The Systematic Review-lution: A Manifesto to Promote Rigour and Inclusivity in Research Synthesis

Katja Rogers
k.s.rogers@uva.nl
University of Amsterdam
Amsterdam, Netherlands

Katie Seaborn
seaborn.k.aa@m.titech.ac.jp
Tokyo Institute of Technology
Tokyo, Japan

Figure 1: Generated images by Dall-E (https://openai.com/dall-e-2/) based on the prompt "a robot drowning among waves of papers, number, and letters, drawn by Monet," visualizing the way it sometimes feels to be a researcher in 2022.

ABSTRACT

The field of human-computer interaction (HCI) is maturing. Systematic reviews, a staple of many disciplines, play an important and often essential role in how each field contributes to human knowledge. On this prospect, we argue that our meta-level approach to research within HCI needs a revolution. First, we echo previous calls for greater rigour in primary research reporting with a view towards supporting knowledge synthesis in secondary research. Second, we must decide as a community how to carry out systematic review work in light of the many ways that knowledge is produced within HCI (rigour in secondary research methods and epistemological inclusivity). In short, our manifesto is this: we need to develop and make space for an inclusive but rigorous set of standards that supports systematic review work in HCI, through careful consideration of both primary and secondary research methods, expectations, and infrastructure. We call for any and all fellow systematic review-lutionaries to join us.

CCS CONCEPTS

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

alt.chi, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3582733

KEYWORDS

research synthesis, systematic review, rigour, literature, epistemology

ACM Reference Format:
Katja Rogers and Katie Seaborn. 2023. The Systematic Review-lution: A Manifesto to Promote Rigour and Inclusivity in Research Synthesis. In *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3544549.3582733

> "*We have reason to fear that the multitude of books which grows every day in a prodigious fashion will make the following centuries fall into a state as barbarous as that of the centuries that followed the fall of the Roman Empire.*"
>
> Adrien Baillet, 1685 (as quoted by Ann Blair) [9]

WHY RESEARCH SYNTHESIS MATTERS

Those of us in human-computer interaction (HCI) are publishing a lot—dare we say *too much*? This is exemplified by the number of papers in the ACM Digital Library (DL) (Figure 2). One person or even a team cannot keep up with the sheer amount of research that we produce. Yet we also need to synthesize this literature. In some fields, the enthusiastic pace of research output has led some researchers to explore automation. Modern computing technologies like machine learning [6, 53, 82] might be able to aid us in the task of keeping up with and synthesizing publications. But is this necessary or desired in HCI, or do we need to rethink the knowledge production and reporting process?

In other large(r) fields, Chu and Evans [18] warn that increased publication output can lead to "*ossification*" because novel ideas

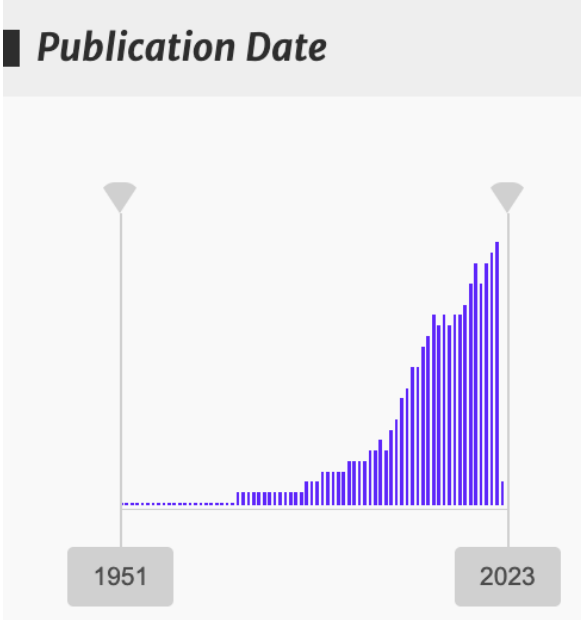

Figure 2: Results per year for the keyword "human-computer interaction" in the ACM Digital Library. A total of 658,883 results were found as of 14:24 on February 7th, 2023. The year 2000 alone featured 8,519 results, with 27,009 for 2010; 35,183 for 2020; and 37,215 results for 2022.

cannot gain traction against the entrenched canon of the papers most often cited. This can have severe consequences for the field as a whole: "*too many papers published each year can lead to stagnation rather than advance [knowledge creation]*" [18]. There are already hints of this trend in HCI: the papers most cited are cited quite a bit more often than the average paper, while the number of citations papers receive per year is declining overall [66]. Based on other fields, this suggests that it is getting more difficult for new ideas to break through and shake up established ones in HCI. However, in HCI, we might actually have the opposite problem (too) because of the *kinds* of papers we publish at this rapid pace. In recent years, some researchers have become concerned that we are focusing too much on novelty [5, 62]. Consider the most recent (2022) proceedings of the Conference on Human Factors in Computing Systems (CHI), where searching for "novel" yields 602 results—based on 697 papers. When 86% of papers are characterized as "novel," we need to ask what this means for knowledge gains and consensus-building. Playing devil's advocate, we might say that our field incentivizes, if not requires, the publication of a never-ending stream of flashy one-offs. Instead of putting effort into rigorous incremental research to confirm evidence across multiple studies, we find ourselves dancing after the novelty carrot[1].

The repercussions of this state of affairs deserve careful and critical attention. Focusing so strongly on novelty may be part of what makes it difficult to provide definitive answers about what we actually know so far in HCI [41]. This also makes it difficult for HCI to situate itself within and participate alongside other fields of study, and limits the kind of research that we do. We echo DiSalvo et al. [24]'s statement on sustainable HCI as relevant for HCI as a whole: "*[t]o avoid reinventing the wheel, there is a need for the field to take stock of what is known and to identify major unknown questions or issues, which arise from what has been established, as a basis for future work.*" Whittaker et al. [84] raised similar criticisms: that the HCI community "*[overemphasizes] "radical invention" at the price of achieving a common research focus.*" They go on to point out that in the absence of "*such a focus, it is difficult to build on previous work, to compare different interaction techniques objectively, and to make progress in developing theory.*" This is not merely a problem of research praxis; it also has practical ramifications when we, as a field of study, cannot provide clear guidelines or implications. We find ourselves in a liminal space, where we all are carrying out research and producing a variety of outcomes, but an outsider looking in may find the overall picture difficult to make out. Such an outsider may then move on to a more clearly defined space, leaving our work unacknowledged and overlooked. From within, we may not be able to see the forest for the trees, leaving no clear path forward. Research synthesis can clarify the work conducted in a field of study, not only for others but ourselves, as well.

Yet replication studies, follow-up work, corrections and expansions, and other forms of explicitly *not novel* forms of inquiry remain sidelined, despite calls to action that go back more than a decade [25, 42, 55, 85, 86]. This has grave implications in light of larger patterns and hiccups in research practice, including p-hacking [38], the replication [2, 72] and publication bias [60] crises, and adverse effects resulting from the preprint server explosion [1]. This is not only a problem for experimental or quantitative work typically housed within positivist frameworks. We recognize that not all research projects within HCI aim for generalization or consensus. Still, many if not all of them hold valuable insights on their own that could be productively synthesized. All epistemological and methodological lenses should be embraced in knowledge synthesis work if we wish to provide a full picture of HCI research.

Globalized computer-based information technology—the very heart of our discipline—has created new drivers and tensions for scholarship. Solutions to this phenomenon might be found by embracing *slow science* [77], to some extent. Yet even if we were to stop publishing altogether tomorrow—and pull the plug on the Internet—we would still need to sift through and synthesize the existing work published so far. Many of us in HCI are taking up this task, but have little guidance or standardization. This is not because guidelines or standards do not exist—they do [13, 37, 45, 63, 67, 73, 75, 79, 80]. However, these are premised around the work developed (and valued) in other fields, e.g., randomized-controlled trials in the medical field [76]. Part of the challenge inherent to HCI is the *sheer variety* of work available (perhaps even leading to what Reeves [69] and Fallman and Stolterman [27] refer to as "*disciplinary anxiety*"). Closely related to this, another key challenge is the lack of consensus on *how* to carry out research synthesis in general, and systematic reviews more specifically: our field has not yet embarked on an explicit conversation about what we expect from systematic reviews, nor how to handle the different kinds of knowledge our field produces.

This is our manifesto. We propose to begin a community-driven conversation to determine how to depart from our typical research praxis to support research synthesis at the meta level. We argue that

[1]Keeping in mind that our field does not necessarily agree on one meaning of "novelty."

a rigorous and inclusive systematic review approach to research synthesis in HCI is the way forward. We pose two critical questions at this juncture: *1) How can we package our work in such a way that meaningful research synthesis can be practiced based on the wonderful diversity of work that we produce in CHI and adjacent spaces?* Secondly, in light of the many forms of knowledge produced in our field, we believe and hope to convince the reader that we must all come together as a community to develop a shared set of research practices for planning, conducting, and reporting research synthesis within HCI: *2) What should research synthesis look like when it is grounded in plurality: quantitative studies, qualitative studies, design research, ethnography, and the development of interactive artifacts and systems?* We raise these concerns, challenges, and desires for a different way in the scholarly tradition of denouncing the "*confusing and harmful abundance*" of literature, a form of self-reflective discourse that dates back several centuries [9]. Our goal is to remind ourselves about the bigger picture and re-orient each other as members of a community of practice.

With the above, we make the case for research synthesis, why it matters for the field of HCI, and why our answers to this issue may differ from other fields. We next present established methodologies for research synthesis, focusing on the current global standard across most fields of study: the systematic review. We then raise critical issues about relying on a systematic review approach for HCI research, and provocations that anchor to its abundance in topics, methods, and epistemological points of view. We end with a call for action on a novel framing of research synthesis: inviting you, dear reader, and everyone participating in HCI research. We aim to start a conversation at this year's conference that we expect will lead into the development of a future workshop, special interest group, committee and/or collaborations with the goal of establishing a community of practice invested in HCI research synthesis. By gathering the multitude of disciplinary voices and epistemological perspectives in our community, we hope to make a disciplinary impact in terms of knowledge creation and methodology that reverberates back to the larger research community.

LAUNCHING PAD: WHAT IS A SYSTEMATIC REVIEW? OR: "BUT I HAVE A PRISMA FIGURE, SURELY THAT MAKES IT SYSTEMATIC"

Systematic reviews are a staple of research synthesis in many fields of study. In medicine, they are considered "*indispensable*" [44] as the "*gold standard*" [59] means of arriving at consensus across individual studies on a specific topic, typically an intervention of some kind, so as to enable decision-making grounded in evidence-based work [44]. Yet there is no clear agreed-on definition for what a systematic review is as an outcome and what it should entail as a process [56]. As Martinic et al. [56] explain, "*definitions of [systematic reviews] are vague and ambiguous, often using terms such as* clear, explicit *and* systematic*, without further elaboration.*" This presents those of us in HCI with a challenge and an opportunity: we may struggle to understand and apply systematic review methodologies to our work, even though we have much to learn from other disciplines in the history and practice of the systematic review. Yet perhaps we do not need to adopt all of these methods as they are, and in some cases perhaps it would even be inappropriate to try; we may instead chart a new path forward.

The first step towards understanding and productive deviation is a definition. Let us begin with what a systematic review is not: It is not merely a description of previous work on a certain topic, within a field of study, or around a certain research question or hypothesis. It is not an annotated bibliography in which we comment on papers that we have read. It is also not conducted ad hoc without an *a priori* plan, especially not if the procedure was changed mid-process so that, in the end, the research question had to be adjusted to fit the method. It is not the summary alone; it cannot be without a description of how the results were constructed, magicking its way from search process to outcomes with no in-between. It is not a narrative of a curated selection of works.

Then, what might a systematic review *be*? In other words, what is its nature as an outcome and method of scholarship? We start with a pair of concepts: primary and secondary research—their relationship, and a basic distinction between the two. Primary research is the *material* of a systematic review. We define *primary research* as any paper[2] that reports directly on collected and analyzed data, e.g., a paper reporting a user study. *Secondary research*, then, is one step removed: a paper that reports on a collected and analyzed sample of primary research papers: systematic reviews are one example of secondary research. We see no issue in carrying these concepts forward for research synthesis within HCI.

With these foundational concepts in hand, how then do we process the material that is primary research into the outcome that is secondary research? Unfortunately, the structure of a typical systematic review process is more contested than ideal [56]. Martinic et al. [56] suggest the following components, listed in procedural order: "*i) a research question; ii) sources that were searched, with a reproducible search strategy (naming of databases, naming of search platforms/engines, search date and complete search strategy); iii) inclusion and exclusion criteria; iv) selection (screening) methods; v) [a critical appraisal of] the quality/risk of bias of the included studies; vi) information about data analysis and synthesis that allows the reproducibility of the results.*" Haddaway and Bilotta [36] instead compare requirements posed by institutions that promote evidence-based research through systematic reviews, e.g., the Cochrane Collaboration. They suggest three basic standards: "*(i) [...] methods should be described in sufficient detail to allow full repeatability and traceability; (ii) [...] a systematic approach to identifying and screening relevant academic and grey literature, and (iii) [...] critical appraisal of the validity (quality and generalisability) of included studies to give greater weight to more reliable studies.*" With the plurality of our field in mind, we draw out the following general characteristics: (i) an *a priori* developed and pre-registered protocol i.e., full documentation of the planned review procedure, as well as clearly and comprehensively articulated research questions, search processes, screening processes, data extraction processes, and the means of quality appraisal (or a rationale for its omission); (ii) data analysis and synthesis methods; and (iii) a discussion that transparently

[2]While we use "paper" here, we do not mean that the paper itself is "the research." We use "paper" for simplicity in writing and in recognition that most output of scholarship is packaged in paper form, particularly in the context of systematic review work and research synthesis.

addresses limitations in both search and synthesis, and for both method choices and results.

The components necessary for a review to be "systematic" remains an open question in HCI. Is pre-registration necessary, especially if a similar registration already exists? Does every review require an assessment of quality of the primary research? Are certain tools or platforms required, such as the use of the ACM Digital Library or IEEE Xplore, which are foundational for primary and secondary research publishing but not without their quirks and outright glitches? Further, the specifics of *how* the steps should be conducted in practice are similarly unclear and in disarray. As one example in HCI, there is no consensus or even weighing in on the trade-offs for the choice between single and double screening—is one person's decision enough, or is at least one other required? Can the other(s) simply review rejected items, i.e., to avoid false negatives? Can the work be divided up between different people? Should there be a "storming and norming" process to get people on the same page or even some form of inter-rater reliability metric? Do we let go of generalizability and accept epistemological diversity? Should we adopt the aspirations to be practical and flexible and simply transparent, as advocated by Braun and Clarke [10] in their *reflexive* thematic approach? On that note, *can* we meaningfully and appropriately draw from other methodologies to inform research synthesis? These are just some of many methodological questions that researchers in other fields have been exploring in recent years [32, 52], yet each one alone already raises an array of questions and provocations for the context of HCI research.

> **Systematic Review.** /sɪstə'matɪk rɪ 'vjuː/.
> `«DefinitionError: term 'systematic review' is not defined».`
>
> *Human-Computer Interaction, Probably*

The systematic review in its modern form can primarily be traced back to the medical field, where the goal is to synthesize the results of multiple randomized controlled trials to better estimate the effect of a specific intervention [76]. When the effect sizes in very similar studies are synthesized via statistical methods, it is considered a systematic review with meta-analysis [22]. The term "meta-analysis" is sometimes used in HCI to refer to review work without statistical aggregation of effect sizes, presumably in a more literal interpretation of the term "meta" to account for a paper that reports on one or more analyses, e.g., [20, 83]. Other fields have adjusted synthesis methodology or created their own to suit their needs, for example fields and subfields that do not conduct (m)any randomized controlled trials [79, 80]. This parallel methodological evolution in multiple fields has led to a dizzying array of closely related but different synthesis methods and review types: scoping review, rapid review, mapping review, review of reviews, (best-fit) framework synthesis, mixed-method synthesis, among many others [78]. Uptake of these methods as well as guidelines for their usage varies wildly, as do opinions on which of these are or are not "systematic." As these fields have matured, they have started to face another flood of papers, this time with secondary rather than primary research [44]. The waterfall does not stop at the pond, but cascades ever further: for a while already, academic research literature has featured tertiary research [4, 23, 74] and even occasional examples of "quarternary" research [58]. We have no reason that this will not be the case for HCI as well; now is the time to act and seek a new path forward.

INPUT COORDINATES: TRACING OUT OPEN QUESTIONS AND COUNTERING OBJECTIONS

The alt.chi website expects submissions to be "*controversial, risk-taking, and boundary pushing*" [15]—so why are we writing about systematic reviews, when they are an established methodology, even a gold standard, that can highlight existing knowledge ("backward-looking" [61]) as well as create new forms of knowledge from what came before ("forward-looking" [61])? Surely this is not a controversial topic? Yet somehow it is: in our own experience when submitting and reviewing papers in HCI, we have come across a broad range of expectations and opinions about:

- whether systematic reviews, as a form of secondary research that heavy relies on primary research, has a place in HCI, since such work does not always lead to a novel outcome in the traditional sense;
- what systematic reviews are for *(providing an objective and comprehensive overview of a subfield vs. providing an opinionated narrative vs. providing an estimation that answers a very specific question; establishing consensus vs. providing a subjective but substantiated perspective)*,
- how they should be conducted *(based on a range of specific guidelines; ignoring or including qualitative research; with or without meta-analysis; with or without critical appraisal or double screening or data extraction forms or ...)*,
- what forms of knowledge they can and should produce *("maps" vs. synthesized effect size estimates vs. taxonomies, theories or frameworks vs. new research questions and directions vs. new primary research or instruments or prototypes)*, and
- basic terminology and definitions *(when should a review be considered systematic; what is a meta-analysis; etc.)*

Let us invoke an imaginary HCI researcher, who sees no benefit to systematic reviews and considers them procrustean:

> **Procrustean.** /prə(ʊ)'krʌstɪən/. Of, relating to, or resembling the practices of Procrustes (see Procrustes n.); (hence) enforcing uniformity or conformity without regard to natural variation or individuality.
>
> *Oxford English Dictionary*

As we have outlined above, there are reasons to come into such a position within HCI, so we give this perspective a platform and trace out likely concerns. This researcher might reasonably ask: Will systematic reviews lead to no one reading the original papers anymore? We again emphasize that it is generally not possible to stay up to date and read all papers in HCI. Sorry. That ship has sailed. Yet it may be too simplistic and disillusioned to respond that "nobody reads anything anyway"—even though it seems that we do not engage with cited work as critically and comprehensively as we should [54]. Systematic reviews *could* indeed shift or divert citations from primary research papers. Reviews are easy to cite for

general overview purposes, and without systematic reviews, the same authors might cite a couple of hand-picked primary research papers instead. However, researchers tackling a particular topic or carrying out work within the same domain will still cite the most relevant papers directly—or should. Still, we acknowledge that an increase systematic reviews might affect citation practices, especially if we consider *who* is writing them (and who is not) as well as *how*: "*citations have politics*" [19]. As noted by Kumar and Karusala: "*How work is written about also matters because it can distort or even erase contributions over time*" [46]. However, a well-conducted systematic review should gather and give platform to a broad and unbiased selection of papers grounded in a comprehensive search strategy and self-reflective quality assessments. It could thus help to reduce biases in how we cite and pay attention to existing research, i.e., be self-correcting in the same spirit as the scientific method. Rather than encouraging us to cite what (or who) we know, which may not represent the diversity of the field but rather our social networks [31], systematic review procedures can broaden our horizons and create greater inclusion in citation practice. Further, a well-conducted systematic review is itself a form of in-depth critical reflection and engagement with the primary research in its corpus. While it may "steal" some citations, it should itself cite the primary work and likely also elicit future citations for it going forward[3].

Our imaginary researcher might next ask: Will systematic reviews lead us to enforce a procrustean norm in our synthesized results that entirely ignores all the beautiful variations in each of the individual papers? This may be true. But maybe such variation does not always help us with our goal in the moment. When seeking a good (enough) answer to a specific question based on the field's currently available research, perhaps those variations are not always useful or relevant at a meta level. In fact, extending the metaphor of Procrustes to user studies can show why these objections should not be an issue. As a field of study, we generally do not shirk the individual user when drawing on the results of a n=30 user study to infer how it might work for the user group as a whole. By posing implications and conclusions about a specific research question based on a single user study, we are not aiming to define or enforce a norm that ignores the beautiful variations of the individual participants. Rather, we offer a slice of the available experience with the resources at our disposal and then turn to other methods to explore and showcase the variations we could not get to in one study. Similarly, we can combine systematic reviews with specific methods of analysis to draw conclusions, and gain nuance and rich, situated understanding.

Finally, we turn towards Blackwell's perspective on HCI as a field of study: Perhaps the goal of HCI should *not* be to "*develop and maintain a stable body of knowledge, but rather to be the catalyst or source of innovation*". This would instead require that we as HCI researchers engage in scholarship that is "*questioning, provocative, disruptive and awkward*" [8]. This could be an argument against systematic reviews, as the goal of synthesis is often to stabilize and find firm ground in the shifting sands of our field. Still, Blackwell [8] also emphasize the importance of "*reflective practice*"—which itself is something that knowledge synthesis through systematic reviews can deliver and structure. We suggest that systematic reviews, with all of their own methodological diversity, have the potential to be part of both the development of stable ground *and* disruptive practice within knowledge production in HCI.

[3]We might also question when citations are truly meaningful or useful, given that they can just as much indicate social power differentials as scholarly engagement. Systematic reviews could help us dodge our natural inclination as social animals towards popularity metrics, as operationalized in citation counts.

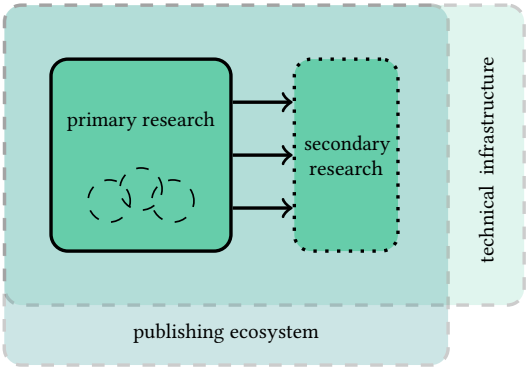

Figure 3: The academic ecosystem consists of primary research based on a broad range of different methods and research paradigms, which can then be synthesized in secondary research. All research is created and shaped by the academic publishing ecosystem—e.g., venues, reviewers, and conference (sub)committees—and technical infrastructure—e.g., databases and their search functions.

CHARTING A NEW TRAJECTORY: CRITICAL ISSUES AND PROVOCATIONS

Based on our experiences of planning, conducting, and publishing several systematic reviews (as well as some less than systematic ones, according to our current understanding of the term), we here present critical issues and provocations that HCI as a field needs to grapple with in the journey to answer the questions and calls raised by this manifesto. These concern the knowledge-building ecology surrounding primary and secondary research in HCI as shaped by the epistemological diversity within the HCI research community. However, we also foresee friction in supporting secondary research within the HCI publishing ecosystem (expectations and requirements of conference proceedings and journals, their reviewers, editors/associate chairs, and subcommittees) and technical infrastructure (e.g., the available and relevant databases and their search functions)—see Figure 3. We map these out next.

Primary Research Reporting. The foundation of research synthesis generally, and systematic reviews specifically, is built upon the reporting of primary research. Yet empirical work in HCI—viewed across the field as a whole—is disparate; how we report is varied and sometimes spotty. This is certainly not a novel criticism; we echo calls from other researchers who are critically reflecting on our research reporting practices, e.g., with regards to the reporting of race and ethnicity data [14], brain signal experiment data [68], participant compensation data [65], inter-rater reliability in

qualitative research [57], specific measures [71] and questionnaires [43, 47], engagement with self-determination theory [81], artifact descriptions [33], and inferential statistics [12], to list just a few. These issues may arise in part due to page limits or efforts to ensure paper length matches perceived contribution, but may also be due to lack of community-driven standardization and education.

This complicates research synthesis in secondary research because it makes results difficult to compare and weigh. Again, we recognize that this may not always be the goal, but it often is in the HCI world. Yet how can we point to what works and what does not if we cannot synthesize results with a high degree of rigour or systematicity? The good news is that using existing guidelines for reporting more will likely also help with secondary research simply by making the reported primary research more comprehensive and comparable. Still, it may be worth examining to what extent existing guidelines for reporting primary research can support follow-up secondary research.

Further, given that criticism of reporting in HCI has been expounded for many years, perhaps it is time to consider pointing authors and reviewers in HCI to such guidelines more explicitly. There are already hints of conferences adding a bit more structure to the submission process. For example, since 2021, at least one ACM conference has required authors to indicate "*the primary and secondary contribution type of their paper*" (empirical-qualitative; empirical-quantitative; empirical-mixed-methods; artifacts-technical; artifacts-design; theoretical; or meta-research[4]), to assist with reviewer fit to assigned papers [17]. We could add a requirement for papers to include a structured abstract of sorts as supplementary material—tailored to the contribution type. This could support not only future secondary research, but also the reviewing process itself, by providing a concise overview of the conducted work that reviewers can quickly and easily digest.

Epistemological Diversity of CHI. We next address the role that our different ways of knowing in HCI play in the synthesis of primary work. This needs to be considered both from the perspective of the diversity of research *within* primary work in HCI, but also in the diversity of methods that we draw on for synthesizing it in secondary research.

How do we approach any kind of formalization of systematic reviews for as broad a field as HCI? Research in other fields has in recent years looked at the methodology of mixed methods reviews in more detail [40, 70]. To conduct a review that accommodates research results from quantitative as well as qualitative methods, we can point to the Joanna Briggs Institute (JBI) guidelines for mixed methods systematic reviews [50] as a starting point that covers some approaches. Currently there are only a few reviews in HCI that use these or related guidelines. Still, we think it deserves more attention from our mixed methods-inclined field of study and could greatly benefit the way we do synthesis.

But HCI also features approaches that are situated more in design research methods, like participatory design and research through design. Wolf et al. [88] describe the field as featuring as an "*inherent tension [that is] reflected in the distinctive practices and disciplinary orientations of engineering and creative design.*" We agree and also highlight that this "*is not an insurmountable conflict [...] both perspectives are valid*" [88]. Excluding the knowledge created through design research methods when we do synthesis decentres a significant section of our field and prevents us from accessing a truly full picture of HCI practice. However, to our knowledge, there are currently no methods or guidelines designed to handle and synthesize evidence and knowledge created through design research methods. We may have to develop new methods to integrate this work into systematic reviews. We call on researchers familar with each approach: "*any notion of rigour has to be developed within a 'firm understanding of the particular purpose of each approach'*" [30] (Frauenberger et al. [30] citing Fallman and Stolterman [27]).

There is no short supply of work in these fields for exploring what rigour means within different HCI approaches and how to evaluate such work for synthesis purposes. For example, Wolf et al. [88] outlines qualities in design praxis that aim to achieve "*design rigor*", among them the design critique: "*a designer's reflective, evaluative and communicative explanation of her design judgments and the activities in which she has engaged*". Similarly, Zimmerman et al. [89]'s criteria or lenses for evaluating interaction design research (process, invention, relevance and extensibility) may be a useful tool for synthesis. For approaches like participatory design, Frauenberger et al. [30] write about how traditionally positivist understandings of rigour need to be re-interpreted: "*accountability and rigour in a post-modern scientific context is delivered through debate, critique and reflection*", and make the case for "*acknowledging different ways of knowing*" [30]. We extend this argument: not only do we need to acknowledge these different ways of knowing, we need to develop methods of synthesizing and integrating different ways of knowing, as well.

Secondary Research Reporting. Perhaps the closest match for existing methodological guidelines towards which synthesis guidance in HCI could be oriented are efforts within software engineering (e.g., Kitchenham and Charters [45]'s work), qualitative health research (e.g., Tong et al. [79]'s ENTREQ, or Cooke et al. [21]'s SPIDER framework), and quite recently, Topor et al. [80]'s NIRO-SR for non-intervention studies (still a preprint). However, we strongly believe that HCI will need to also draw on synthesis methods that more explicitly *combine* quantitative and qualitative work: To quote Reeves [69], we need "*more reviews of and reflections upon the landscape of different forms of reasoning in HCI and through this better ways of managing how potentially competing disciplinary perspectives meet together.*" Guidance for pulling together evidence from different disciplines and methodologies *does* exist (e.g., [50]) although it is rare. Yet how well this works in HCI is an open question; currently, there is essentially no uptake in our field.

When existing systematic reviews at CHI cite a guideline for their method, they primarily reference PRISMA [63] (e.g., [7, 39, 51]). The PRISMA figure, specifically, is popular, as it can be a great way to illustrate the search and screening process. However, the PRISMA guidelines as a whole were made for reporting systematic reviews and meta-analyses of intervention studies in the medical field [63]. Most HCI reviews—even if they state that they followed the PRISMA guidelines!—do not actually answer all (or even most) of the PRISMA checklist items [63, 64]. For example, how often have you seen a systematic review at CHI report a quality assessment and/or risk of

[4]Adapted from Wobbrock and Kientz [87]'s classification of research contributions in HCI.

bias in each study[5] or certainty in the evidence as a whole[6]? Further, because of the medical world's focus on meta-analyses, several PRISMA items are designed for statistical synthesis methods that reviews in HCI only very rarely employ (e.g., explorations of causes of statistical heterogeneity[7]), and are thus simply not applicable to the kinds of reviews that we (can) conduct. The PRISMA figure may be useful, but the guidelines are, for the most part, not actually appropriate for our field—at least not past the search procedure when it comes to the synthesis methods at the heart of the review.

A quick search for "systematic review" in the ACM DL shows a sharp increase in systematic reviews being produced. This means that now is an important moment to **STOP** and reflect on the methods we use for systematic reviews in our field. We need to figure out what we mean when we use the term "*systematic*" in the context of review work, and what we expect in terms of best practices. We need to report methods clearly and comprehensively, including how we adapted guidelines to our own use. We need to look more deeply into synthesis methods and carefully choose, name, and rationalize our choices. We may also want to look into structured abstracts as supplementary materials for secondary research (for example, Haddaway et al. [37]'s ROSES could either be borrowed directly or adapted for HCI research).

Oulasvirta and Hornbæk [62] put forth that HCI needs more "*conceptual contributions that link empirical findings and the design of technology*" to make our research findings actionable and create "*integrative types of knowledge.*" We argue that by putting effort into developing and upholding guidelines and standards for review synthesis and its reporting that works for HCI specifically, we will be able to improve the conceptual contributions that HCI can make as a field. If we view HCI as a field defined by its problem-solving capacity [62], then systematic reviews—when done rigorously—can directly help to improve several of the criteria they propose as important for problem-solving: it can help us develop a better understanding of how well solutions *transfer* and inform our degree of *confidence* in them.

Venues and Subcommittees. Marshall et al. [55] lamented that HCI has few explicit publication formats that invite critical discussion: "*none of the major [venues] have any format for critical response to published articles [...] once a piece of HCI work is in publication, it is unlikely to attract any critical discussion.*" Critical discussion instead is more likely to take place in social media, Slack workplaces and Discord channels, and other unofficial venues. Systematic reviews could perform the function of critical discussion in a rigorous and formalized way, accessible to the community of practice as a whole. Yet there is no clear place for them, either. Perhaps the only publication venue in HCI that explicitly welcomes reviews ("*survey papers*") is the ACM Computing Surveys (CSUR) journal, but they make no mention of systematicness in their author guidelines [3]. CHI as the "*flagship conference of the discipline*" [49] features only one subcommittee—Health—that mentions (systematic) reviews as a method in their description [16]. Even subcommittees that describe themselves as "*epistemologically pluralistic [and] welcoming of a range of perspectives, approaches, and contributions*" [16] can recruit associate chairs and reviewers that do not consider systematic reviews as a methodology per se and may be inclined to reject them for that reason alone. Reviewers in HCI as a whole have wildly different expectations and methodological expertise when it comes to reviews; a little more agreement would go a long way.

A perception of systematic reviews not producing "*novel*" work may be a partial reason for this issue. For example, the TOCHI journal warns that they "*rarely publish[...] survey papers unless they offer a major original contribution.*" We note that reviews *absolutely* can produce original contributions based on the synthesis, e.g., intermediate-level knowledge like taxonomies [11]. When it should be considered "*major*", and whether or when a systematic overview of existing work should be considered an "*original contribution*" is something that might be helpful for TOCHI to describe in more detail for potential authors, and indeed something that we should discuss as a field.

Infrastructure: Digital Libraries and Machine Learning Approaches. Our digital libraries are poorly documented and barely evaluated. Results can vary wildly over time. This is sometimes expected (i.e., numbers go up as more research is published); however, it sometimes also *decreases* due to adjustments in metadata[8]. Metadata in publication databases often has errors and cannot necessarily be relied on [28, 29]. Additionally, what databases cover is not always entirely clear and can vary based on institutional access—e.g., the "*Web of Science Core Collection*" consists of different subdata sets depending on university subscription [48]. This makes one of the fundamental goals of systematic reviews—namely, that it should be possible to reproduce the results—rather difficult. It is considered best practice in other fields to conduct searches on multiple databases. Perhaps we need to consider doing multiple searches over several days to try to mitigate database fluctuation. However, perhaps we also need to re-consider or be clearer about what we require for a systematic review to be "reproducible": What do we mean when talk about reproducing results? For example, as long as the search queries themselves are reported, and the records of papers found in each step, then perhaps we should not require the search to yield the same number of results, simply because we cannot rely on the databases to be consistent.

Still, there are additional issues with designing multiple searches to be *comparable* across databases. Databases use a variety of different keyword and filter options, and often they are only poorly documented. Guidance for creating comparable searches across, for example, ACM and Scopus, would be highly beneficial for synthesis in our field. Current ACM DL tutorials are not sufficient for this purpose, and contacting the ACM DL team about database and search specifics has been unproductive. One option is to work in concert with publishers in a participatory design project with ourselves as the target "end-users" directing the design of these systems in a more fruitful direction for supporting review work.

[5] "*Item 18. Present assessments of risk of bias for each included study*" [64]
[6] "*Item 15. Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome*" [64], e.g., via the GRADE framework for evaluating quality of evidence in a review [35].
[7] "*Item 13e. Describe any methods used to explore possible causes of heterogeneity among study results (such as subgroup analysis, meta-regression)*" [64]

[8] And on some days, databases are simply buggy: on one memorable occasion, we noted the ACM DL reporting 0, then 200+, then 0, then 500+ results for the same search within a single day.

Another option to consider is automation. With the growth in artificial intelligence and machine learning, the landscape of digital infrastructure surrounding databases and publication searches now also features tools for (semi-)automated search (e.g., Research Rabbit[9]) or screening (e.g., ASReview[10]). These may be of interest for reviewing the field, but to what extent they can and should be used in formal systematic reviews is an open question—especially as the exact data sources and how often they are updated is often not made explicit. Perhaps a participatory design approach can again be useful as a starting point.

Finally, we note that there is little information on what kinds of publications relevant to HCI are found within which databases. Gusenbauer [34] created a discipline-based coverage map of a wide range of academic databases, giving us a first hint. However, HCI was not included in this disciplinary coverage map; it may be worth creating a disciplinary coverage map of databases for HCI specifically. This would give us a better idea of what kind of HCI research can be found in which database, and provide guidance on which databases to chose for specific research questions.

ENGAGE: A CALL TO ACTION FOR RESEARCH SYNTHESIS IN HCI

As a research community, we need to come together and decide what actions need to be taken towards building a set of standards that is rigorous yet inclusive of the diversity of work that we do in HCI. We do not aim to be prescriptive in this manifesto, but we do offer some ideas for what to aim for based on the discussion so far:

- a shared understanding of what should be considered a systematic review, the desired and possible outcomes of systematic reviews, and the forms that systematic reviews can take when exploring diverse evidence resulting from different research paradigms (quantitative, qualitative, mixed-methods, as well as design research methods)
- a shared understanding of what best practices we want to encourage in secondary research methods: double screening, extraction, critical appraisal, protocol development and preregistration, etc., specifically through an agreement on standards (e.g., for critical appraisals of primary research: what kind and when)
- unearthing how the digital libraries relevant to HCI work (e.g., query filters) and what they cover
- better infrastructure in our publishing ecosystem: is it time for a subcommittee or track for research synthesis and meta science? Should we require structured abstracts or checklists for primary and/or secondary research?
- robust descriptions of and/or access to the interactive artifacts reported on in primary research papers to support research synthesis about them
- exploration of the design and use of living reviews [26]—as interactive systems, HCI expertise could be particularly beneficial here

[9]https://www.researchrabbit.ai/, last accessed: 23 Nov, 2022
[10]https://asreview.nl/, last accessed 15 Dec, 2022

Our goal is to begin a discussion and gather different experiences and opinions of researchers on the role that systematic reviews should play, on what a systematic review should look like, and how systematic reviews are currently valued and received within the CHI community—and more broadly, within HCI as a whole.

ACKNOWLEDGMENTS

Many thanks to Maximilian Altmeyer for feedback on an earlier draft of this alt.chi paper, and to all of our colleagues for the many lively discussions on these topics over the years.

REFERENCES

[1] 2020. Rise of the preprints. *Nature Cancer* 1, 11 (Nov. 2020), 1025–1026. https://doi.org/10.1038/s43018-020-00151-y Number: 11 Publisher: Nature Publishing Group.
[2] 2022. Replication studies hold the key to generalization. *Nature Communications* 13, 1 (Nov. 2022). https://doi.org/10.1038/s41467-022-34748-x
[3] ACM Computing Surveys (CSUR). 2022. CSUR Author Guidelines. Retrieved November 23, 2022 from https://dl.acm.org/journal/csur/author-guidelines
[4] 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 (Sept. 2015), 132–140. https://doi.org/10.1097/xeb.0000000000000055
[5] Mara Balestrini, Yvonne Rogers, and Paul Marshall. 2015. Civically Engaged HCI: Tensions between Novelty and Social Impact. In *Proceedings of the 2015 British HCI Conference* (Lincoln, Lincolnshire, United Kingdom) *(British HCI '15)*. Association for Computing Machinery, New York, NY, USA, 35–36. https://doi.org/10.1145/2783446.2783590
[6] Elaine Beller, , Justin Clark, Guy Tsafnat, Clive Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, Jun Xia, Karen Robinson, and Paul Glasziou. 2018. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews* 7, 1 (May 2018). https://doi.org/10.1186/s13643-018-0740-7
[7] 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 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 533, 20 pages. https://doi.org/10.1145/3411764.3445193
[8] Alan F. Blackwell. 2015. HCI as an Inter-Discipline. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 503–516. https://doi.org/10.1145/2702613.2732505
[9] Ann Blair. 2003. Reading Strategies for Coping with Information Overload ca. 1550-1700. *Journal of the History of Ideas* 64, 1 (Jan. 2003), 11. https://doi.org/10.2307/3654293
[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806 Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1628806.
[11] Frederik Brudy, Christian Holz, Roman Rädle, Chi-Jui Wu, Steven Houben, Clemens Nylandsted 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–28. https://doi.org/10.1145/3290605.3300792
[12] Paul Cairns. 2007. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*. 1–7.
[13] Mhairi Campbell, Joanne E McKenzie, Amanda Sowden, Srinivasa Vittal Katikireddi, Sue E Brennan, Simon Ellis, Jamie Hartmann-Boyce, Rebecca Ryan, Sasha Shepperd, James Thomas, Vivian Welch, and Hilary Thomson. 2020. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* (Jan. 2020), l6890. https://doi.org/10.1136/bmj.l6890
[14] Yiqun T. Chen, Angela D. R. Smith, Katharina Reinecke, and Alexandra To. 2022. Collecting and Reporting Race and Ethnicity Data in HCI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 327, 8 pages. https://doi.org/10.1145/3491101.3519685

[15] CHI 2023 alt.chi committee. 2022. alt.chi page on CHI 2023 website. Retrieved September 29, 2022 from https://chi2023.acm.org/for-authors/alt-chi/

[16] CHI 2023 alt.chi committee. 2022. Selecting a Subcommittee; CHI 2023 website. Retrieved November 23, 2022 from https://chi2023.acm.org/subcommittees/selecting-a-subcommittee/

[17] CHI PLAY 2021 committee. 2021. Full Papers page on the CHI PLAY 2021 website. Retrieved November 23, 2022 from https://chiplay.acm.org/2021/full-papers/

[18] Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* 118, 41 (Oct. 2021). https://doi.org/10.1073/pnas.2021636118

[19] Citational Justice Collective, Gabriela Molina León, Lynn Kirabo, Marisol Wong-Villacres, Naveena Karusala, Neha Kumar, Nicola Bidwell, Pedro Reynolds-Cuéllar, Pranjal Protim Borah, Radhika Garg, Sushil K. Oswal, Tee Chuanromanee, and Vishal Sharma. 2021. Following the Trail of Citational Justice: Critically Examining Knowledge Production in HCI. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 360–363. https://doi.org/10.1145/3462204.3481732

[20] Mark Colley, Taras Kränzle, and Enrico Rukzio. 2022. Accessibility-Related Publication Distribution in HCI Based on a Meta-Analysis. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 299, 28 pages. https://doi.org/10.1145/3491101.3519701

[21] Alison Cooke, Debbie Smith, and Andrew Booth. 2012. Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis. *Qualitative Health Research* 22, 10 (July 2012), 1435–1443. https://doi.org/10.1177/1049732312452938

[22] Harris Cooper. 2015. *Research synthesis and meta-analysis: A step-by-step approach.* Vol. 2. Sage Publications.

[23] Fabio Q.B. da Silva, André L.M. Santos, Sérgio Soares, A. César C. França, Cleviton V.F. Monteiro, and Felipe Farias Maciel. 2011. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology* 53, 9 (Sept. 2011), 899–913. https://doi.org/10.1016/j.infsof.2011.04.004

[24] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the Landscape of Sustainable HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1975–1984. https://doi.org/10.1145/1753326.1753625

[25] Florian Echtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395

[26] 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. https://doi.org/10.1016/j.jclinepi.2017.08.010

[27] Daniel Fallman and Erik Stolterman. 2010. Establishing criteria of rigour and relevance in interaction design research. *Digital Creativity* 21, 4 (Dec. 2010), 265–272. https://doi.org/10.1080/14626268.2010.548869

[28] 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. https://doi.org/10.1016/j.joi.2016.07.003

[29] Fiorenzo Franceschini, Domenico Maisano, and Luca Mastrogiacomo. 2016. The museum of errors/horrors in Scopus. *Journal of Informetrics* 10, 1 (Feb. 2016), 174–182. https://doi.org/10.1016/j.joi.2015.11.006

[30] Christopher Frauenberger, Judith Good, Geraldine Fitzpatrick, and Ole Sejer Iversen. 2015. In pursuit of rigour and accountability in participatory design. *International Journal of Human-Computer Studies* 74 (Feb. 2015), 93–106. https://doi.org/10.1016/j.ijhcs.2014.09.004

[31] Riccardo Gallotti and Manlio De Domenico. 2019. Effects of homophily and academic reputation in the nomination and selection of Nobel laureates. *Scientific Reports* 9, 1 (Nov. 2019), 17304. https://doi.org/10.1038/s41598-019-53657-6 Number: 1 Publisher: Nature Publishing Group.

[32] Gerald Gartlehner, Lisa Affengruber, Viktoria Titscher, Anna Noel-Storr, Gordon Dooley, Nicolas Ballarini, and Franz König. 2020. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of Clinical Epidemiology* 121 (May 2020), 20–28. https://doi.org/10.1016/j.jclinepi.2020.01.005

[33] Kathrin Gerling and Max V. Birk. 2022. Reflections on Rigor and Reproducibility: Moving Toward a Community Standard for the Description of Artifacts in Experimental Games Research. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play* (Bremen, Germany) *(CHI PLAY '22)*. Association for Computing Machinery, New York, NY, USA, 266–267. https://doi.org/10.1145/3505270.3558360

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

[35] Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 336, 7650 (April 2008), 924–926. https://doi.org/10.1136/bmj.39489.470347.ad

[36] Neal R. Haddaway and Gary S. Bilotta. 2016. Systematic reviews: Separating fact from fiction. *Environment International* 92-93 (July 2016), 578–584. https://doi.org/10.1016/j.envint.2015.07.011

[37] 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 (March 2018). https://doi.org/10.1186/s13750-018-0121-7

[38] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLOS Biology* 13, 3 (March 2015), e1002106. https://doi.org/10.1371/journal.pbio.1002106 Publisher: Public Library of Science.

[39] 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 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 530, 14 pages. https://doi.org/10.1145/3411764.3445361

[40] Quan Nha Hong, Pierre Pluye, Mathieu Bujold, and Maggy Wassef. 2017. Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews* 6, 1 (March 2017). https://doi.org/10.1186/s13643-017-0454-2

[41] Kasper Hornbæk. 2015. We Must Be More Wrong in HCI Research. *Interactions* 22, 6 (oct 2015), 20–21. https://doi.org/10.1145/2833093

[42] Kasper Hornbæk, Søren S. Sander, Javier Andrés 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* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[43] Nathan G.J. Hughes, Josephine R. Flockton, and Paul Cairns. 2023. Growing together: An analysis of measurement transparency across 15 years of player motivation questionnaires. *International Journal of Human-Computer Studies* 169 (Jan. 2023), 102940. https://doi.org/10.1016/j.ijhcs.2022.102940

[44] John P.a. Ioannidis. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly* 94, 3 (2016), 485–514. https://doi.org/10.1111/1468-0009.12210 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0009.12210.

[45] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing systematic literature reviews in software engineering.* Technical Report. EBSE Technical report, ver. 2.3..

[46] Neha Kumar and Naveena Karusala. 2021. Braving Citational Justice in Human-Computer Interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for

Computing Machinery, New York, NY, USA, Article 11, 9 pages. https://doi.org/10.1145/3411763.3450389

[47] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. 2018. Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 257–270. https://doi.org/10.1145/3242671.3242683

[48] Weishu Liu. 2019. The data source of this study is Web of Science Core Collection? Not enough. *Scientometrics* 121, 3 (Sept. 2019), 1815–1824. https://doi.org/10.1007/s11192-019-03238-1

[49] 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* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3553–3562. https://doi.org/10.1145/2556288.2556969

[50] Lucylynn Lizarondo, Cindy Stern, Judith Carrier, Christina Godfrey, Kendra Rieger, Susan Salmond, Joao Apostolo, Pamela Kirkpatrick, and Heather Loveday. 2020. Chapter 8: Mixed methods systematic reviews. In *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/jbimes-20-09 Available from https://synthesismanual.jbi.global.

[51] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 401, 15 pages. https://doi.org/10.1145/3411764.3445701

[52] Kamal R Mahtani, Carl Heneghan, and Jeffrey Aronson. 2019. Single screening or double screening for study selection in systematic reviews? *BMJ Evidence-Based Medicine* 25, 4 (Nov. 2019), 149–150. https://doi.org/10.1136/bmjebm-2019-111269

[53] Iain J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 8, 1 (July 2019). https://doi.org/10.1186/s13643-019-1074-9

[54] Joe Marshall, Conor Linehan, Jocelyn Spence, and Stefan Rennick Egglestone. 2017. Throwaway Citation of Prior Work Creates Risk of Bad HCI Research. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 827–836. https://doi.org/10.1145/3027063.3052751

[55] Joe Marshall, Conor Linehan, Jocelyn C. Spence, and Stefan Rennick Egglestone. 2017. A Little Respect: Four Case Studies of HCI's Disregard for Other Disciplines. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 848–857. https://doi.org/10.1145/3027063.3052752

[56] 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, 1 (Nov. 2019). https://doi.org/10.1186/s12874-019-0855-0

[57] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174

[58] Alexios-Fotios A. Mentis, Efthimios Dardiotis, Vasiliki Efthymiou, and George P. Chrousos. 2021. Non-genetic risk and protective factors and biomarkers for neurological disorders: a meta-umbrella systematic review of umbrella reviews. *BMC Medicine* 19, 1 (Jan. 2021). https://doi.org/10.1186/s12916-020-01873-7

[59] Robert Andrew Moore, Emma Fisher, and Christopher Eccleston. 2022. Systematic reviews do not (yet) represent the 'gold standard' of evidence: A position paper. *European Journal of Pain* 26, 3 (2022), 557–566. https://doi.org/10.1002/ejp.1905 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejp.1905.

[60] Nature Human Behaviour. 2019. The importance of no evidence. *Nature Human Behaviour* 3, 3 (March 2019), 197–197. https://doi.org/10.1038/s41562-019-0569-7 Number: 3 Publisher: Nature Publishing Group.

[61] Ilkka Niiniluoto. 2019. Scientific Progress. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[62] Antti Oulasvirta and Kasper Hornbæk. 2016. HCI Research as Problem-Solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4956–4967. https://doi.org/10.1145/2858036.2858283

[63] 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. *Systematic Reviews* 10, 1 (March 2021). https://doi.org/10.1186/s13643-021-01626-4

[64] 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, and Joanne E McKenzie. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* (March 2021), n160. https://doi.org/10.1136/bmj.n160

[65] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 141, 16 pages. https://doi.org/10.1145/3411764.3445734

[66] Henning Pohl and Aske Mottelson. 2019. How We Guide, Write, and Cite at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310429

[67] Andrew S. Pullin and Gavin B. Stewart. 2006. Guidelines for Systematic Review in Conservation and Environmental Management. *Conservation Biology* 20, 6 (Dec. 2006), 1647–1656. https://doi.org/10.1111/j.1523-1739.2006.00485.x

[68] Felix Putze, Susanne Putze, Merle Sagehorn, Christopher Micek, and Erin T. Solovey. 2022. Understanding HCI Practices and Challenges of Experiment Reporting with Brain Signals: Towards Reproducibility and Reuse. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 31 (mar 2022), 43 pages. https://doi.org/10.1145/3490554

[69] Stuart Reeves. 2015. Human-computer interaction as science. *Aarhus Series on Human Centered Computing* 1, 1 (Oct. 2015), 12. https://doi.org/10.7146/aahcc.v1i1.21296

[70] Margarete Sandelowski, Corrine I Voils, and Julie Barroso. 2006. Defining and designing mixed research synthesis studies. *Research in the schools: a nationally refereed journal sponsored by the Mid-South Educational Research Association and the University of Alabama* 13, 1 (2006), 29. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2809982/

[71] Katie Seaborn and Jacqueline Urakami. 2021. Measuring voice UX quantitatively: A rapid review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3411763.3451712

[72] Patrick E. Shrout and Joseph L. Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology* 69, 1 (2018), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845 _eprint: https://doi.org/10.1146/annurev-psych-122216-011845.

[73] 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 (2019), 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

[74] K. Slim and T. Marquillier. 2022. Umbrella reviews: A new tool to synthesize scientific evidence in surgery. *Journal of Visceral Surgery* 159, 2 (April 2022), 144–149. https://doi.org/10.1016/j.jviscsurg.2021.10.001

[75] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* 104 (Nov. 2019), 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[76] Mark Starr, Iain Chalmers, Mike Clarke, and Andrew D. Oxman. 2009. The origins, evolution, and future of The Cochrane Database of Systematic Reviews. *International Journal of Technology Assessment in Health Care* 25, S1 (2009), 182–195. https://doi.org/10.1017/S026646230909062X

[77] Isabelle Stengers. 2018. *Another Science is Possible: A Manifesto for Slow Science*. John Wiley & Sons. Google-Books-ID: oxJSDwAAQBAJ.

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

[79] 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). https://doi.org/10.1186/1471-2288-12-181

[80] Marta Topor, Jade Pickering, Ana Barbosa Mendes, Dorothy Bishop, Fionn Cléirigh Büttner, Mahmoud Elsherif, Thomas Rhys Evans, Emma L Henderson, Tamara Kalandadze, Faye Nitschke, et al. 2020. An integrative framework for planning and conducting Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR). (2020). Preprint / not yet published. https://doi.org/10.31222/osf.io/8gu5z.

[81] April Tyack and Elisa D. Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. In *Proceedings of the 2020 CHI*

Conference on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3313831.3376723

[82] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence* 3, 2 (Feb. 2021), 125–133. https://doi.org/10.1038/s42256-020-00287-7

[83] Sarah Theres Völkel, Christina Schneegass, Malin Eiband, and Daniel Buschek. 2020. What is "Intelligent" in Intelligent User Interfaces? A Meta-Analysis of 25 Years of IUI. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 477–487. https://doi.org/10.1145/3377325.3377500

[84] Steve Whittaker, Loren Terveen, and Bonnie A. Nardi. 2000. Let's Stop Pushing the Envelope and Start Addressing It: A Reference Task Agenda for HCI. *Human–Computer Interaction* 15, 2-3 (Sept. 2000), 75–106. https://doi.org/10.1207/s15327051hci1523_2

[85] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: From a Panel to a New Submission Venue for Replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 1185–1188. https://doi.org/10.1145/2212776.2212419

[86] Max L. Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. RepliCHI - CHI Should Be Replicating and Validating Results More: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 463–466. https://doi.org/10.1145/1979742.1979491

[87] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (apr 2016), 38–44. https://doi.org/10.1145/2907069

[88] Tracee Vetting Wolf, Jennifer A. Rode, Jeremy Sussman, and Wendy A. Kellogg. 2006. Dispelling "Design" as the Black Art of CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 521–530. https://doi.org/10.1145/1124772.1124853

[89] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704