	✎ “Partial” if multi-extraction but no statements regarding disagreements; or if unclear who was involved in the extraction. 💬 Synonyms: coding, extraction, analysis, screening, categorization; this was inferred by coders. ❯ 5× yes, 18× partial, and 1× no
P10a. Data items (a)	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	✎ Outcomes vs. other variables; distinguished based on relevance to key research question. 💬 Lacking descriptions for variables; relied on coder subjectivity (uncertainty of replication led to “partial”). ❯ 13× yes 9× partial, and 2× no
P10b. Data items (b)	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	✎ Outcomes vs. other variables; distinguished based on irrelevance to key research question (but still expect data beyond citation reference). ❯ 3× yes, 2× partial, and 19× no
P11. Study risk of bias assessment	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	✎ Also accept quality assessment (not just risk of bias). 💬 Informal/subjective quality assessment; yes if method can be replicated rather than specific results. ❯ 1× yes, 2× partial, and 21× no
P18. Risk of bias in studies	Present assessments of risk of bias for each included study.	✎ Also accept quality assessment (not just risk of bias). 💬 Aggregate reporting; these were rated as “partial”. ❯ 2× yes, 1× partial, and 21× no
E10. Rationale for appraisal	Describe the rationale and approach used to appraise the included studies or selected findings (e.g. assessment of conduct (validity and robustness), assessment of reporting (transparency), assessment of content and utility of the findings).	✎ “Approach” likely expects characterisation of “quality”; we accepted more implicit indications as well (but note that explicit statements would be ideal).“Partial” if no rationale provided. ❯ 0× yes, 2× partial, and 31× no
E11. Appraisal items	State the tools, frameworks and criteria used to appraise the studies or selected findings (e.g. Existing tools: CASP, QARI, COREQ, Mays and Pope [25]; reviewer developed tools; describe the domains assessed: research team, study design, data analysis and interpretations, reporting).	💬 Definition of tools/frameworks/criteria; accepted informal constructs as well. ❯ 1× yes, 1× partial, and 31× no
E12. Appraisal process	Indicate whether the appraisal was conducted independently by more than one reviewer and if consensus was required.	✔. ❯ 1× yes, 1× partial, and 31× no
E13. Appraisal results	Present results of the quality assessment and indicate which articles, if any, were weighted/excluded based on the assessment and give the rationale.	💬 Individual vs. aggregate results; the latter was rated as “partial”. ❯ 1× yes, 1× partial, and 31× no
E14. Data extraction	Indicate which sections of the primary studies were analysed and how were the data extracted from the primary studies? (e.g. all text under the headings “results/conclusions” were extracted electronically and entered into a computer software).	✎ “Full-paper”-analysis assumed as default. Expect indication of single/multi extraction vs. shared extraction and statements regarding disagreements if multi extraction. Expect characterisation of extracted data. 💬 Distinction between data extraction/collection, coding, data annotation, and data generation; added “partial” options for implicit extraction in case of synonym usage/vague terminology. ❯ 2× yes, 29× partial (7× implicit terms but all details, 10× implicit terms and missing details, 12× explicit extraction but missing details), and 2× no

The application column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The checkmark (✔) indicates that an item was applied without modification. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 7. Categorization of PRISMA Items Based on the AS Review Process Domain (for the ENTREQ Items of This Domain, Refer to Table 9)

● AS: Analysis and Synthesis, PRISMA–Part 1

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P12. Effect measures	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	💬 Application in reviews that do not explore effect on outcome; coded it literally, note need for new item. "Yes" for clear effect size measures e.g., in meta-analysis, "partial" for informal, aggregate specifications of effect sizes. ❯ 2× yes, 2× partial, and 20× no
P13a. Synthesis methods (a)	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5: eligibility criteria)).	✎ Single- vs. multiple-synthesis review; decision based on screening criteria (P5) for single-synthesis vs. requires additional criteria for split into subsyntheses "partial" if eligibility for sub-synthesis unclear to coders. 💬 Synthesis identification. ❯ 17× yes (13× only implicitly: 9× implicitly understood as one synthesis and 4× as synthesis split), 6× partial, and 1× no
P13b. Synthesis methods (b)	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	✎ Also accept discussion of other classification challenges. Scored data annotation in labelling/coding/categorization as "partial." 💬 Definitions of data conversion. ❯ 13× yes, 6× partial, and 5× no
P13c. Synthesis methods (c)	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	✎ Re-interpreted: Are figures/tables clear, and is there at least one presenting results-related information. ❯ 22× yes, 2× partial, and 0× no
P13d. Synthesis methods (d)	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	✎ Analysis vs. synthesis considered synonymous for this item. 💬 Level of detail required for synthesis description and rationale; required specification beyond "coding." ❯ 5× yes (2 meta-analysis, 3 other); 14× partial (1 meta analysis no rationale, 1 other but only some descriptions/rationale, 12 other and no rationale); and 5× no
P13e. Synthesis methods (e)	Describe any methods used to explore possible causes of heterogeneity among study results (e.g., subgroup analysis, meta-regression).	💬 Non-statistical heterogeneity; did not count it. ❯ 1× yes, 2× partial, and 21× no
P13f. Synthesis methods (f)	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	✔. ❯ 1× yes, 0× partial, and 23× no
P17. Study characteristics	Cite each included study and present its characteristics.	✎ Citation metadata vs. study characteristics; require 2+ content characteristics for "yes." Accepted external sources e.g., supplementary materials. ❯ 6× yes (2× in paper, 4× outside), 8× partial (4× in paper, 4× outside), and 10× no
P19. Results of individual studies	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	✎ Item label vs. description mismatch (studies vs. outcomes); followed description (outcomes). First item half: descriptive, percentage, and absolute reporting count for summary statistics. ❯ 3× yes, 20× partial, and 1× no
P20a. Results of syntheses (a)	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	✎ Single- vs. multiple-synthesis review; combination of P17 and P18 vs. requires additional summaries for subsyntheses. ❯ 2× yes, 12× partial, and 10× no
P20b. Results of syntheses (b)	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	✎ Require effect size and direction reporting. Descriptive statistics, percentages and absolute reporting scored as "partial." 💬 Definition of statistical synthesis; reviews that identify as "meta-analysis" but aren't. ❯ 3× yes (2× meta-analysis, 1× other), 19× partial (15× only descriptive, 4× incomplete inferential), and 2× no

...

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The checkmark (✔) indicates that an item was applied without modification. This table is continued in Table 8. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 8. Categorization of PRISMA Items Based on the AS Review Process Domain (for the ENTREQ Items of This Domain, Refer to Table 9)

● AS: Analysis and Synthesis, PRISMA—Part 2

Item	Description (from [108, 142])	Modifications/Discussions and Scores
	…	
P20c. Results of syntheses (c)	Present results of all investigations of possible causes of heterogeneity among study results.	✔. ❯ 1× yes, 2× partial, and 21× no
P20d. Results of syntheses (d)	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	✔. ❯ 1× yes, 0× partial, and 23× no
P23a. Discussion (a)	Provide a general interpretation of the results in the context of other evidence.	✎ Require references in discussion section. 💬 Papers that mix results and discussion/interpretation; we coded these as "partial." ❯ 21× yes, 2× partial, and 1× no
P23d. Discussion (d)	Discuss implications of the results for practice, policy, and future research.	✎ Interpreted "practice, policy, and future research" as "at least two separate domains, e.g., researchers and practitioners, or theory vs. practice. This item also requires "explicit recommendations" [109]; we accepted implicit recommendations as well. 💬 Distinction between implicit and explicit recommendations. ❯ 14× yes, 9× partial, and 1× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The checkmark (✔) indicates that an item was applied without modification. This is a continuation of Table 7. The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

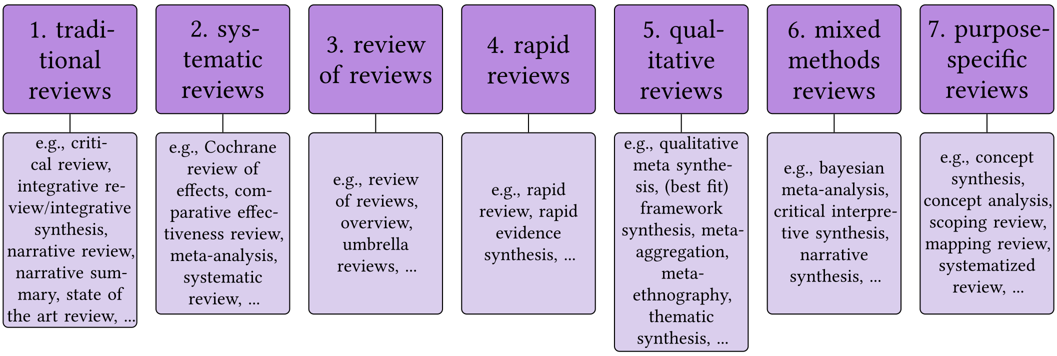

1. traditional reviews
e.g., critical review, integrative review/integrative synthesis, narrative review, narrative summary, state of the art review, ...
2. systematic reviews
e.g., Cochrane review of effects, comparative effectiveness review, meta-analysis, systematic review, ...
3. review of reviews
e.g., review of reviews, overview, umbrella reviews, ...
4. rapid reviews
e.g., rapid review, rapid evidence synthesis, ...
5. qualitative reviews
e.g., qualitative meta synthesis, (best fit) framework synthesis, meta-aggregation, meta-ethnography, thematic synthesis, ...
6. mixed methods reviews
e.g., bayesian meta-analysis, critical interpretive synthesis, narrative synthesis, ...
7. purpose-specific reviews
e.g., concept synthesis, concept analysis, scoping review, mapping review, systematized review, ...

Fig. 9. Families of review types according to Sutton et al. [139].

Table 9. Categorization of ENTREQ Items Based on the AS Review Process Domain (for the PRISMA Items, Refer to Table 7 and Table 8)

● AS: Analysis and Synthesis, ENTREQ

Item	Description (from [108, 142])	Modifications/Discussions and Scores
E2. Synthesis methodology	Identify the synthesis methodology or theoretical framework which underpins the synthesis, and describe the rationale for choice of methodology (e.g., meta- ethnography, thematic synthesis, critical interpretive synthesis, grounded theory synthesis, realist synthesis, meta-aggregation, meta-study, framework synthesis).	✎ Analysis vs. synthesis considered synonymous for this item. 💬 Distinction between "identify" here vs. "describe" in P13d. ❯ 0× yes, 28× partial (17× identification without rationale, 1× no identification but description with rationale, 10× description without rationale), and 5× no
E8. Study characteristics	Present the characteristics of the included studies (e.g., year of publication, country, population, number of participants, data collection, methodology, analysis, research questions).	✎ Distinction between paper-related and content/study-related; required 2+ of the latter for "yes" to match P17. Accepted external sources e.g., supplementary materials. ❯ 8× yes, 9× partial, and 16× no
E15. Software	State the computer software used, if any.	💬 Uncertainty what stage this refers to (e.g., data extraction/analysis/other stage); accepted specific software mention for any stage. ❯ 12× yes, 0× partial, and 21× no
E16. Number of reviewers	Identify who was involved in coding and analysis.	✎ Coding vs. analysis considered synonymous for this item. Did not require specific identification ("two coders" is enough). ❯ 8× yes, 14× partial, and 11× no
E17. Coding	Describe the process for coding of data (e.g., line by line coding to search for concepts).	✎ Require type of coding, indication of inductive/deductive, and code examples for "yes". Do not require number of reviewers for this "process" item as covered by E16. ❯ 5× yes, 17× partial, and 11× no
E18. Study comparison	Describe how were comparisons made within and across studies (e.g., subsequent studies were coded into pre-existing concepts, and new concepts were created when deemed necessary).	✎ Accepted mentions of iterating, grouping/clustering, pattern-finding as part of the coding/analysis/synthesis methodology for "partial". 💬 What counts as clear description of study/paper comparison. ❯ 7× yes, 18× partial, and 8× no
E19. Derivation of themes	Explain whether the process of deriving the themes or constructs was inductive or deductive.	✎ Accepted also descriptions of theme-like constructs without explicit mention of themes for "partial". ❯ 9× yes, 9× partial (3× themes mentioned but process unclear, 6× unlabelled themes), and 15× no
E20. Quotations	Provide quotations from the primary studies to illus- trate themes/constructs, and identify whether the quotations were participant quotations of the author's interpretation.	✎ No participant quotations; ignored second item half. Accepted use of quotations in general (without connection to themes) for "partial". 💬 Distinction between actual quotation and quotation marks for emphasis; did not count single-word quotations. ❯ 5× yes, 19× partial (theme connection unclear), and 9× no
E21. Synthesis output	Present rich, compelling and useful results that go beyond a summary of the primary studies (e.g., new interpretation, models of evidence, conceptual models, analytical framework, development of a new theory or construct).	✎ Required clear new framework, taxonomy, model, or construct for "yes"; coded less clear forms of knowledge as "partial" (recommendations, guidelines, implications, design considerations, lessons learned, ...) 💬 Definitions of "go[ing] beyond a summary" and "rich, compelling and useful". ❯ 7× yes (6× framework or similar; 1× framework or similar + living document, 15× partial (13× recommendations or similar; 2× living document), and 11× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 10. Categorization of PRISMA Items Based on the ROB Review Process Domain (no ENTREQ Were Applicable)

● ROB: Risk of Bias in Synthesis

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P14. Reporting bias assessment	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	✎ Expanded statement [109] expects reporting of number of reviewers involved in assessing risk of bias; we ignored this aspect for this item. 💬 Considered broadening item to quality assessment in general; decided against it as it specifies the type of risk of bias. ❯ 1× yes, 1× partial, and 22× no
P15. Certainty assessment	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	✎ Interpreted risk of bias for this item more broadly; accept any quality assessment. Based on [109], we require process information (number of reviewers and how disagreements were resolved). 💬 Confidence intervals and p values; count only as “partial” unless framed in the context of quality or bias. ❯ 2× yes, 1× partial, and 21× no
P21. Reporting biases	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	💬 Considered broadening item to quality assessment in general; decided against it as it specifies the type of risk of risk (and to match P14). ❯ 1× yes, 0× partial, and 23× no
P22. Certainty of evidence	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	✎ Accept aggregate quality assessment for “yes”; accept for confidence intervals only when framed in context of trust/certainty/confidence or quality/risk of bias. ❯ 1× yes, 2× partial, and 21× no
P23b. Discussion (b)	Discuss any limitations of the evidence included in the review.	✎ Require explicit label of limitation (section header, terms “limitation”, “limiting” or synonym. Implicit limitations addressed in discussion/reflection coded as “partial.” ❯ 9× yes, 11× partial, and 4× no
P23c. Discussion (c)	Discuss any limitations of the review processes used.	✎ Require limitations addressal of at least two parts of review procedure (e.g., search/selection and analysis/synthesis). Require explicit limitation to match P23b. ❯ 6× yes, 11× partial, and 7× no
P25. Support	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	✎ Would have accepted statements of non-support in line with E&E paper [109]. Required rough implication of type of support for “yes.” ❯ 16× yes, 4× partial, and 4× no
P26. Competing interests	Declare any competing interests of review authors.	✎ More broadly interpreted “competing interest” than the examples listed in the E&E paper [109]. 💬 Noted inclusions of authors’ own work in review corpus and whether this was addressed; but did not code it. ❯ 1 yes, 0× partial and 23× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 11. Categorization of PRISMA Items Based on the MISC Review Process Domain (no ENTREQ Were Applicable)

🟢 MISC: Miscellaneous/Uncategorized

Item	Description (from [108, 142])	Modifications/Discussions and Scores
P1. Title	Identify the report as a systematic review.	✎ E&E paper requires this to be in the title; identifications in abstract were coded as "partial." We ignored requirement in E&E paper [109] on providing key information on main objective. ❯ 8× yes, 5× partial, and 11× no
P2. Abstract	PRISMA 2020 Abstracts checklist [106]	💬 Impossible given CHI's abstract length restrictions; decided against interpreting item to apply across whole paper, i.e., coded literally. ❯ 0× yes, 0× partial, and 24× no/not applicable
P24a. Registration and protocol (a)	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	💬 General mentions of using PRISMA as a protocol were not counted; item refers to *a priori* protocols developed for specific review. ❯ 0× yes, 0× partial, and 24× no
P24b. Registration and protocol (b)	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	💬 Same as P24a: General mentions of using PRISMA as a protocol were not counted; item refers to *a priori* protocols developed for specific review. ❯ 0× yes, 0× partial, and 24× no
P24c. Registration and protocol (c)	Describe and explain any amendments to information provided at registration or in the protocol.	✎ Also accepted any indications of changes to planned review procedure (even without protocol) for "partial." ❯ 0× yes, 2× partial (no registration/protocol but changes explained), and 22× no
P27. Availability of data, code, and other materials	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	✎ Focused on supplementary materials and appendix; ignored in-manuscript tables. Require each of the materials specifically mentioned for "yes" ("template data collection forms; data extracted from included studies; data used for all analyses; analytic code"). Coded some materials being provided, or statements of that materials would be available upon request, or intent to provide materials (e.g., broken link to external data) as different "partials." ❯ 0× yes, 10× partial (7× some material; 1× on request; 2× intent but broken access), and 14× no

The final column notes discussions (💬) and when/how items were re-interpreted (✎), as well as resulting scores (❯). The review process domain is denoted by the abbreviation and a coloured circle, for easier reference between the different tables throughout the article.

Table 12. Basic Descriptive Statistics for Review Papers in Our Main Corpus Analysis for This Paper (n = 41)

Variable	Descriptive Summary		
	Median	Interquartile Range	Range
Number of Authors	4	3–5	1–11
Range of Years Covered[a]	11	6.5–19.25	1–38
Number of Papers in Each Corpus[b]	100	60–270	25–2,768
Number of Papers in Initial Search[c]	597.5	323.5–2,038.5	164–64,249
Papers by Country of University Affiliations	United States of America: 17 United Kingdom: 8 Denmark, Germany: 7 Australia: 5 Sweden, Switzerland: 4 Canada, Finland: 3 Austria, France: 2 China, Cyprus, Ireland, Netherlands, Portugal, Qatar: 1		
Papers that Received an Award *(39%)*[d]	Honourable Mention: 15 Best Paper: 1		

[a] Range only explicitly determinable for 24 corpus papers, i.e., 58.54%.

[b] Number only explicitly determinable for 37 review papers, i.e., 90.24%.

[c] Number only explicitly determinable for 34 review papers, i.e., 82.93%.

[d] This was 36% for the 50 papers in the initial corpus (literature and non-literature) with 16 honourable mentions and 2 best papers.

Table 13. Non-Exhaustive List of Resources for Broadest Level Review Types

Broad Review Types	Protocol	Conduct	Appraisal of Units of Analysis	Reporting
Quantitative SRs	PRISMA-P [98] (also endorsed by JBI [6, Ch.1.3]), PROSPERO [14], COCHRANE MECIR C1-C23 *(and formerly PR1-PR44—now retired)* [56], NIRO-SR [144, Pt. A] ...	JBI [6] *, in usage though not intent: PRISMA 2020 [108], ...*	[74, Appx. A], CONSORT [128], STROBE [148], CASP [119], [26, p.4], ...	PRISMA 2020 [108][a], QUORUM [99], NIRO-SR [144, Pt. B], ...
Qualitative SRs	JBI [6, Ch. 3.6], ...	JBI [6, Ch. 3.7], ...	Kmet et al. [74, Appx. B], CASP [119], COREQ [143], JBI [6, Appx. 3.1], ...	ENTREQ [142][a], ...
Mixed methods SRs	JBI [6, Ch. 8.4], ...	JBI [6, Ch. 8.5], ...	*pick from the 2 cells above*	JBI [6, Ch. 8.5]
Umbrella reviews	JBI [6, Ch. 9.2] , ...	JBI [6, Ch. 9], ...	ROSES [52], AMSTAR [130], AMSTAR 2 [131], ROBIS [150], MOOSE [138], JBI [6, Appx. 9.1], PRISMA 2020 [108] / ENTREQ [142])[a], ...	JBI [6, Ch. 9.3], PRIOR [41] , ...
Scoping reviews	Peters et al. [112], JBI [6, Ch. 10.2], ...	Arksey and O'Malley [4], Levac et al. [85], ...	*some of the qualitative appraisal methods may be applicable*	PRISMA ScR [145], Scoping Review Checklist (SCR) [29], JBI [6, Ch. 10.3]

[a] As discussed in the paper, while PRISMA and ENTREQ are intended as guidelines for reporting quality, we applied them as critical appraisal tools due to the nature of our research questions (i.e., reporting quality as the focus of our appraisal), and as the other tools for appraisal of systematic reviews as units of analysis are too specialized to cover the variety of studies and reviews encountered in HCI. Further, PRISMA is notably often referred to as guidance for how to conduct reviews, although it was designed for guidance on reporting. Readers should note that some resources are intended only for specific subtypes of the broader review types listed (e.g., PRISMA is intended for interventional research questions and meta-analyses). Additionally, we recommend the more extensive list of resources (including software tools) curated by Marshall and Sutton and maintained by the Evidence Synthesis Group (Newcastle University) and the School of Health and Related Research (University of Sheffield) [93].

Acknowledgement

Sincere thanks go to our Waterloo HCI colleagues who provided feedback on early drafts, as well as colleagues across the world who were willing to discuss epistemologies with the first author.

References

[1] Jacob Abbott, Haley MacLeod, Novia Nurain, Gustave Ekobe, and Sameer Patil. 2019. Local standards for anonymization practices in health, wellness, accessibility, and aging research at CHI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, 1–14. DOI: https://doi.org/10.1145/3290605.3300692

[2] Elie A. Akl, Joerg J. Meerpohl, Julian Elliott, Lara A. Kahale, Holger J. Schünemann, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, Charlotte Pestridge, Lauren Albrecht, Tanya Horsley, Joanne Platt, Rebecca Armstrong, Phi Hung Nguyen, Robert Plovnick, Anneliese Arno, Noah Ivers, Gail Quinn, Agnes Au, Renea Johnston, Gabriel Rada, Matthew Bagg, Arwel Jones, Philippe Ravaud, Catherine Boden, Lara Kahale, Bernt Richter, Isabelle Boisvert, Homa Keshavarz, Rebecca Ryan, Linn Brandt, Stephanie A. Kolakowsky-Hayner, Dina Salama, Alexandra Brazinova, Sumanth Kumbargere Nagraj, Georgia Salanti, Rachelle Buchbinder, Toby Lasserson, Lina Santaguida, Chris Champion, Rebecca Lawrence, Nancy Santesso, Jackie Chandler, Zbigniew Les, Holger J. Schünemann, Andreas Charidimou, Stefan Leucht, Ian Shemilt, Roger Chou, Nicola Low, Diana Sherifali, Rachel Churchill, Andrew Maas, Reed Siemieniuk, Maryse C. Cnossen, Harriet MacLehose, Mark Simmonds, Marie-Joelle Cossi, Malcolm Macleod, Nicole Skoetz, Michel Counotte, Iain Marshall, Karla Soares-Weiser, Samantha Craigie, Rachel Marshall, Velandai Srikanth, Philipp Dahm, Nicole Martin, Katrina Sullivan, Alanna Danilkewich, Laura Martínez García, Anneliese Synnot, Kristen Danko, Chris Mavergames, Mark Taylor, Emma Donoghue, Lara J. Maxwell, Kris Thayer, Corinna Dressler, James McAuley, James Thomas, Cathy Egan, Steve McDonald, Roger Tritton, Julian Elliott, Joanne McKenzie, Guy Tsafnat, Sarah A. Elliott, Joerg Meerpohl, Peter Tugwell, Itziar Etxeandia, Bronwen Merner, Alexis Turgeon, Robin Featherstone, Stefania Mondello, Tari Turner, Ruth Foxlee, Richard Morley, Gert van Valkenhoef, Paul Garner,

Marcus Munafo, Per Vandvik, Martha Gerrity, Zachary Munn, Byron Wallace, Paul Glasziou, Melissa Murano, Sheila A. Wallace, Sally Green, Kristine Newman, Chris Watts, Jeremy Grimshaw, Robby Nieuwlaat, Laura Weeks, Kurinchi Gurusamy, Adriani Nikolakopoulou, Aaron Weigl, Neal Haddaway, Anna Noel-Storr, George Wells, Lisa Hartling, Annette O'Connor, Wojtek Wiercioch, Jill Hayden, Matthew Page, Luke Wolfenden, Mark Helfand, Manisha Pahwa, Juan José Yepes Nuñez, Julian Higgins, Jordi Pardo Pardo, Jennifer Yost, Sophie Hill, and Leslea Pearson. 2017. Living systematic reviews: 4. Living guideline recommendations. *Journal of Clinical Epidemiology* 91 (Nov. 2017), 47–53. DOI: https://doi.org/10.1016/j.jclinepi.2017.08.009

[3] Ferran A. Bertran, Samvid Jhaveri, Rosa Lutz, Katherine Isbister, and Danielle Wilde. 2019. *Making Sense of Human-Food Interaction.* ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3290605.3300908

[4] Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology* 8, 1 (2005), 19–32. DOI: https://doi.org/10.1080/1364557032000119616

[5] Edoardo Aromataris, Ritin Fernandez, Christina M. Godfrey, Cheryl Holly, Hanan Khalil, and Patraporn Tungpunkom. 2015. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *International Journal of Evidence-Based Healthcare* 13, 3 (Sep. 2015), 132–140. DOI: https://doi.org/10.1097/xeb.0000000000000055

[6] Edoardo Aromataris and Zachary Munn. 2020. JBI Manual for Evidence Synthesis. JBI, 2020. Retrieved from https://synthesismanual.jbi.global

[7] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. 2021. Are you open? A content analysis of transparency and openness guidelines in HCI journals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* ACM, New York, NY, Article 176, 10 pages. DOI: https://doi.org/10.1145/3411764.3445584

[8] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. *Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience.* ACM, New York, NY, 2689–2698. DOI: https://doi.org/10.1145/1978942.1979336

[9] Elaine Barnett-Page and James Thomas. 2009. Methods for the synthesis of qualitative research: A critical review. *BMC Medical Research Methodology* 9, 1 (Aug. 2009), 1–11. DOI: https://doi.org/10.1186/1471-2288-9-59

[10] Joanna Bergström, Tor-Salve Dalsgaard, Jason Alexander, and Kasper Hornbæk. 2021. How to evaluate object selection and manipulation in VR? Guidelines from 20 years of studies. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '21).* ACM, New York, NY, Article 533, 20 pages. DOI: https://doi.org/10.1145/3411764.3445193

[11] Lisa Bero. 2017. Systematic review: A method at risk for being corrupted. *American Journal of Public Health* 107, 1 (2017), 93–96. DOI: https://doi.org/10.2105/AJPH.2016.303518

[12] Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. MetaGen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, 1653–1656. DOI: https://doi.org/10.1145/3397271.3401190

[13] Sebastian K. Boell and Dubravka Cecez-Kecmanovic. 2010. Literature reviews and the hermeneutic circle. *Australian Academic & Research Libraries* 41, 2 (Jun. 2010), 129–144. DOI: https://doi.org/10.1080/00048623.2010.10721450

[14] Alison Booth, Mike Clarke, Gordon Dooley, Davina Ghersi, David Moher, Mark Petticrew, and Lesley Stewart. 2012. The nuts and bolts of PROSPERO: An international prospective register of systematic reviews. *Systematic Reviews* 1, 1 (Feb. 2012), 1–8. DOI: https://doi.org/10.1186/2046-4053-1-2

[15] Frederik Brudy, Christian Holz, Roman Rädle, Chi-Jui Wu, Steven Houben, Clemens N. Klokmose, and Nicolai Marquardt. 2019. Cross-device taxonomy: Survey, opportunities and challenges of interactions spanning across multiple devices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–28. DOI: https://doi.org/10.1145/3290605.3300792

[16] Emeline Brulé, Brianna J. Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. 2020. Review of quantitative empirical evaluations of technology for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–14. DOI: https://doi.org/10.1145/3313831.3376749

[17] Nina Buscemi, Lisa Hartling, Ben Vandermeer, Lisa Tjosvold, and Terry P. Klassen. 2006. Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology* 59, 7 (Jul. 2006), 697–703. DOI: https://doi.org/10.1016/j.jclinepi.2005.11.010

[18] Matthew Butler, Leona M. Holloway, Samuel Reinders, Cagatay Goncu, and Kim Marriott. 2021. Technology developments in touch-based accessible graphics: A systematic review of research 2010-2020. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY. DOI: https://doi.org/10.1145/3411764.3445207

[19] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 981–992. DOI: https://doi.org/10.1145/2858036.2858498

[20] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–15. DOI: https://doi.org/10.1145/3290605.3300733

[21] Christopher Carroll, Andrew Booth, and Katy Cooper. 2011. A worked example of "best fit" framework synthesis: A systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology* 11, 1 (2011), 1–9. DOI: https://doi.org/10.1186/1471-2288-11-29

[22] Christopher Carroll, Andrew Booth, Joanna Leaviss, and Jo Rick. 2013. "Best fit" framework synthesis: Refining the method. *BMC Medical Research Methodology* 13, 1 (Mar. 2013), 1–16. DOI: https://doi.org/10.1186/1471-2288-13-37

[23] Iain Chalmers, Larry V. Hedges, and Harris Cooper. 2002. A brief history of research synthesis. *Evaluation & the Health Professions* 25, 1 (Mar. 2002), 12–37. DOI: https://doi.org/10.1177/0163278702025001003

[24] Sheung-Tak Cheng and Fan Zhang. 2020. A comprehensive meta-review of systematic reviews and meta-analyses on nonpharmacological interventions for informal dementia caregivers. *BMC Geriatrics* 20, 1 (Apr. 2020), 1–24. DOI: https://doi.org/10.1186/s12877-020-01547-2

[25] John A. Collins and Bart C. J. M. Fauser. 2005. Balancing the strengths of systematic and narrative reviews. *Human Reproduction Update* 11, 2 (Mar. 2005), 103–104. DOI: https://doi.org/10.1093/humupd/dmh058

[26] Thomas M. Connolly, Elizabeth A. Boyle, Ewan MacArthur, Thomas Hainey, and James M. Boyle. 2012. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education* 59, 2 (2012), 661–686. DOI: https://doi.org/10.1016/j.compedu.2012.03.004

[27] Deborah J. Cook, Cynthia D. Mulrow, and R. B. Haynes. 1997. Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine* 126, 5 (1997), 376–380. DOI: https://doi.org/10.7326/0003-4819-126-5-199703010-00006

[28] Harris Cooper. 2016. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. SAGE, Los Angeles.

[29] Simon Cooper, Robyn Cant, Michelle Kelly, Tracy Levett-Jones, Lisa McKenna, Philippa Seaton, and Fiona Bogossian. 2021. An evidence-based checklist for improving scoping review quality. *Clinical Nursing Research* 30, 3 (2021), 230–240. DOI: https://doi.org/10.1177/1054773819846024

[30] Daniela S. Cruzes and Tore Dybå. 2010. Synthesizing Evidence in Software Engineering Research. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '10)*. ACM, New York, NY, Article 1, 10 pages. DOI: https://doi.org/10.1145/1852786.1852788

[31] Mary Dixon-Woods. 2011. Using framework-based synthesis for conducting reviews of qualitative studies. *BMC Medicine* 9, 1 (Apr. 2011), 1–2. DOI: https://doi.org/10.1186/1741-7015-9-39

[32] Mary Dixon-Woods, Shona Agarwal, David Jones, Bridget Young, and Alex Sutton. 2005. Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy* 10, 1 (Jan. 2005), 45–53. DOI: https://doi.org/10.1177/135581960501000110

[33] Julian H. Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A. Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, David Tovey, Ian Shemilt, James Thomas, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, Charlotte Pestridge, Lauren Albrecht, Tanya Horsley, Joanne Platt, Rebecca Armstrong, Phi Hung Nguyen, Robert Plovnick, Anneliese Arno, Noah Ivers, Gail Quinn, Agnes Au, Renea Johnston, Gabriel Rada, Matthew Bagg, Arwel Jones, Philippe Ravaud, Catherine Boden, Lara Kahale, Bernt Richter, Isabelle Boisvert, Homa Keshavarz, Rebecca Ryan, Linn Brandt, Stephanie A. Kolakowsky-Hayner, Dina Salama, Alexandra Brazinova, Sumanth Kumbargere Nagraj, Georgia Salanti, Rachelle Buchbinder, Toby Lasserson, Lina Santaguida, Chris Champion, Rebecca Lawrence, Nancy Santesso, Jackie Chandler, Zbigniew Les, Holger J. Schünemann, Andreas Charidimou, Stefan Leucht, Ian Shemilt, Roger Chou, Nicola Low, Diana Sherifali, Rachel Churchill, Andrew Maas, Reed Siemieniuk, Maryse C. Cnossen, Harriet MacLehose, Mark Simmonds, Marie-Joelle Cossi, Malcolm Macleod, Nicole Skoetz, Michel Counotte, Iain Marshall, Karla Soares-Weiser, Samantha Craigie, Rachel Marshall, Velandai Srikanth, Philipp Dahm, Nicole Martin, Katrina Sullivan, Alanna Danilkewich, Laura Martínez García, Anneliese Synnot, Kristen Danko, Chris Mavergames, Mark Taylor, Emma Donoghue, Lara J. Maxwell, Kris Thayer, Corinna Dressler, James McAuley, James Thomas, Cathy Egan, Steve McDonald, Roger Tritton, Julian Elliott, Joanne McKenzie, Guy Tsafnat, Sarah A. Elliott, Joerg Meerpohl, Peter Tugwell, Itziar Etxeandia, Bronwen Merner, Alexis Turgeon, Robin Featherstone, Stefania Mondello, Tari Turner, Ruth Foxlee, Richard Morley, Gert van Valkenhoef, Paul Garner, Marcus Munafo, Per Vandvik, Martha Gerrity, Zachary Munn, Byron Wallace, Paul Glasziou, Melissa Murano, Sheila A. Wallace, Sally Green, Kristine Newman, Chris Watts, Jeremy Grimshaw, Robby Nieuwlaat, Laura Weeks, Kurinchi Gurusamy, Adriani Nikolakopoulou, Aaron Weigl, Neal Haddaway, Anna Noel-Storr, George Wells, Lisa Hartling, Annette O'Connor, Wojtek Wiercioch, Jill Hayden, Matthew Page, Luke Wolfenden, Mark Helfand, Manisha Pahwa, Juan José Yepes Nuñez, Julian Higgins, Jordi Pardo Pardo, Jennifer Yost, Sophie Hill, and Leslea Pearson. 2017. Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology* 91 (Nov. 2017), 23–30. DOI: https://doi.org/10.1016/j.jclinepi.2017.08.010

[34] Julian H. Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian P. T. Higgins, Chris Mavergames, and Russell L. Gruen. 2014. Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. *PLoS Medicine* 11, 2 (Feb. 2014), e1001603. DOI: https://doi.org/10.1371/journal.pmed.1001603

[35] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel P. Robert. 2021. A meta-analysis of human personality and robot acceptance in human-robot interaction. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 711, 18 pages. DOI: https://doi.org/10.1145/3411764.3445542

[36] Guy Faulkner, Matthew J. Fagan, and Jacqueline Lee. 2021. Umbrella reviews (systematic review of reviews). *International Review of Sport and Exercise Psychology* (Jun. 2021), 1–18. DOI: https://doi.org/10.1080/1750984x.2021.1934888

[37] Collaboration for Environmental Evidence. 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management. Published at Environmental Evidence, Version 4.2. Retrieved from www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.

[38] Fiorenzo Franceschini, Domenico Maisano, and Luca Mastrogiacomo. 2016. Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics* 10, 4 (Nov. 2016), 933–953. DOI: https://doi.org/10.1016/j.joi.2016.07.003

[39] Christopher Frauenberger. 2019. Entanglement HCI the next wave? *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 1 (Nov. 2019), Article 2, 27 pages. DOI: https://doi.org/10.1145/3364998

[40] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (Feb. 2019), 101–121. DOI: https://doi.org/10.1016/j.infsof.2018.09.006

[41] Michelle Gates, Allison Gates, Dawid Pieper, Ricardo M. Fernandes, Andrea C. Tricco, David Moher, Sue E. Brennan, Tianjing Li, Michelle Pollock, Carole Lunny, Dino Sepúlveda, Joanne E McKenzie, Shannon D. Scott, Karen A. Robinson, Katja Matthias, Konstantinos I. Bougioukas, Paolo Fusar-Poli, Penny Whiting, Stephana J. Moss, and Lisa Hartling. 2022. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. *BMJ* (Aug. 2022), e070849. DOI: https://doi.org/10.1136/bmj-2022-070849

[42] Sneha Gathani, Peter Lim, and Leilani Battle. 2020. Debugging database queries: A survey of tools, techniques, and users. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–16. DOI: https://doi.org/10.1145/3313831.3376485

[43] David Gough, James Thomas, and Sandy Oliver. 2012. Clarifying differences between review designs and methods. *Systematic Reviews* 1, 1 (Jun. 2012), 1–9. DOI: https://doi.org/10.1186/2046-4053-1-28

[44] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal* 26, 2 (May 2009), 91–108. DOI: https://doi.org/10.1111/j.1471-1842.2009.00848.x

[45] Trisha Greenhalgh, Sally Thorne, and Kirsti Malterud. 2018. Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation* 48, 6 (Apr. 2018), e12931. DOI: https://doi.org/10.1111/eci.12931

[46] Egon G. Guba and Yvonna S. Lincoln. 1994. Competing paradigms in qualitative research. *Handbook of Qualitative Research* 2, 163–194 (1994), 105.

[47] Katie E. Gunnell, Veronica J. Belcourt, Jennifer R. Tomasone, and Laura C. Weeks. 2022. Systematic review methods. *International Review of Sport and Exercise Psychology* 15, 1 (Jan. 2022), 5–29. DOI: https://doi.org/10.1080/1750984x.2021.1966823

[48] Michael Gusenbauer. 2022. Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics* 127, 5 (May 2022), 2683–2745. DOI: https://doi.org/10.1007/s11192-022-04289-7

[49] Neal R. Haddaway, Alison Bethel, Lynn V. Dicks, Julia Koricheva, Biljana Macura, Gillian Petrokofsky, Andrew S. Pullin, Sini Savilaakso, and Gavin B. Stewart. 2020. Eight problems with literature reviews and how to fix them. *Nature Ecology & Evolution* 4, 12 (Oct. 2020), 1582–1589. DOI: https://doi.org/10.1038/s41559-020-01295-x

[50] Neal R. Haddaway, Magnus Land, and Biljana Macura. 2017. "A little learning is a dangerous thing": A call for better understanding of the term 'systematic review'. *Environment International* 99 (Feb. 2017), 356–360. DOI: https://doi.org/10.1016/j.envint.2016.12.020

[51] Neal R. Haddaway and Biljana Macura. 2018. The role of reporting standards in producing robust literature reviews. *Nature Climate Change* 8, 6 (May 2018), 444–447. DOI: https://doi.org/10.1038/s41558-018-0180-3

[52] Neal R. Haddaway, Biljana Macura, Paul Whaley, and Andrew S. Pullin. 2018. ROSES reporting standards for systematic evidence syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence* 7, 1 (Mar. 2018), 1–8. DOI: https://doi.org/10.1186/s13750-018-0121-7

[53] Karin Hannes, Andrew Booth, Janet Harris, and Jane Noyes. 2013. Celebrating methodological challenges and changes: Reflecting on the emergence and importance of the role of qualitative evidence in Cochrane reviews. *Systematic Reviews* 2, 1 (Oct. 2013), 1–10. DOI: https://doi.org/10.1186/2046-4053-2-84

[54] Lon Å. E. J. Hansson, Teresa Cerratto Pargman, and Daniel S. Pargman. 2021. A decade of sustainable HCI: connecting SHCI to the sustainable development goals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–19 DOI: https://doi.org/10.1145/3411764.3445069

[55] Julia Hertel, Sukran Karaosmanoglu, Susanne Schmidt, Julia Bräker, Martin Semmann, and Frank Steinicke. 2021. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '21)*, 431–440. DOI: https://doi.org/10.1109/ISMAR52148.2021.00060

[56] J. P. T. Higgins, T. Lasserson, J. Thomas, E. Flemyng, and R. Churchill. 2023. Standards for the Conduct of New Cochrane Intervention Reviews. Protocol development: C1–C23; conduct: C24–C75. pages #55

[57] Julia Himmelsbach, Stephanie Schwarz, Cornelia Gerdenitsch, Beatrix Wais-Zechmann, Jan Bobeth, and Manfred Tscheligi. 2019. Do we care about diversity in human computer interaction: A comprehensive content analysis on diversity dimensions in research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–16. DOI: https://doi.org/10.1145/3290605.3300720

[58] Sebastian Hinde and Eldon Spackman. 2014. Bidirectional citation searching to completion: An exploration of literature searching methods. *PharmacoEconomics* 33, 1 (Aug. 2014), 5–11. DOI: https://doi.org/10.1007/s40273-014-0205-3

[59] Teresa Hirzle, Maurice Cordts, Enrico Rukzio, Jan Gugenheimer, and Andreas Bulling. 2021. A critical assessment of the use of SSQ as a measure of general discomfort in VR head-mounted displays. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 530, 14 pages. DOI: https://doi.org/10.1145/3411764.3445361

[60] Kai Holländer, Mark Colley, Enrico Rukzio, and Andreas Butz. 2021. A taxonomy of vulnerable road users for HCI based on a systematic literature review. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 158, 13 pages. DOI: https://doi.org/10.1145/3411764.3445480

[61] Kristina Höök and Jonas Löwgren. 2012. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 3 (Oct. 2012), Article 23, 18 pages. DOI: https://doi.org/10.1145/2362364.2362371

[62] Kasper Hornbæk, Søren S. Sander, Javier A. Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, 3523–3532. DOI: https://doi.org/10.1145/2556288.2557004

[63] Xin Huang, He Zhang, Xin Zhou, Muhammad Ali Babar, and Song Yang. 2018. Synthesizing qualitative research in software engineering: A critical review. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*. ACM, New York, NY, 1207–1218. https://doi.org/10.1145/3180155.3180235

[64] Lee Humphreys, Neil A. Lewis, Katherine Sender, and Andrea S. Won. 2021. Integrating qualitative methods and open science: Five principles for more trustworthy research. *Journal of Communication* 71, 5 (Aug. 2021), 855–874. DOI: https://doi.org/10.1093/joc/jqab026

[65] Netta Iivari, Leena Ventä-Olkkonen, Sumita Sharma, Tonja Molin-Juustila, and Essi Kinnunen. 2021. CHI against bullying: Taking stock of the past and envisioning the future. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 357, 17 pages. DOI: https://doi.org/10.1145/3411764.3445282

[66] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. "I can't get no sleep": Discussing #insomnia on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1501–1510. DOI: https://doi.org/10.1145/2207676.2208612

[67] Paweł Jemioło, Dawid Storman, Barbara Giżycka, and Antoni Ligęza. 2021. Emotion elicitation with stimuli datasets in automatic affect recognition studies – Umbrella review. In *Human-Computer Interaction (INTERACT '21)*. Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.), Springer International Publishing, Cham, 248–269.

[68] R. B. Johnson and Anthony J. Onwuegbuzie. 2004. Mixed methods research: A research paradigm whose time has come. *Educational Researcher* 33, 7 (2004), 14–26.

[69] Tuomas Kari. 2014. Can exergaming promote physical fitness and physical activity? A systematic review of systematic reviews. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)* 6, 4 (Oct. 2014), 59–77. DOI: https://doi.org/10.4018/ijgcms.2014100105

[70] Sathya Karunananthan, Lara J. Maxwell, Vivian Welch, Jennifer Petkovic, Jordi P. Pardo, Tamara Rader, Marc T. Avey, John Baptiste-Ngobi, Ricardo Batista, Janet A. Curran, Elizabeth Tanjong Ghogomu, Ian D. Graham, Jeremy M. Grimshaw, John PA. Ioannidis, Zoe Jordan, Janet Jull, Anne Lyddiatt, David Moher, Mark Petticrew, Kevin Pottie, Gabriel Rada, Larissa Shamseer, Beverley Shea, Konstantinos C. Siontis, Naomi Tschirhart, Brigitte Vachon, George A. Wells, Howard White, and Peter Tugwell. 2020. PROTOCOL: When and how to replicate systematic reviews. *Campbell Systematic Reviews* 16, 2 (May 2020), e1087. DOI: https://doi.org/10.1002/cl2.1087

[71] Barbara Kitchenham. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University and University of Durham. EBSE Technical Report EBSE-2007-01.

[72] B. A. Kitchenham, T. Dyba, and M. Jorgensen. 2004. Evidence-based software engineering. In *Proceedings of the 26th International Conference on Software Engineering*. IEEE Computer Society, 1–9. DOI: https://doi.org/10.1109/icse.2004.1317449

[73] Barbara Kitchenham, Rialette Pretorius, David Budgen, O. Pearl Brereton, Mark Turner, Mahmood Niazi, and Stephen Linkman. 2010. Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology* 52, 8 (Aug. 2010), 792–805. DOI: https://doi.org/10.1016/j.infsof.2010.03.006

[74] Leanne M. Kmet, Linda S. Cook, and Robert C. Lee. 2004. Standard quality assessment criteria for evaluating primary research papers from a variety of fields. (2004). University of Alberta's ERA: Education and Research Archive, 1–20, DOI: https://doi.org/10.7939/R37M04F16

[75] Nancy Knechel. 2019. What's in a sample? Why selecting the right research participants matters. *Journal of Emergency Nursing* 45, 3 (2019), 332–334. DOI: https://doi.org/10.1016/j.jen.2019.01.020

[76] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. 2020. Social acceptability in HCI: A survey of methods, measures, and design strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–19. DOI: https://doi.org/10.1145/3313831.3376162

[77] Erwin Krauskopf. 2017. Call for caution in the use of bibliometric data. *Journal of the Association for Information Science and Technology* 68, 8 (May 2017), 2029–2032. DOI: https://doi.org/10.1002/asi.23809

[78] Erwin Krauskopf. 2019. Missing documents in Scopus: The case of the journal Enfermeria Nefrologica. *Scientometrics* 119, 1 (Mar. 2019), 543–547. DOI: https://doi.org/10.1007/s11192-019-03040-z

[79] Laura Krefting. 1991. Rigor in qualitative research: The assessment of trustworthiness. *The American journal of occupational Therapy* 45, 3 (1991), 214–222.

[80] Klaus Krippendorff. 1989. *Content Analysis*, Vol. 1. Oxford University Press, New York, NY, 403–407.

[81] Udo Kuckartz. 2014. *Three Basic Methods of Qualitative Text Analysis*. SAGE Publications Ltd, 65–120. DOI: https://doi.org/10.4135/9781446288719.n4

[82] Larry Laudan. 1978. *Progress and Its Problems: Towards a Theory of Scientific Growth*, Vol. 282. University of California Press.

[83] Effie L.-C. Law, Marc Hassenzahl, Evangelos Karapanos, Marianna Obrist, and Virpi Roto. 2014. Tracing links between UX frameworks and design practices: Dual carriageway. In *Proceedings of the HCI Korea (HCIK '15).* Hanbit Media, Inc., Seoul, KOR, 188–195.

[84] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation strategies for HCI toolkit research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–17. DOI: https://doi.org/10.1145/3173574.3173610

[85] Danielle Levac, Heather Colquhoun, and Kelly K. O'Brien. 2010. Scoping studies: Advancing the methodology. *Implementation Science* 5, 1 (Sep. 2010), 1–9. DOI: https://doi.org/10.1186/1748-5908-5-69

[86] Yang Li, Sayan Sarcar, Yilin Zheng, and Xiangshi Ren. 2021. Exploring text revision with backspace and caret in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3411764.3445474

[87] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* ACM, New York, NY, Article 143, 14 pages. DOI: https://doi.org/10.1145/3411764.3445488

[88] Julia Littell. 2008. *Systematic Reviews and Meta-Analysis*. Oxford University Press, Oxford, New York.

[89] Weishu Liu, Meiting Huang, and Haifeng Wang. 2021. Same journal but different numbers of published records indexed in scopus and web of science core collection: Causes, consequences, and solutions. *Scientometrics* 126, 5 (Mar. 2021), 4541–4550. DOI: https://doi.org/10.1007/s11192-021-03934-x

[90] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).* ACM, New York, NY, 3553–3562. https://doi.org/10.1145/2556288.2556969

[91] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–16. DOI: https://doi.org/10.1145/3313831.3376727

[92] Cayley MacArthur, Arielle Grinberg, Daniel Harley, and Mark Hancock. 2021. You're making me sick: A systematic review of how virtual reality research considers gender & cybersickness. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–15. DOI: https://doi.org/10.1145/3411764.3445701

[93] Christopher Marshall and Anthea Sutton. 2021. The Systematic Review Toolbox. Retrieved September 7, 2021 from http://www.systematicreviewtools.com/

[94] Marina Krnic Martinic, Dawid Pieper, Angelina Glatt, and Livia Puljak. 2019. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. *BMC Medical Research Methodology* 19, 203 (Nov. 2019), 1–12. DOI: https://doi.org/10.1186/s12874-019-0855-0

[95] Joseph A. Maxwell and Kavita Mittapalli. 2010. *Realism as a Stance for Mixed Methods Research*, Vol. 2. Sage Thousand Oaks, CA, 145–168.

[96] Tim May and Beth Perry. 2014. *Reflexivity and the Practice of Qualitative Research*, Vol. 109. Sage Los Angeles.

[97] Elisa D. Mekler, Julia Ayumi Bopp, Alexandre N. Tuch, and Klaus Opwis. 2014. A systematic review of quantitative studies on the enjoyment of digital entertainment games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, 927–936. https://doi.org/10.1145/2556288.2557078

[98] David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A. Stewart. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews* 4, 1 (Jan. 2015), 1–9. DOI: https://doi.org/10.1186/2046-4053-4-1

[99] David Moher, Deborah J. Cook, Susan Eastwood, Ingram Olkin, Drummond Rennie, and Donna F. Stroup. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *The Lancet* 354, 9193 (Nov. 1999), 1896–1900. DOI: https://doi.org/10.1016/s0140-6736(99)04149-5

[100] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman and. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6, 7 (Jul. 2009), e1000097. DOI: https://doi.org/10.1371/journal.pmed.1000097

[101] David Moher, Jennifer Tetzlaff, Andrea C. Tricco, Margaret Sampson, and Douglas G. Altman. 2007. Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine* 4, 3 (Mar. 2007), e78. DOI: https://doi.org/10.1371/journal.pmed.0040078

[102] David Moher and Alexander Tsertsvadze. 2006. Systematic reviews: When is an update an update? *The Lancet* 367, 9514 (Mar. 2006), 881–883. DOI: https://doi.org/10.1016/s0140-6736(06)68358-x

[103] Kevin J. Munro and Garreth Prendergast. 2019. Encouraging pre-registration of research studies. *International Journal of Audiology* 58, 3 (Mar. 2019), 123–124. DOI: https://doi.org/10.1080/14992027.2019.1574405

[104] Shinichi Nakagawa, Gihan Samarasinghe, Neal R. Haddaway, Martin J. Westgate, Rose E. O'Dea, Daniel W.A. Noble, and Malgorzata Lagisz. 2019. Research weaving: Visualizing the future of research synthesis. *Trends in Ecology & Evolution* 34, 3 (Mar. 2019), 224–238. DOI: https://doi.org/10.1016/j.tree.2018.11.007

[105] Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359.

[106] PRISMA Group / Ottawa Hospital Research Institute (OHRI). 2021. PRISMA Statement Website - PRISMA 2020 Abstracts Checklist. Retrieved from http://www.prisma-statement.org/Extensions/Abstracts

[107] Antti Oulasvirta and Kasper Hornbæk. 2016. HCI research as problem-solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 4956–4967. DOI: https://doi.org/10.1145/2858036.2858283

[108] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* (Mar. 2021), n71. DOI: https://doi.org/10.1136/bmj.n71

[109] Matthew J. Page, David Moher, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, David Moher. 2021. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *Bmj* 372 (2021), 1–36. DOI: https://doi.org/10.1136/bmj.n160

[110] Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52, 2 (Mar. 2015), 183–199. DOI: https://doi.org/10.1016/j.im.2014.08.008

[111] Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing reporting of participant compensation in HCI: A systematic literature review and recommendations for the field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 141, 16 pages. DOI: https://doi.org/10.1145/3411764.3445734

[112] Micah D.J. Peters, Christina Godfrey, Patricia McInerney, Hanan Khalil, Palle Larsen, Casey Marnie, Danielle Pollock, Andrea C. Tricco, and Zachary Munn. 2022. Best practice guidance and reporting items for the development of scoping review protocols. *JBI Evidence Synthesis* 20, 4 (Feb. 2022), 953–968. DOI: https://doi.org/10.11124/jbies-21-00242

[113] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64 (Aug. 2015), 1–18. DOI: https://doi.org/10.1016/j.infsof.2015.03.007

[114] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A bermuda triangle? A review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, 1–16. DOI: https://doi.org/10.1145/3173574.3174035

[115] Mark Petticrew and Helen Roberts. 2008. *Systematic Reviews in the Social Sciences: A Practical Guide*. John Wiley & Sons.

[116] Mai T. Pham, Andrijana Rajić, Judy D. Greig, Jan M. Sargeant, Andrew Papadopoulos, and Scott A. McEwen. 2014. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods* 5, 4 (2014), 371–385. DOI: https://doi.org/10.1002/jrsm.1123

[117] Henning Pohl, Andreea Muresan, and Kasper Hornbæk. 2019. Charting subtle interaction in the HCI literature. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–15. DOI: https://doi.org/10.1145/3290605.3300648

[118] Kylie Porritt, Judith Gomersall, and Craig Lockwood. 2014. JBI's systematic reviews: Study selection and critical appraisal. *AJN The American Journal of Nursing* 114, 6 (2014), 47–52. DOI: https://doi.org/10.1097/01.NAJ.0000450430.97383.64

[119] Critical Appraisal Skills Programme. 2021. CASP Checklists. Retrieved September 9, 2021 from https://casp-uk.net/casp-tools-checklists/

[120] Isabel P. S. Qamar, Rainer Groh, David Holman, and Anne Roudaut. 2018. HCI meets material science: A literature review of morphing materials for the design of shape-changing interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, 1–23. DOI: https://doi.org/10.1145/3173574.3173948

[121] Katja Rogers, Sukran Karaosmanoglu, Maximilian Altmeyer, Ally Suarez, and Lennart E. Nacke. 2022. Much realistic, such wow! A systematic literature review of realism in digital games. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, Article 190, 21 pages. DOI: https://doi.org/10.1145/3491102.3501875

[122] Katja Rogers, Sukran Karaosmanoglu, Dennis Wolf, Frank Steinicke, and Lennart E. Nacke. 2021. A best-fit framework and systematic review of asymmetric gameplay in multiplayer virtual reality games. *Frontiers in Virtual Reality* 2 (Jul. 2021). DOI: https://doi.org/10.3389/frvir.2021.694660

[123] Yvonne Rogers. 2012. HCI theory: Classical, modern, and contemporary. *Synthesis Lectures on Human-Centered Informatics* 5, 2 (2012), 1–129. DOI: https://doi.org/10.2200/S00418ED1V01Y201205HCI014

[124] Tanja Rombey, Livia Puljak, Katharina Allers, Juan Ruano, and Dawid Pieper. 2020. Inconsistent views among systematic review authors toward publishing protocols as peer-reviewed articles: An international survey. *Journal of Clinical Epidemiology* 123 (Jul. 2020), 9–17. DOI: https://doi.org/10.1016/j.jclinepi.2020.03.010

[125] Johnny Saldaña. 2013. *The Coding Manual for Qualitative Researchers* (2nd. ed.). Sage Publications, Thousand Oaks, California.

[126] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the U.S. child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–15. DOI: https://doi.org/10.1145/3313831.3376229

[127] Kara Schick-Makaroff, Marjorie MacDonald, Marilyn Plummer, Judy Burgess, and Wendy Neander. 2016. What synthesis methodology should I use? A review and analysis of approaches to research synthesis. *AIMS Public Health* 3, 1 (2016), 172. DOI: https://doi.org/10.3934/publichealth.2016.1.172

[128] K. F. Schulz, D. G. Altman, and D. Moher and. 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, mar23 1 (Mar. 2010), c332–c332. DOI: https://doi.org/10.1136/bmj.c332

[129] Hasti Seifi, Farimah Fazlollahi, Michael Oppermann, John A. Sastrillo, Jessica Ip, Ashutosh Agrawal, Gunhyuk Park, Katherine J. Kuchenbecker, and Karon E. MacLean. 2019. Haptipedia: Accelerating haptic device discovery to support interaction & engineering design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3290605.3300788

[130] Beverley J. Shea, Jeremy M. Grimshaw, George A. Wells, Maarten Boers, Neil Andersson, Candyce Hamel, Ashley C. Porter, Peter Tugwell, David Moher, and Lex M. Bouter. 2007. Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology* 7, 1 (Feb. 2007), 1–7. DOI: https://doi.org/10.1186/1471-2288-7-10

[131] Beverley J. Shea, Barnaby C. Reeves, George Wells, Micere Thuku, Candyce Hamel, Julian Moran, David Moher, Peter Tugwell, Vivian Welch, Elizabeth Kristjansson, and David A. Henry. 2017. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* (Sep. 2017), j4008. DOI: https://doi.org/10.1136/bmj.j4008

[132] Andy P. Siddaway, Alex M. Wood, and Larry V. Hedges. 2019. How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology* 70, 1 (Jan. 2019), 747–770. DOI: https://doi.org/10.1146/annurev-psych-010418-102803

[133] Valerie Smith, Declan Devane, Cecily M. Begley, and Mike Clarke. 2011. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology* 11, 15 (Feb. 2011). DOI: https://doi.org/10.1186/1471-2288-11-15

[134] Kai Standvoss, Vartan Kazezian, Britta R. Lewke, Kathleen Bastian, Shambhavi Chidambaram, Subhi Arafat, Ubai Alsharif, Ana Herrera-Melendez, Anna-Delia Knipper, Bruna M. S. Seco, Nina Nitzan Soto, Orestis Rakitzis, Isa Steinecker, Philipp van Kronenberg Till, Fereshteh Zarebidaki, and Tracey L. Weissgerber. 2022. Taking shortcuts: Great for travel, but not for reproducible methods sections. arXiv:2022.08.08.503174. Retrieved from https://doi.org/10.1101/2022.08.08.503174

[135] Evropi Stefanidi, Marit Bentvelzen, Paweł W. Woźniak, Thomas Kosch, Mikołaj P. Woźniak, Thomas Mildner, Stefan Schneegass, Heiko Müller, and Jasmin Niess. 2023. Literature reviews in HCI: A review of reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, New York, NY, Article 509, 24 pages. DOI: https://doi.org/10.1145/3544548.3581332

[136] Crystal N. Steltenpohl, Hilary Lustick, Melanie S. Meyer, Linsday E. Lee, Sondra M. Stegenga, Laurel S. Reyes, and Rachel L. Renbarger. 2023. Rethinking transparency and rigor from a qualitative open science perspective. *Journal of Trial and Error* 4, 1 (May 2023), 1–13. DOI: https://doi.org/10.36850/mr7

[137] Elizabeth Stowell, Mercedes C. Lyson, Herman Saksono, Reneé C. Wurth, Holly Jimison, Misha Pavel, and Andrea G. Parker. 2018. Designing and evaluating mhealth interventions for vulnerable populations: A systematic review. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, 1–17. DOI: https://doi.org/10.1145/3173574.3173589

[138] Donna F. Stroup. 2000. Meta-analysis of observational studies in epidemiology - A proposal for reporting. *JAMA* 283, 15 (Apr. 2000), 2008. DOI: https://doi.org/10.1001/jama.283.15.2008

[139] Anthea Sutton, Mark Clowes, Louise Preston, and Andrew Booth. 2019. Meeting the review family: Exploring review types and associated information retrieval requirements. *Health Information & Libraries Journal* 36, 3 (Sep. 2019), 202–222. DOI: https://doi.org/10.1111/hir.12276

[140] Paige L. Sweet. 2020. Who knows? Reflexivity in feminist standpoint theory and bourdieu. *Gender & Society* 34, 6 (Nov. 2020), 922–950. DOI: https://doi.org/10.1177/0891243220966600

[141] James Thomas and Angela Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* 8, 1 (2008), 45. DOI: https://doi.org/10.1186/1471-2288-8-45

[142] Allison Tong, Kate Flemming, Elizabeth McInnes, Sandy Oliver, and Jonathan Craig. 2012. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Medical Research Methodology* 12, 1 (Nov. 2012). DOI: https://doi.org/10.1186/1471-2288-12-181

[143] A. Tong, P. Sainsbury, and J. Craig. 2007. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* 19, 6 (Sep. 2007), 349–357. DOI: https://doi.org/10.1093/intqhc/mzm042

[144] Marta Topor, Jade S. Pickering, Ana Barbosa Mendes, Dorothy V. M. Bishop, Fionn Büttner, Mahmoud M. Elsherif, Thomas R. Evans, Emma L. Henderson, Tamara Kalandadze, Faye T. Nitschke, Janneke P. C. Staaks, Olmo R. Van den Akker, Siu K. Yeung, Mirela Zaneva, Alison Lam, Christopher R. Madan, David Moreau, Aoife O'Mahony, Adam J. Parker, Amy Riegelman, Meghan Testerman, and Samuel J. Westwood. 2023. An integrative framework for planning and conducting non-intervention, reproducible, and open systematic reviews (NIRO-SR). *Meta-Psychology* 7 (Jul. 2023), 1–14. DOI: https://doi.org/10.15626/mp.2021.2840

[145] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D. J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine* 169, 7 (Oct. 2018), 467–473. DOI: https://doi.org/10.7326/m18-0850

[146] Andrea C. Tricco, Jennifer Tetzlaff, and David Moher. 2011. The art and science of knowledge synthesis. *Journal of Clinical Epidemiology* 64, 1 (Jan. 2011), 11–20. DOI: https://doi.org/10.1016/j.jclinepi.2009.11.007

[147] Raphael Velt, Steve Benford, and Stuart Reeves. 2017. A survey of the trajectories conceptual framework: Investigating theory use in HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, 2091–2105. DOI: https://doi.org/10.1145/3025453.3026022

[148] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. 2007. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ* 335, 7624 (Oct. 2007), 806–808. DOI: https://doi.org/10.1136/bmj.39335.541782.ad

[149] Marilyn D. White and Emily E. Marsh. 2006. Content analysis: A flexible methodology. *Library Trends* 55, 1 (2006), 22–45.

[150] Penny Whiting, Jelena Savović, Julian P. T. Higgins, Deborah M. Caldwell, Barnaby C. Reeves, Beverley Shea, Philippa Davies, Jos Kleijnen, and Rachel Churchill. 2016. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology* 69 (Jan. 2016), 225–234. DOI: https://doi.org/10.1016/j.jclinepi.2015.06.005

[151] Paul Woodcock, Andrew S. Pullin, and Michel J. Kaiser. 2014. Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: A methodology. *Biological Conservation* 176 (Aug. 2014), 54–62. DOI: https://doi.org/10.1016/j.biocon.2014.04.020

[152] Qiushi Zhou, Cheng Cheng Chua, Jarrod Knibbe, Jorge Goncalves, and Eduardo Velloso. 2021. Dance and Choreography in HCI: A Two-Decade Retrospective. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, Article 262, 14 pages. DOI: https://doi.org/10.1145/3411764.3445804

[153] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An analysis and critique of *research through design*: Towards a formalization of a research approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)*. ACM, New York, NY, 310–319. DOI: https://doi.org/10.1145/1858171.1858228

Received 11 July 2022; revised 12 April 2024; accepted 1 May 2024