Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022

Kavous Salehzadeh Niksirat
kavous.salehzadehniksirat@unil.ch
University of Lausanne
Lausanne, Switzerland

Lahari Goswami
lahari.goswami@unil.ch
University of Lausanne
Lausanne, Switzerland

Pooja S. B. Rao
pooja.rao@unil.ch
University of Lausanne
Lausanne, Switzerland

James Tyler
james.tyler@unil.ch
University of Lausanne
Lausanne, Switzerland

Alessandro Silacci
alessandro.silacci@unil.ch
University of Lausanne
Lausanne, Switzerland
School of Management of Fribourg,
HES-SO University of Applied
Sciences and Arts Western
Switzerland
Fribourg, Switzerland

Sadiq Aliyu
sadiq.aliyu@unil.ch
University of Lausanne
Lausanne, Switzerland

Annika Aebli
nikaaebli@gmail.com
University of Lausanne
Lausanne, Switzerland

Chat Wacharamanotham
chat@acm.org
Swansea University
Swansea, United Kingdom

Mauro Cherubini
mauro.cherubini@unil.ch
University of Lausanne
Lausanne, Switzerland

ABSTRACT

In recent years, various initiatives from within and outside the HCI field have encouraged researchers to improve research ethics, openness, and transparency in their empirical research. We quantify how the CHI literature might have changed in these three aspects by analyzing samples of 118 CHI 2017 and 127 CHI 2022 papers—randomly drawn and stratified across conference sessions. We operationalized research ethics, openness, and transparency into 45 criteria and manually annotated the sampled papers. The results show that the CHI 2022 sample was better in 18 criteria, but in the rest of the criteria, it has no improvement. The most noticeable improvements were related to research transparency (10 out of 17 criteria). We also explored the possibility of assisting the verification process by developing a proof-of-concept screening system. We tested this tool with eight criteria. Six of them achieved high accuracy and F1 score. We discuss the implications for future research practices and education.

This paper and all supplementary materials are freely available at https://doi.org/10.17605/osf.io/n25d6.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → *Codes of ethics.*

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580848

KEYWORDS

replicability, reproducibility, transparency, ethics, open science, data availability, CHI

ACM Reference Format:
Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3544548.3580848

1 INTRODUCTION

Empirical research is one of the cornerstones of the Human-Computer Interaction (HCI) field. Since HCI research examines human experiences, ethical research has long been at the heart of planning and conducting studies. In the last decade, many scholarly fields increasingly recognized the value of openness and transparency in research. The field of HCI also participates in this broader discourse through various movements and research works. Let us look at these three values—Research Ethics, Openness, and Transparency.

Research ethics aims to protect research participants and foster socially responsible collaboration between science and society [75]. Research ethics in HCI studies include having study plans vetted by an institutional review board (IRB), obtaining informed consent from participants, implementing measures to ensure participant safety, and protecting data collected from study participants [14, 18, 38]. Within the ACM SIGCHI community, several research publications (e.g., [1, 64, 78]) and events (e.g., [18, 37]) were dedicated to discourses on research ethics. In 2016, the SIGCHI

Executive Committee appointed an Ethics Committee to facilitate the discourses and review related policies and procedures.

The UNESCO Recommendation on Open Science defines the term "Open Science" as "an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone" [90]. Although we appreciate the inclusiveness of this definition, for the reason that will be apparent in the next paragraph, we use a narrower definition in this paper: *Openness* refers to precisely the availability of research publications and materials. Openness initiatives have led research institutions and funding agencies to renegotiate their relationships with scientific publishers—including the ACM.[1] Consequently, ACM SIGCHI also made papers in selected conference proceedings from 2016 freely downloadable.

Transparency is closely related to openness and is often mentioned together, such as in The Center of Open Science's Transparency and Openness Promotion Guideline [65]. For this paper, we distinguish transparency from openness. We define *transparent research practices* as researchers' actions in disclosing details of methods, data, and other research artifacts. A transparent practice does not guarantee openness and vice versa. For example, describing statistical results in detail is transparent, but when the paper is behind a paywall, the results are also not open. In the HCI community, the discourse on transparency manifests in community-led events, such as RepliCHI [104–107] and Transparent Research [25, 49, 50], surveys [94, 95], and opinion pieces [27, 88].

Despite being regarded as desirable qualities, research ethics, openness, and transparency could be challenging to achieve. The limitation of research resources—finance and human resources—and the misalignment of incentives can be barriers to openness and transparency [90, 95]. Specifically for HCI, some research settings may cause tensions between these values. For example, research projects with participants from a vulnerable population might need to prioritize ethics over transparency. In other cases, researchers may need to sacrifice these values to ensure the quality of the knowledge. For example, a research project could emphasize transparency by creating a social network to learn about people's behavior on social media sites. The ecological validity of the findings from this study would be less than if the study were conducted on Facebook or Twitter where transparency of research data is limited.

Previous work either investigated specific aspects of the HCI literature such as statistical reporting [94], sample size reporting [19], or replication [45]. Other works indirectly assess the situation through self-reported surveys [95], and content-analysis of journal guidelines [11]. To determine how the field of HCI evolved in these aspects and where the community should focus improvement efforts, we need an assessment across these aspects based on actual published papers and their research artifacts.

Towards this goal, this paper makes three contributions:

- We collected criteria in research ethics, openness, and transparency and operationalized them for evaluation based on published papers and research materials.
- We sampled 118 and 127 papers from CHI 2017 and 2022 and evaluated them with these criteria to provide snapshots of research practices and discuss the implications of the results.
- We explored the possibility of assisting the assessment by developing a proof-of-concept screening system.

2 RELATED WORK

In this section, we first review the existing work on research ethics. Next, we review the relevant studies on practices related to openness. Finally, we review studies focused on the principle of transparency. For all three practices, we review studies conducted in the HCI community and those in adjacent fields that contribute to general guidelines.

2.1 Research Ethics

The ethical guidelines of *responsibly* conducting experiments are often informed by national or state laws and institutional regulations. Additionally, different science communities design their own domain-specific codes of ethics [99].[2] Munteanu et al. [64] make the point that although the formal process of establishing the ethical approval of a study can vary by country, the underlying principles are universal. However, they also note that new technologies present challenges to existing ethical review processes, which may need mitigating. An example is the raw power of data collection afforded by technologies, where opinions on the kind (or extent) that is acceptable are subject to changing attitudes [93].

Some researchers, such as Punchoojit and Hongwarittorrn [78], have attempted to understand how ethical concerns have evolved. They offer categories ranging from broad issues to some highly specific to HCI. It is vital that such concerns or conflicts are not oversimplified or proceduralized to an extent that researchers refrain from engaging with the issues [17]. Instead of simply writing that they followed the institutional safeguards, researchers should describe research ethic issues they faced and how they were addressed. Such ethical considerations can also help researchers inoculate themselves against biases.

Well-defined standards may help researchers engage with and report on the ethical dimensions of their work. Ethical standards in HCI can be related to both the data collection & analysis and reporting & dissemination of results.[3] For data collection & analysis, practices such as acquiring *ethical approval* and collecting participants' *consent* are discussed in HCI textbooks (see, for example, [55, section 15]). Some aspects are studied in more detail. For example, Pater et al. [71] assessed ethical challenges in *compensating* participants. Their systematic literature review of papers from four HCI venues (CHI, CSCW, Ubicomp/IMWUT, UIST) found that 84.2% of the studies did not sufficiently report essential decisions in participant compensation. For the ethics of reporting and the dissemination of results, Abbott et al. [1] examined reporting trends with regard to *anonymization practices* in CHI. They studied 509 CHI papers for health, wellness, accessibility, and aging research and found that codes and pseudonyms were the most

[1] See ACM Plan S Compliance statement at https://authors.acm.org/open-access/plan-s-compliance, last accessed January 2023.

[2] See, for example, ACM Code of Ethics and Professional Conduct at https://www.acm.org/code-of-ethics, last accessed January 2023.

[3] In this work, we focus on research ethics. To read about design ethics, see a literature survey by Nunes Vilaza et al. [66].

used techniques to protect participant privacy. They offered further suggestions to the community that facilitate data reporting while limiting privacy risks.

Finally, several studies discussed the ethical precautions that HCI researchers should consider when dealing with *vulnerable populations*. For example, Walker et al. [97] proposed heuristics for HCI research with vulnerable populations. This heuristic includes several actions to be conducted *before* research (e.g., understanding the needs and interests of vulnerable communities), *during* research (e.g., considering if collected data can be harmful to participants), and *after* research (e.g., considering researchers' positionality in relation to the vulnerable community when presenting the results). On a different note, Antle [9] reflected on their experience in doing research with children who live in poverty and asked five questions to consider when working with vulnerable populations, for example, *"How can we feel relatively certain that we are providing benefits to the population we are working with?"* [9]. Furthermore, Gautam et al. [40] described the tension they experienced in running a participatory design study with a vulnerable population and McDonald et al. [58] discussed how privacy researchers should consider the *power dynamics* that may impact vulnerable populations.

Despite these studies, the adherence of HCI researchers to different practices regarding research ethics still requires investigation.

2.2 Openness

In comparison to the studies on transparency and research ethics, the HCI literature lacks sufficient studies on openness to understand to what extent researchers publish their papers and materials freely—without locking them behind a paywall—and whether they face any challenges in meeting open science standards. More than a decade ago, several articles in ACM magazines discussed open-access publication models and their benefits for computer science [57, 98]. In order to publish open-access, authors had to pay a so-called article processing charges (APCs) fee. While APCs are mostly sponsored by the authors' institutions or funding agencies, researchers without such support might face difficulties [22]. Furthermore, awareness of open science is not globally distributed and some researchers, from developing countries, might face difficulties when seeking for funding for open access. Spann et al. [86] discussed the benefits of an alternative publication model (used by some publishers) called Pay What You Want (PWYW) where researchers are allowed to pay any amount that they can afford. Some publishers (e.g., ACM) support green open access and allow authors to publish the author version of their article publicly on their personal or institutional website [4]. However, some authors might also use *commercial* social networking websites such as ResearchGate. Jamali [46] showed that almost half of the authors who publish their non-open-access articles on ResearchGate infringe the copyrights of their publishers. Thus, ACM strictly prohibited sharing on such websites [4].

Besides the use of open access for sharing articles, several researchers studied different practices for sharing supplementary materials (e.g., [15]). One of the most typical practices for sharing supplementary materials is promising to share *upon request*. Krawczyk and Reuben [52] showed that the compliance rate for such requests is low. Vines et al. [92] showed that it can be even lower when papers are published far in the past. The standard approach for material sharing is the use of platforms that are compatible with FAIR principles [101], namely being Findable (e.g., having unique identifiers), Accessible (e.g., not being locked behind a paywall), Interoperable (e.g., providing ReadMe files to clarify the structure), and Reusable (e.g., providing metadata that can support readers to understand the data and reuse it). Two well-known FAIR-compatible platforms are OSF and Zenodo.

2.3 Transparency

Transparent research practices disclose details of methods, data, and other research artifacts. In quantitative research, these practices usually lead to reproducibility and increase the likelihood of replicability [65]. Reproducibility means that re-running the same analysis on the same data yields the same results [72]. Replicability means that re-running the study to produce new data—analyzed in the same or different manner—should yield a similar result [72]. No study can be reproduced or replicated without having access to its detailed methodology, procedures, and materials.

Replication studies—where the explicit intent is to confirm or challenge the results of prior work—are infrequent in the field of HCI. Hornbæk et al. [45] examined 891 studies across four different HCI outlets and found that only 3% attempted to replicate a prior result. Upon closer examination, they found that authors of *non-replication* studies could have often corroborated earlier work by, for example, analyzing data differently and collecting additional data. Often, these choices would have required *minimal* additional effort [45]. That many HCI studies overlook these kinds of opportunities has led some to question the culture of the field. Nevertheless, outside HCI, the consensus on general practices for research transparency boils down to a 36-item checklist [5].

In qualitative research, the discussion on research transparency is more complex. The term transparency has another semantics. In the Introduction chapter in an influential ethnographic text—*The Religion of Java*—Geertz describes a desirable characteristic of ethnographic reports, where the "ethnographer is able to get out of the way of his data, to make himself translucent" [41, p. 7]. Clifford disagrees with this portrayal of objectivity as "too simple notions of transparency". The word "transparency" was used as a paraphrase of Geertz's "translucency". To avoid confusion, we will refer to this semantics with Geertz's original term: *translucency*. In our definition, research transparency does not require or preclude translucency. In fact, despite disagreeing with translucency, Clifford praised Geertz's practice of sharing his ethnographic field notes extensively [26, p. 61], which is a transparency practice.

In a panel discussion about transparency in qualitative research at CHI 2020 [88], the panelists concurred that in qualitative research, transparency in the method should be emphasized over transparency in data. In addition to this separation of transparency between data and method, Moravcsik [62] points out the third aspect: *production transparency*, which demonstrates how arguments and citations are drawn fairly from different points of view in the literature. We set aside production transparency because it is not possible to evaluate this aspect within each paper. In the following subsections, we distinguish transparency in method, results in the paper, data beyond the paper, and other non-data research artifacts.

2.3.1 Transparency of research methods. Numerous guidelines for reporting research methods are evidence of the importance of research method transparency. In quantitative research, there is a list of 34 research decisions that could be pertinent to *p*-hacking [100]. More specifically, there are guidelines for reporting decisions on sample size [19, 53], measurements, and constructs [6]. Quantitative data analysis could also be transparent by sharing the analysis code. In a survey of CHI 2018–2019 authors [95], around 25% of the respondents shared quantitative analysis procedures.

In qualitative research, the *Standards for Reporting Qualitative Research* (SRQR) standard is extensive in methodological decisions [67]. More specific guides are also available for interview and focus group research [8, 89], reflexive thematic analysis [16], and for using inter-rater reliability [59]. The survey of CHI 2018–2019 authors found around 25% of the respondents shared qualitative analysis procedures; this percentage is similar to quantitative research [95].

Another practice to foster research transparency is the *preregistration* of study objectives and methods before collecting or analyzing data. Cockburn et al. [27] promote preregistering HCI experiments. They argue that preregistration will clarify the intent to do exploratory research and reduce the misuse of null hypothesis significance testing (NHST). Preregistration is also helpful in qualitative research. Haven et al. [43] conducted a Delphi study with 295 qualitative researchers; the results of their study culminated in 13 items for preregistration of qualitative studies. In the field of HCI, preregistration is rare. Pang et al. [68] systematically reviewed CHI 2018–21 papers and found only 32 papers with preregistration. Another novel method to promote methodological transparency is *Registered Report*, where the research method is written and peer-reviewed before data collection [23]. Despite over 300 journals supporting this format[4], none of them is HCI.

2.3.2 Transparency of research results. In addition to research methods, the research results reported in the paper contribute to its transparency. In quantitative HCI research, problems in statistical reporting persist. In 2006, Cairns [20] looked at the use of inferential statistics in BCS HCI conferences over two years and the output of two leading HCI journals in the same year. Of the 80 papers analyzed, 41 used inferential statistics, and only *one* conducted inferential statistics appropriately. All others had errors in their reporting or analysis. Still, in 2020, Vornhagen et al. [94] looked at the quality of reporting statistical significance testing in CHI PLAY 2014–19. More than half of the papers employed NHST *without* adequate specificity in their research questions of statistical hypotheses [94]. To address these problems, several HCI books are dedicated to statistical practices and reporting, for example, [21, 80].

In qualitative research, how research results are transparent depends on the research methods. The SRQR standard only requires the results to (1) describe an analysis and (2) support with evidence [67]. The *Consolidated criteria for reporting qualitative research* (COREQ) checklist—for interviews and focus group studies—adds consistency and clarity as criteria [89]. Braun & Clarke also highlight that the results must fit the assumptions made in the analysis method and the epistemology [16, Table 2].

[4]See https://www.cos.io/initiatives/registered-reports, last accessed January 2023.

2.3.3 Transparency of data. In Wacharamanotham et al. [95]'s survey, they found that around 40% shared some data, with around 21% sharing raw data. The respondents of their survey reported key concerns about protecting data that may be sensitive and that they had not obtained permission from the participants to share data. A recent study supports these concerns: VandeVusse et al. [91] found that participants in qualitative studies volunteer to share data to be helpful. However, their participants misunderstood "sharing" as disseminating research findings instead of sharing the interview transcripts [91]. HCI research has looked into challenges in Research Data Management [34, 35] and has come up with an innovative approach to facilitate sharing despite these challenges [63].

2.3.4 Transparency of research artifacts. In addition to sharing methods, results, and data, researchers also generate other artifacts. In the survey of CHI 2018–2019 authors [95], slightly above 30% of respondents reported that they shared study materials, such as stimuli or interview guides. A slightly higher percentage —around 40%—reported sharing hardware or software. One worrisome result is that many respondents indicated that they did not see the benefits of sharing these materials. In another analysis of CHI 2016–17 papers, only around 2% of papers publicly share source code [31].

The proliferation of guidelines, discussions, and empirical studies in the last few years might have changed the transparency practices in HCI research. In fact CHI conferences have added a *Transparency* section to the Guide to Authors and Reviewers[5] since CHI 2020 [42]. For the time-being, empirical studies about research transparency in the HCI literature are self-reported survey [95], and focus on individual aspects [31, 94] or policies [11]. We need a comprehensive study into how transparency is actually practiced in order to take stock of where the field currently stands, and which directions the effort to improve should be focused.

While the previous research studied different aspects of research ethics, openness, and transparency, none provided a *comprehensive picture* of these practices in HCI. In particular, it is necessary to inquire into the *status quo* of the adopted practices and understand how much progress the field has made and which areas are lacking. One way to objectively measure this is by collecting criteria for these practices and analyzing the text of published research articles and their supplementary materials. To address this gap, we *operationalize* 45 criteria related to research ethics, openness, and transparency. We evaluate the HCI literature by comparing two samples of papers published in ACM CHI 2017 and CHI 2022. Additionally, given the lack of a *screening* tool to assess HCI articles, we explore the potential for such a system.

3 CRITERIA FOR RESEARCH ETHICS, OPENNESS, AND TRANSPARENCY

Towards the goal of evaluating the research ethics, openness, and transparency of HCI publications, we developed a comprehensive set of criteria for assessing published papers and their published research artifacts. This section describes the development process and highlights the insights we gained.

[5]See https://chi2020.acm.org/authors/papers/guide-to-a-successful-submission/, last accessed January 2023.

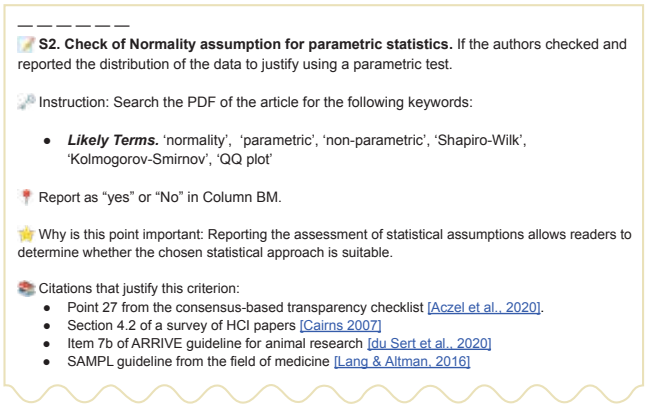

Figure 1: An excerpt from an instruction note provided to coders for data collection for STAT-NORMALITY criterion. The note involved a title, step-by-step instructions, keywords, rationale, and references.

3.1 Development process

We drew some criteria that are already operationalized in prior works, for example, statistical reporting criteria [94]. Other criteria were inspired by high-level principles, self-report checklists, survey questionnaires, and textbook recommendations. From these sources, two co-authors created a set of distinct criteria and worked out how to inspect them solely from the papers and their published research artifacts. This initial version was discussed and refined together with two other co-authors. The second version was used to create detailed coding instructions (half A4 page per criterion on average). Figure 1 demonstrated an excerpt from the instruction note provided for the authors, which includes a title, step-by-step instructions, keywords, rationale, and references (for detailed examples, see Sup. 1).[6] The coding instructions were refined in a collaborative coding process as detailed in Section 4.2.

The study was preregistered at OSF Registries[7]. In the preregistered study design, we identified 44 criteria. The total number of criteria evolved during the course of the study, as explained in Sup. 2. Table 1 presents an overview of the final version with 45 criteria. Some criteria (marked with an asterisk *) apply to a subset of empirical paper. For example, SHARE-INTERVIEW-GUIDE is only applicable to qualitative papers that use interviews and STAT-DESCRIPTIVE is only applicable to quantitative or mixed papers that uses frequentist statistics. The criteria are related to the distinct phases of research including study design, data collection, data analysis, and reporting (see Table 1). The specific subset of each criterion is listed in *criteria definition document* (see Tables 1–6 in this document in Sup. 3). We also provide Table 1 in the Excel format (Sup. 4) for authors, reviewers, and teachers to adapt them to their purposes.

The criteria definition document (Sup. 3) also provides the rationale behind each criterion in detail with additional citations. We hope that knowing the rationale will better encourage the practices related to research ethics, openness, and transparency. For example, a statistical guideline prescribes reporting degrees of freedom in statistical tests [54] (see STAT-PARAMETERS). In the supplement, we explain that readers could use the degrees of freedom to determine whether the choice of statistical tests and the input data are appropriate. In a different example, for the criterion about study preregistration (see PREREG), we explain that preregistration is a useful practice to avoid HARKing (i.e., Hypothesizing After the Results are Known) and we provide resources for the most commonly used services for preregistration.

3.2 Insights

Below, we describe notable insights from the criteria and the development process. Some insights are facts that—we believe—are not well known. Others are caveats for future researchers who will use this criteria set.

3.2.1 Downloading CHI papers for free (for a limited time), if you know where to look. Since 2016 SIGCHI have made the CHI proceedings available without any paywall restriction at this open-proceedings page.[8] Although this page indicates that the proceedings are "permanent open access," the availability is subject to the ACM OpenTOC program that is still in the pilot phase and could be discontinued in the future [3]. Additionally, it seems that OpenTOC pages are not indexed by search engines, which limits the discoverability of this access channel.

3.2.2 Supplementary materials are free on the ACM Digital Library. The ACM policy [2] indicates that supplementary materials on the ACM Digital Library can be downloaded for free, even if the paper itself is not. This fact makes the supplementary materials on ACM Digital Library compatible with the FAIR principles. Nevertheless, the supplementary materials for each paper are displayed as one zip file. This presentation impairs the discoverability of its content, especially when the paper is behind a paywall.

3.2.3 Nuances among openness and transparency terms. The terms "free," "open," "public," and "transparent" are closely related. However, we found two cases where their nuances matter. In the first case, the ACM Digital Library is marked at the top-left corner of some paper webpage with either "Open Access," "Free Access," or "Public Access." Only the Open Access paper can be accessed without a paywall at the time of publication in perpetuity. Public Access papers are eventually open after an embargo period—mandated by the funding agencies. For the last category, Free Access papers are freely accessible for a limited period—determined by ACM—before being locked behind a paywall.

The second case highlights the difference between transparency and openness. Some papers share research artifacts, such as questionnaires, in an appendix of the paper. Although this practice is transparent, the questionnaire is not open if the paper is behind a paywall. To avoid depending on the availability of the paper, research artifacts should be shared as separate materials in an open repository. In the criterion EXTRA-FAIR, we assess whether research artifacts are shared at a location that meets the FAIR principles. A paper may meet this criterion by publishing its supplementary

[6] All supplementary materials of the paper are publicly available on OSF at https://doi.org/10.17605/osf.io/n25d6.
[7] See preregistration document at https://doi.org/10.17605/osf.io/k35w4
[8] https://sigchi.org/conferences/conference-proceedings/, last accessed January 2023.

Table 1: A summary of research ethics, openness, and transparency criteria for evaluating research papers. See Sup. 3 for full definitions. This table is also available in Excel format in Sup. 4.

CODE	Criterion	Sources	Phase‡	Auto§
Criteria for Research Ethics				
IRB	Did the study receive approval from an institutional review board?	[55]	D	Def
CONSENT (reported)*	Was written consent obtained from study participants?	[55]	D/C	Def
CONSENT (form shared)*	Do supplementary materials include the consent form?	[55]	D/C	Def
STUDY-COMPENSATION*	Was participants' compensation explained in the paper?	[55]	D/R	Def
ANON	Was any data anonymization used?	[1, 103]	R	Scr
FACE-PHOTO*	Are facial photos in the paper shared with consent? Is privacy being protected?	[1, 24, 87]	R	PP
VULNERABLE*	Were any ethical measures taken to support vulnerable participants?	[79, 97]	D	No
ANIMAL*	Were any ethical measures taken to support animals?	[30]	D	Scr
Criteria for Openness				
PAYWALL-ACMDL†	Is the paper in ACM DL available as open access?	[4]	R	No
FREE-PDF-EXTERN†	Is the paper PDF available on external platforms other than ACM DL?	[46]	R	PP
EXTRA	Are any research artifacts beyond the paper provided anywhere?	[96]	R	Scr
EXTRA-EXIST*	Do all provided research artifacts exist at the location specified in the paper?	[101]	R	Scr
EXTRA-FAIR*†	Do any of the locations of provided artifacts satisfy the FAIR principle?	[101]	R	PP
Criteria for Transparency				
PREREG	Was the study preregistered?	[27, 65]	D	Def
SHARE-STIMULI*	Are study stimuli (except survey questionnaires) archived?	[95]	D/R	Scr
SHARE-SURVEY*	Are questionnaires or surveys archived?	[95]	D/R	Scr
SHARE-INTERVIEW-GUIDE*	Is interview guide archived?	[95]	D/R	Scr
SHARE-STUDY-PROTOCOL	Is the study protocol archived?	[73]	D/R	Scr
JUSTIFY-N-QUAL*	Was the sample size justified (qualitative studies)?	[19]	D	Scr
JUSTIFY-N-QUAN*	Was the sample size justified (quantitative studies)?	[53, 74]	D	Def
DEMOGRAPHICS*	Was the demographics information of the participants described?	[39]	C/R	Def
CONDITION-ASSIGNMENT*	Did the study properly explain study design (e.g., grouping, IDVs)?	[94]	D/R	Scr
SPECIFY-QUAL-ANALYSIS*	Is qualitative data analysis approach named or explicitly described?	[95][65]	A/R	Scr
SHARE-ANALYSIS-CODE*	Is quantitative data analysis code shared?	[95][65]	A/R	Scr
QUAL-DATA-RAW*	Is raw qualitative data shared?	[95][65]	R	Scr
QUAL-DATA-PROCESSED*	Is processed qualitative data shared?	[95][65]	R	Scr
QUAN-DATA-RAW*	Is raw quantitative data shared?	[95][65]	R	Scr
QUAN-DATA-PROCESSED*	Is processed quantitative data shared?	[95][65]	R	Scr
SHARE-SOFTWARE*	Is the source code of the software shared?	[95]	R	Scr
SHARE-HARDWARE*	Is the code of the hardware shared?	[95]	R	Scr
SHARE-SKETCH*	Is any hand-drawn sketch shared?	—	R	Scr
Criteria for Reporting (i.e., frequentist analysis, estimation analysis, qualitative reporting)				
STAT-DESCRIPTIVE (cen. tend.)*	For each key dependent variable on the interval or ratio scale, were their sample central tendency reported?	[28][54]	A/R	Scr
STAT-DESCRIPTIVE (variability)*	For each key dependent variable was their sample variability reported?	[28][54]	A/R	Scr
STAT-DESCRIPTIVE (cat. data)*	Were their sample reported for each key dependent variable on the nominal or ordinal scale? (categorical data)	[28][54]	A/R	Scr
STAT-CLEAR-PROCEDURE*	Is the statistical procedure for data analysis clearly named?	[28]	A/R	Scr
STAT-NORMALITY*	When the normality assumption is required by the statistical procedure, was the assumption assessed?	[5, 20, 54, 94]	A/R	Scr
STAT-OTHER-ASSUMPTIONS*	When the statistical procedure requires additional assumptions, were they assessed?	[54, 94]	A/R	Scr
STAT-PARAMETERS (df)*	Were degree of freedom reported?	[54, 94]	A/R	Scr
STAT-PARAMETERS (test value)*	Were the test statistic and all test parameters reported? (e.g., F-value)	[54, 94]	A/R	Scr
STAT-PARAMETERS (p-value)*	Were p-value reported?	[54, 94]	A/R	Scr
STAT-EFFECT-SIZE*	For the effects that were tested, were effect sizes reported?	[54, 94, 110]	R	Scr
STAT-CI*	For the effects that were tested, were their confidence intervals reported?	[28, 94]	R	Scr
ESTIMATES-INTERVAL*	Were interval estimates reported?	[29]	R	Scr
ESTIMATES-VIS-UNCERTAINTY*	Was the uncertainty of the effect visualized?	[29]	R	Scr
QUAL-INTERVIEW-REPORT*	Did the study properly report themes and quotes?	[55]	R	No

*Evaluated on applicable subset of empirical papers. See Section 3.1 for explanation.

†See additional discussion about these openness criteria in Section 3.2.

‡ Study Phase: D, C, A, and R stand for Study Design, Data Collection, Data Analysis, and Reporting, respectively.

§ Potential for Automation: **Def**: Definitely, **Scr**: Screening, **PP**: Potentially Possible, and **No**: Difficult to Automate. Details in Section 3.2.4.

materials on FAIR repositories (e.g., OSF) or on the ACM Digital Library (as discussed in the previous subsection). Papers that share research artifacts only in the appendix meet this criterion only when the paper itself is either open-access or public-access.

3.2.4 A potential for screening system. For the majority of the criteria, it is possible to narrow down parts of a paper for assessment based on keywords (for a complete list of keywords, see Sup. 5). This insight indicates the potential to automate (fully or partially) the assessment of some criteria. We describe a proof-of-concept system in Section 6. Based on this system, we indicate the potential of a screening system for each criterion in the fourth column of Table 1. We labeled them as 'definitely' (i.e., for criteria with high accuracy in our system), 'potentially possible' (i.e., for criteria that might require advanced techniques like Computer Vision, not attempted in our tool), 'screening' (i.e., for criteria where automation is possible to narrow down some papers or parts of them, but manual checks are required), and 'no' (i.e., for criteria that we believe require manual inspection). Six out of the eight criteria we attempted could be checked automatically with high accuracy (> 0.80) and F1 scores (> 0.75). For one of the criteria (CONDITION-ASSIGNMENT), our proof-of-concept system yielded a high accuracy of 0.81 but an F1 score of 0.74 narrowly missing our desired 0.75 threshold. One criterion (ANON) might benefit from machine-screening, but the content requires humans to manually do the checking. The proportions reported in Section 5, are solely based on the manual review effort. In the study below, we did not rely on the results of the screener tool for reporting the result section.

4 METHOD

To investigate the changes in research ethics, openness, and transparency practices in HCI, we applied the criteria above to assess papers from two proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). We chose CHI for three reasons: (1) Its once-per-year camera-ready deadline is a single cut-off point. The cut-off point provides a clear separation between years—unlike journal publications where the duration between initial submission and the publication varies across papers. (2) CHI conferences have considerable numbers of papers that span a broad range of HCI application areas. (3) For many years, CHI conferences hosted many events (SIG discussions, workshops, research presentations) that contributed to the discourse on research ethics, openness, and transparency. These events might have changed the awareness and understanding of these issues among their attendees.

In this study, we investigate how the field of HCI has progressed in addressing issues related to research ethics, openness, and transparency. This study will help us understand the extent to which practices in research ethics, openness, and transparency have been reported and implemented in the CHI literature.

Additionally, given the tension between practices in research ethics versus transparency [25, 36, 88, 95], we exploratorily investigate how transparency practices can differ between papers that deal with more ethical constraints (e.g., studies with vulnerable populations) and papers that deal with lesser ethical constraints (e.g., studies without vulnerable populations). This finding will provide an understanding of whether tension is actually reflected in researchers' practices.

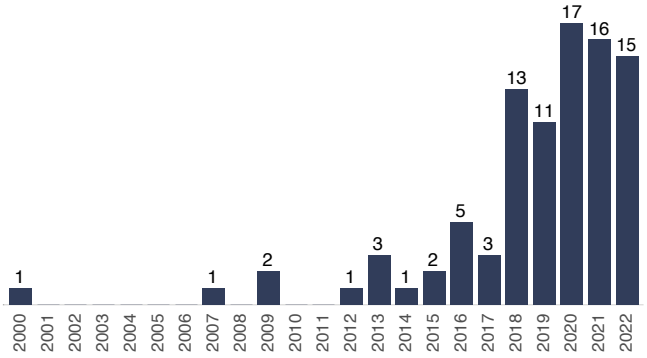

Figure 2: The search results of 'open science,' 'reproducibility,' 'replicability,' 'replication crisis,' or 'research ethics' from the ACM DL. The y-axis shows the number of matched papers. The majority of the matched papers were from after 2017.

Methodological deviations from the preregistered study plan are explained in Sup. 6. The study protocol had institutional review board (IRB) approval.

4.1 Samples

Proceeding Selection. We used proceedings of CHI 2022, which was the most recent volume at the time of this research. Additionally, we searched the abstracts of SIGCHI Sponsored Conferences between 2000–2022 with any of the following terms: open science, reproducibility, replicability, replication crisis, or research ethics. These searches resulted in 91 papers (full search results are listed in Sup. 7). As shown in Figure 2, 80% of these were published after 2017, suggesting it to be a watershed moment. Therefore, we selected the proceedings of CHI 2017 and CHI 2022. We used only the "Paper" publication type because the papers have undergone rigorous referee vetting processes.[9]

Sample Sizes. The sheer number of papers each year (600 in CHI 2017 and 637 in CHI 2022) exceeds our resources. For this study, we analyzed samples of papers. To determine the sample size, we considered the effect size from past surveys of transparent research practices among CHI authors [95]. Among the respondents of their surveys, the transparent research practices across all dimensions were, on average, 27.6% among CHI 2017 and 31% among CHI 2018 authors. The difference is 3%. We used this information to conduct an *a priori* power analysis based on the z-test of the difference between two independent proportions in G*Power [33] at $\alpha = 0.05$ and $\beta = 0.80$ (for details see Sup. 8.) The power analysis suggested sampling 119 papers from CHI 2017 and 127 papers from CHI 2022.

Sampling Procedure. The paper sampling procedure is demonstrated in Figure 3. The organization of sessions at CHI conferences groups together thematically related papers [51]. We used this fact to inform a stratified sampling [69], ensuring we drew across the application areas covered by the conference. The number of sessions (149 and 139 in CHI 2017 and 2022, respectively) is higher than the

[9]See https://www.acm.org/publications/policies/pre-publication-evaluation, last accessed January 2023.

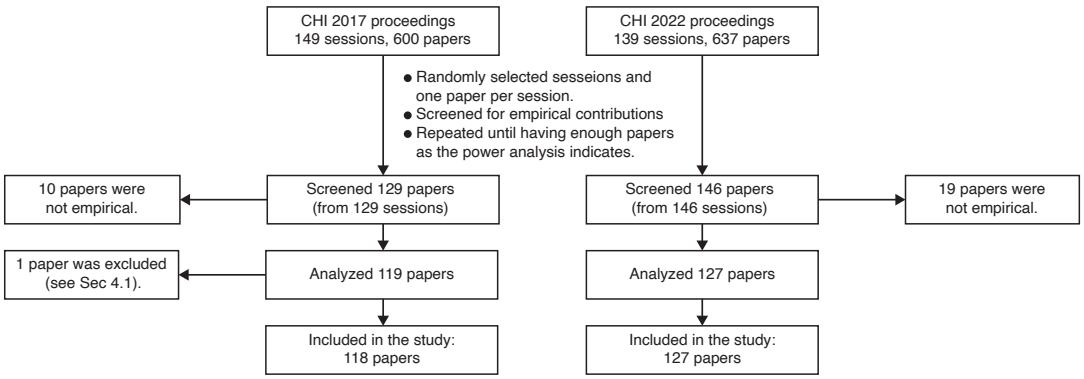

Figure 3: A flow diagram showing the paper sampling process.

planned sample size. Therefore, we randomly sampled the sessions, and for each session randomly sampled a paper.

Then, seven co-authors read each paper's title and abstract and coded its contribution type according to Wobbrock and Kientz [108]'s taxonomy of HCI research contributions. The coding of contribution types was later re-checked by one of the co-authors (i.e., different from the person who initially coded it). In case of any mismatches, the coding was refined.[10] If there were no empirical contributions, paper replacements occurred through subsequent rounds of sampling and coding with the same procedure. In one case, during data analysis, (as explained in deviations from pre-registration, Sup. 6), while conducting consistency checks on the articles, we found one article from CHI 2017 that was an experience report of case studies of design processes. Although the cases contain empirical studies, the article did not report on those empirical results and rather reported on the designers' experience working on these cases. Thus, we excluded this article, and our sample size was reduced from 246 to 245.

4.2 Coding procedure

Based on the title and the abstract of each paper, we coded the broad types of the method (qualitative, quantitative, mixed-method), research questions (exploratory or confirmatory), and the participants (human or animal). These broad codes allow us to subsequently subset the papers for each set of criteria. The assessment of the relevant subset of papers followed the procedure described in Sup. 3.

We use papers as the unit of analysis. For papers with multiple studies, a criterion can be satisfied by *any* of the studies described in the paper. In contrast, a criterion was marked as violated, only when *all* studies failed to met that. This hysteresis and the assessment at the paper-level is a lower bar to meet than assessing each study individually. Nevertheless, these choices are necessary for us to avoid making judgments about the relative importance of the studies in each paper. These choices also avoid the page limit constraint that was present only in CHI 2017.

Seven co-authors contributed to coding and were assigned to work on $Med = 35$ papers. The seven coders were two postdocs (in HCI and psychology) and five PhD students (all in HCI). The PhD students have 2-4 years of experience working on HCI research. The overall process of criteria definition and coding was supervised by two HCI professors who are experts on topics related to transparency, openness, and research ethics. This assignment allowed each coder to be familiar with the structure and context of their papers. To prevent overload, we worked in rounds; each round focused on 4–11 criteria drawn from similar aspects. Each round comprised these steps:

(1) An expert coder created a detailed procedure (see Figure 1 or Sup. 1).
(2) Each coder independently coded their paper.
(3) Each coder independently coded additional five papers randomized from other coders.
(4) We calculated an agreement score [61] from these twice-coded papers.
  (a) If the agreement score was lower than 90%, each coder coded three additional random papers and calculated the second agreement score from this set.
  (b) If the second agreement score was still lower than 90%, two expert coders inspected all of the twice-coded papers and resolved the inconsistencies.
(5) The resolutions were discussed and resolved in group meetings.
(6) Each coder then updated their work accordingly.
(7) Finally, each coder checked the work of another coder. The pairing of each round rotated according to a Latin Square to avoid systematic influences between coders.
(8) The detailed procedure (Figure 1 or Sup. 1) was updated to incorporate insights from the discussion.

Two coders with statistical knowledge created the codes for statistical criteria (e.g., STAT-DESCRIPTIVE and ESTIMATES-INTERVAL). Each coder worked on half of the papers with NHST statistics (a total of 117). After the first round of coding, 23.4% of the papers were unclear. We discussed these papers with a co-author who is an expert in statistics. After the consultation, the coders revised their work. Finally, each coder independently coded five random papers from another coder.

The agreement score of the twice-coded papers was 96.4% for the statistical criteria. For other criteria, the agreement scores were

[10] A reviewer pointed out that we could have better controlled this step by calculating inter-rater reliability, and we agree. We disclose that we overlooked this decision.

95.6% on average (SD=7.5%). In total, 30 review meetings were conducted, and 45 criteria were extracted out of these activities.

4.3 Data Analysis

As explained in deviations from preregistration (see Sup. 6), in the preregistration, we planned to use a two-sample Z-test for proportions [113] to compare the two years in each criterion. However, several criteria have boundary probabilities (close to 0 or 1) because the cell frequencies differ greatly. Z-tests and their confidence intervals are therefore not reliable in these cases [7, p. 164]. Instead, we calculated the confidence intervals using the Miettinen-Nurminen asymptotic score method—which does not suffer from the boundary cases [32, p. 250]. We use the implementation in the `diffscoreci()` function from the `PropCI` package [83] in R. The analysis script and data are provided in Sup. 9.

If a criterion was met, we coded it as "Yes", otherwise as "No". The proportions for each criterion were calculated based on the applicable denominator subset as mentioned in Tables 1–6 in Sup. 3. For two criteria (SHARE-STUDY-PROTOCOL & SHARE-SURVEY), we used the label "partially." For both, we treated "partially" as "Yes" to consider bare minimum practices in survey and protocol sharing. For STUDY-COMPENSATION, we coded "Paid with the amount mentioned," "Paid without the amount mentioned", and "Not paid (or voluntary)" as "Yes,"—as a sign of transparency in the compensation policy, and "Not mentioned" as "No." For FACE-PHOTO, we coded "Face is not clear," "Face is masked or cropped," and "Consent collected" as "Yes" since they support participants' photo privacy. For the criterion VULNERABLE, if any additional ethical measures were reported to protect the well-being of the concerned vulnerable population other than *general practices*, we coded the criterion as "Yes," otherwise "No."

To understand any potential trade-off between research ethics and transparency practices, we focused on the factor of vulnerability. Researchers usually consider data collected from vulnerable participants as sensitive and they are concerned that transparency may disclose participants' identities and cause negative consequences for them [25, 36, 88, 95]. Therefore, we distinguished between papers that deal with more ethical constraints (i.e., studies with vulnerable populations, coded as "Yes") and papers that deal with lesser ethical constraints (i.e., studies without vulnerable populations, coded as "No"). To determine the relation between ethical constraints and comparable transparency practices, we consider data sharing to be a relevant dimension of transparency since it might include sensitive information. A paper's data sharing is coded as "Yes" if either raw or processed data has been shared, irrespective of the paper being quantitative, qualitative, or mixed-method. We visualize the relationship between participant types (i.e., being vulnerable or not) with data sharing practices through a mosaic plot using the `geom_mosaic` function from the `ggplot2` package in R.

Additionally, to check for potential selection bias due to our sampling approach, we compared the proportions of papers with Best Paper awards or Honorable Mention between the two years using a two-sample Z-test.

Finally, while we defined and extracted 45 criteria, we test and visualize 41 criteria. ANIMAL, ESTIMATES-INTERVAL, and ESTIMATES-VIS-UNCERTAINTY had only $n = 1$ paper in their respective subsets. Also, for SHARE-SKETCH, we did not test differences between years because determining a meaningful denominator of this criterion requires a deep understanding of the paper's contributions and research methods.

5 RESULTS

Table 2 summarizes the characteristics of the selected papers. Most papers were mixed-method (40%) or qualitative (35%), while almost one-fourth of the papers were quantitative. Moreover, there were more qualitative papers in the CHI 2022 sample (39%) compared with quantitative papers (20%), whereas in the CHI 2017[11] sample, there were equal amounts of qualitative and quantitative papers. The increase in the number of qualitative over quantitative papers might be due to the COVID-19 pandemic which might have limited the quantitative empirical research practices during the lockdown (e.g., in-person lab experiments).

In both years, around one-fifth of the papers conducted confirmatory research while the rest conducted exploratory research. In terms of contribution, all the papers were empirical. Some papers also had other contributions, with artifact contribution being the next most common in both samples (i.e., 54% of CHI'17 and 59% of CHI'22 papers). The vast majority of the papers in both years (i.e., > 98%) recruited human participants for data collection or data annotation, whereas the rest used datasets (i.e., human data collected earlier). At least one-fifth of the selected papers received either a Best Paper award or a Honorable Mention. Although the number of awarded papers was greater in the CHI'22 subset (25%) compared with the CHI'17 subset (21%), this difference was not statistically significant ($Z = 0.74$, $p = .46$). The results showed that our samples did not bias toward a higher-quality paper in one year than the other.

5.1 Changes in Research Ethics

Among the criteria for research ethics, we found good improvements in CHI'22, where four out of seven criteria showed better adherence to research ethics (see Figure 4). Practices about acquiring IRB approval (IRB), *reporting* consent collection (CONSENT), and being transparent with participant compensation (STUDY-COMPENSATION) were all almost doubled during the last five years. While these findings show substantial improvements in ethics criteria, these practices still have more room for improvement as they were observed only in around half of the CHI'22 papers. We also observed evidence of transparency and ethics in CHI'22, where four papers shared their complete consent form as supplementary material. With regards to participant compensation (see Figure 5A), the most common approach was mentioning the exact amount or type of compensation (32%), whereas a few papers reported payment without any detail (3%).

Practices regarding preserving photo privacy (FACE-PHOTO) and anonymization (ANON) did not change between CHI'17 and CHI'22 (see Figure 4). 27% of the papers used participants' photos in their paper figures. Among the papers with participant's photos, 42% did not show any protective measures (see Figure 5B). The two top measures were (i) not depicting participants' faces clearly (e.g.,

[11] Henceforth, in the results section, we will refer to CHI 2017 and CHI 2022 as CHI'17 and CHI'22, respectively.

Table 2: Characteristics of the paper samples in CHI 2017 and CHI 2022.

		CHI 2017	CHI 2022	Total
Method	Mixed-method papers	44 (37.3%)	53 (41.7%)	97 (39.6%)
	Qualitative papers	37 (31.4%)	49 (38.6%)	86 (35.1%)
	Quantitative papers	37 (31.4%)	25 (19.7%)	62 (25.2%)
Hypothesis testing	Exploratory research	94 (79.7%)	100 (78.7%)	194 (79.2%)
	Confirmatory research	24 (20.3%)	27 (21.3%)	51 (20.8%)
Contribution	Empirical	118 (100.0%)	127 (100.0%)	245 (100.0%)
	Artifact	64 (54.2%)	75 (59.1%)	139 (56.7%)
	Methodological	7 (5.9%)	8 (6.3%)	15 (6.1%)
	Theoretical	0 (0.0%)	9 (7.1%)	9 (3.7%)
	Literature survey	5 (4.2%)	1 (0.8%)	6 (2.4%)
	Dataset	0 (0.0%)	3 (2.4%)	3 (1.2%)
	Opinion	0 (0.0%)	3 (2.4%)	3 (1.2%)
Participant	Papers with human participants	116 (98.3%)	125 (98.4%)	241 (98.4%)
	Papers with animal participants	0 (0.0%)	1 (0.8%)	1 (0.4%)
Award	Papers with award	25 (21.2%)	32 (25.2%)	57 (23.3%)
Total	Total	118	127	245

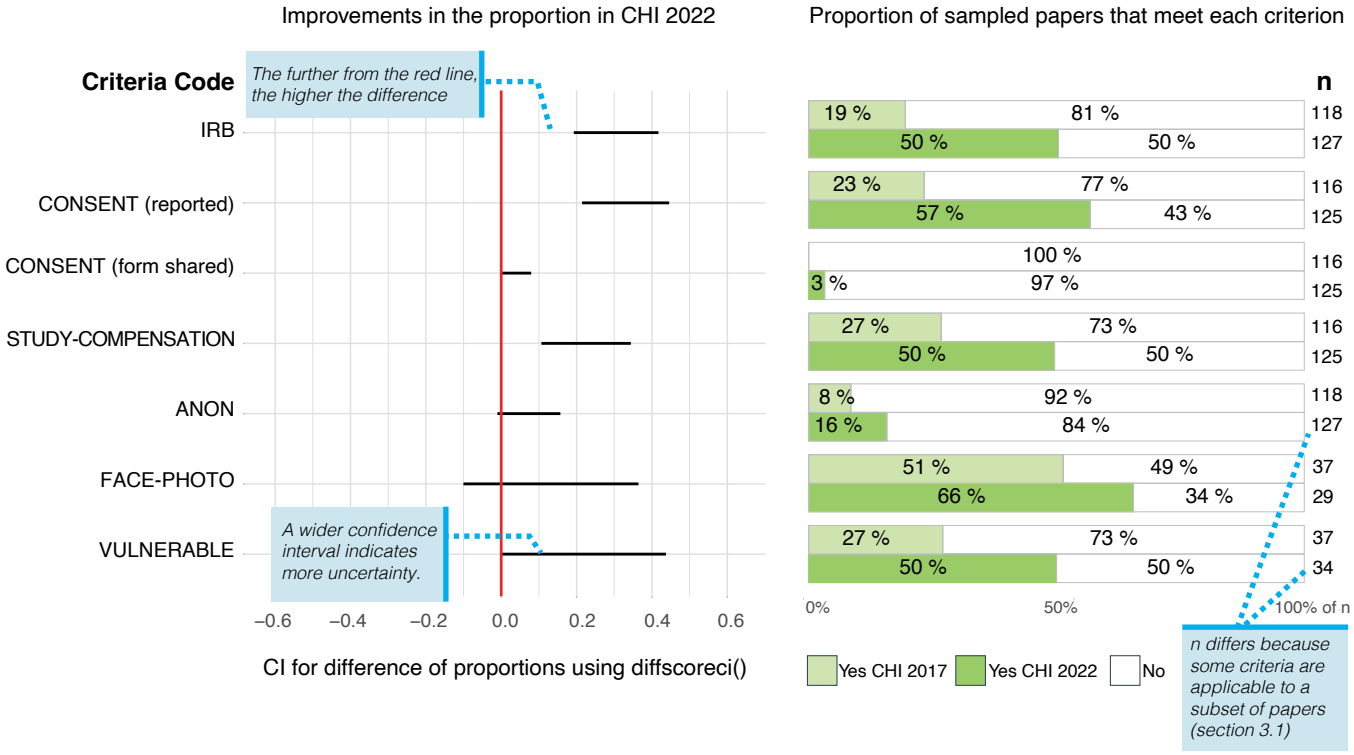

Figure 4: Research Ethics: (Right) Proportion of sampled papers meeting each of the ethics-related criterion. (Left) the difference in CI of the proportions between CHI'17 and CHI'22. CI on the right of the red line indicates improvements in CHI'22. n represents the number of papers applicable to each criterion.

photos taken from the back side) and (ii) obfuscating their faces (e.g., masking). Surprisingly, only a few papers (8%) reported collecting consent from participants before publishing their photos.

Regarding research with vulnerable populations, we found that 29% of the papers used data of participants from a vulnerable population (VULNERABLE) such as minorities or children. Among the papers with vulnerable populations, awareness about research ethics in the CHI'22 papers was higher than in the CHI'17 papers (50% vs. 27%), however the confidence interval is close to zero, suggesting that at best the improvement is negligible. Figure 5C shows the details of these vulnerabilities. The most frequent types of participants were people with disabilities (27%), potentially vulnerable students (14%), and children (11%).

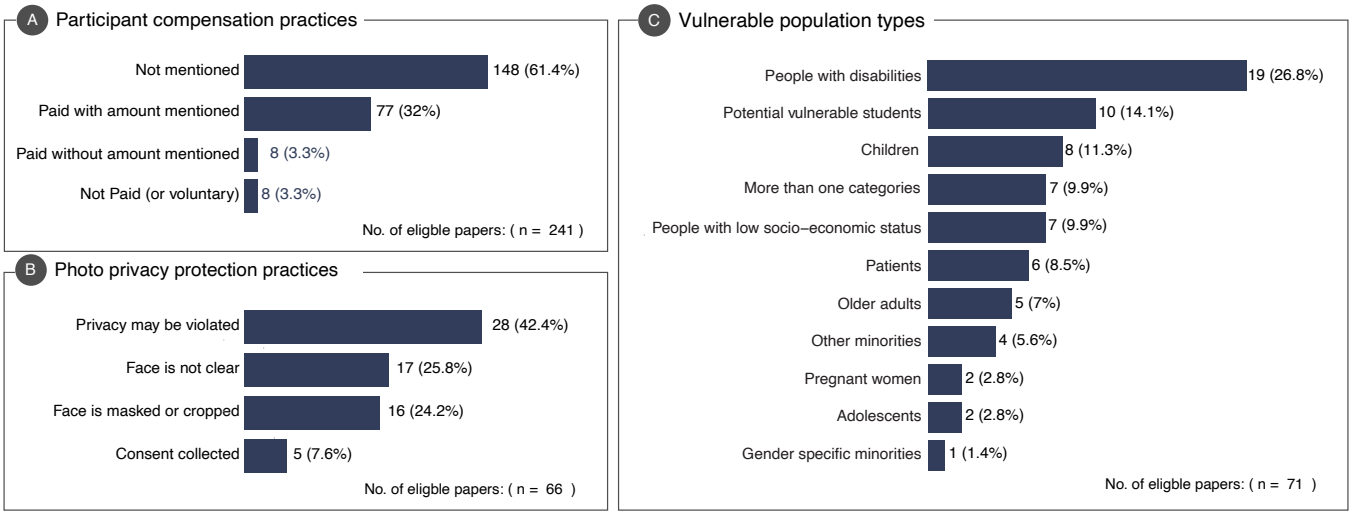

Figure 5: A. Summary of participant compensation practices in CHI'17 and CHI'22 papers . B. Authors' practices with regard to photo privacy. We acknowledge that consent for publishing photos might be collected verbally, but potential consent collection was not reported in the paper. C. Summary of the vulnerability identified in papers involving participants from vulnerable populations.

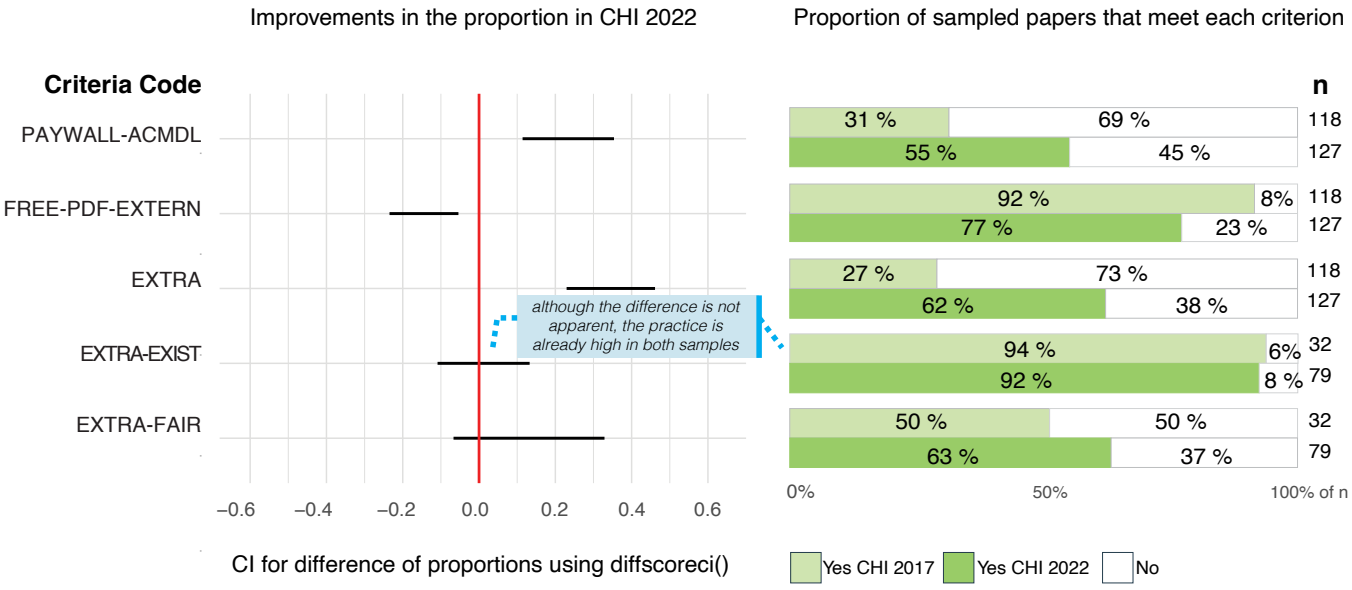

Figure 6: Openness practices: (Left) The difference in CI of the proportions between CHI'17 and CHI'22. CI on the right of the red line indicates improvements in CHI'22. (Right) Proportion of sampled papers meeting each of the openness-related criterion. n represents the number of papers applicable to each criterion.

In our sample, we only found one paper with animal participants (ANIMAL). The paper did not report any ethical measures.

5.2 Changes in Openness Practices

Figure 6 summarizes our findings about openness practices. On the ACM DL, papers will eventually be available without a paywall (PAYWALL-ACMDL) if they are either open access or public access (i.e., eventually publicly accessible after an embargo period). There are 31% from CHI'17 and 55% from CHI'22, with either open access or public access. The number of open-access papers in both samples outnumbered public-access papers (82 vs. 25). The number of papers that were accessible on other platforms (FREE-PDF-EXTERN) was relatively higher. 77% of the papers from CHI'22 and 92% from CHI'17 were available in external platforms. Table 3

Table 3: Summary of external sources used for sharing PDFs.

Type	CHI 2017	CHI 2022	Total
Long-term archival plan (e.g., ArXiv)	37 (34.3%)	34 (34.7%)	71 (34.5%)
Transitory (e.g., personal website)	37 (34.3%)	39 (39.8%)	76 (36.9%)
Commercial (e.g., ResearchGate)	34 (31.5%)	14 (14.3%)	48 (23.3%)
Long-term but accidental (e.g., Wayback Machine)	0 (0%)	11 (11.2%)	11 (5.3%)
Total	108	98	206

shows that among these external sources, only 34% of papers share on repositories with a *long-term archival plan*, for example, university/institutional/library research information systems, OSF, or ArXiv. In contrast, a slightly higher percentage, 37%, share on personal/lab/company websites, GitHub, or Google Drive, which do not guarantee longevity (i.e., labeled as *transitory*). 23% share on *commercial* social networking websites such as ResearchGate or Semantic Scholar, which is not permitted by the ACM publication policy [4] and might incur a copyright infringement [46]. Interestingly, for 11 papers, their PDF on the ACM DL were cached by the Wayback Machine and can be found by web search. In the long term, these papers will remain publicly available. However, it is unclear why these papers were crawled and cached. For this reason, we do not recommend depending on the Wayback Machine for archiving and disseminating research. The complete breakdown for Table 3 can be found in Sup. 10.

From the readers' perspective, accessing most CHI papers should be possible, as the papers are published open access, public access, or can be found somewhere else by searching in Google Scholar.

The reason for more CHI'17 papers being accessible on external platforms can be the short period between our data collection and the release of the CHI'22 proceedings (April to July 2022). Therefore, the authors did not have an opportunity to upload their work on external platforms, or the search engines did not crawl them, prior to our data collection. Moreover, given the higher rate of open access among the CHI'22 papers, some authors might not be interested in sharing their paper elsewhere.

Next, for sharing any additional research artifacts beyond the paper (EXTRA), we found a substantial increase in sharing practices between the two CHIs, where a higher proportion of CHI'22 papers (62%) shared additional materials (vs. 27% in CHI'17), through supplementary materials using ACM DL, supplementary materials shared in external repositories such as OSF/GitHub, or appendices at the end of the papers. Additionally, with regard to the existence of purportedly shared material (EXTRA-EXIST), the ratio between the two years was very close. Among the papers that shared additional research artifacts, 94% of CHI'17 and 92% of CHI'22 papers properly provided the promised materials. We discovered nine cases of missing data from either shared repositories or appendices, as follows: project website ($n = 3$), GitHub ($n = 2$), Harvard Dataverse ($n = 1$), ACM DL ($n = 1$), appendix ($n = 1$), and a broken link ($n = 1$). Seven of these locations were not FAIR-compatible. Finally, regarding the availability of the shared research artifacts (EXTRA-FAIR), despite the higher percentage of availability in CHI'22 (63% vs. 50% in CHI'17), the confidence interval crossed zero, suggesting that at best the improvement is negligible.

5.3 Changes in Transparency Practices

We observed improvement in CHI'22 compared with CHI'17 for 10 out of 17 transparency-related criteria (see Figure 7). We found great improvements in the sharing of interview protocols of qualitative papers (SHARE-INTERVIEW-GUIDE). While only 2% of the CHI'17 papers shared their interview protocols, this ratio increased to 25% in CHI'22. Such improvements were also seen in other aspects of qualitative papers where more papers from the CHI'22 sample clearly specify their data analysis procedure (SPECIFY-QUAL-ANALYSIS: 79% in CHI'22 vs. 58% in CHI'17).

A higher proportion of the CHI'22 sample shared qualitative data (QUAL-DATA-RAW: 7% vs. 0% and QUAL-DATA-PROCESSED: 17% vs. 4%). Similarly, more quantitative studies from the CHI'22 samples shared data analysis procedures (SHARE-ANALYSIS-CODE: 10% vs. 1%). Also, sharing raw and processed data (QUAN-DATA-RAW & QUAN-DATA-PROCESSED) increased in CHI'22. Despite this improvement, the overall data sharing is still low in both qualitative and quantitative studies. With regards to clarifying the sample size, for both qualitative and quantitative studies, the confidence interval capturing zero suggests that the difference is inconclusive (JUSTIFY-N-QUAL: 95% CI [-0.046, 0.144], JUSTIFY-N-QUAN: 95% CI [-0.010, 0.131]). In both samples, the majority of the papers described the demographic of their participants (~ 90%), however the difference between the two samples was negligible (DEMOGRAPHICS: 95% CI [-0.019, 0.139]). In most papers (~ 81%), from both samples, the authors clearly explained their study design, but the difference between the two samples was negligible (CONDITION-ASSIGNMENT: 95% CI [-0.106, 0.192]).

We also found improvements in sharing study protocols (SHARE-STUDY-PROTOCOL) or multimedia stimuli (SHARE-STIMULI) that could facilitate replicability, however, the ratios are still very low (9–13%). Surprisingly, there is no increase in sharing surveys or questionnaire materials (SHARE-SURVEY: 95% CI [-0.056, 0.221]).

Additionally, based on the confidence interval, we cannot be certain about the improvement in sharing software and hardware (SHARE-SOFTWARE: 95% CI [-0.058, 0.199]; SHARE-HARDWARE: 95% CI [-0.059, 0.079]). For SHARE-SKETCH, we found eight CHI'17 and five CHI'22 papers that shared their sketches (i.e., we did not test the difference, see the last paragraph of Section 4.3). Finally, while there were no CHI'17 papers which preregistered their studies, we found seven cases in CHI'22 with preregistration (PREREG). This improvement equates to only 6% of CHI'22 papers preregistering their study, indicating that this practice is still far from perfect.

Are papers that involved more ethically concerning entities less likely to adhere to transparency practices? Earlier debates in the CHI community [25, 36, 88, 95] revealed that some transparency practices and research ethics might be at odds with each other. Researchers in sensitive domains may need to contend with research decisions that sacrifice transparency for ethical practices. For instance, researchers might need to forgo sharing data when conducting research with vulnerable populations to minimize the likelihood of making the participants identifiable. For this reason, the prevalence of transparency practices could be dramatically different for research where ethical concerns are dominant as opposed to other research.

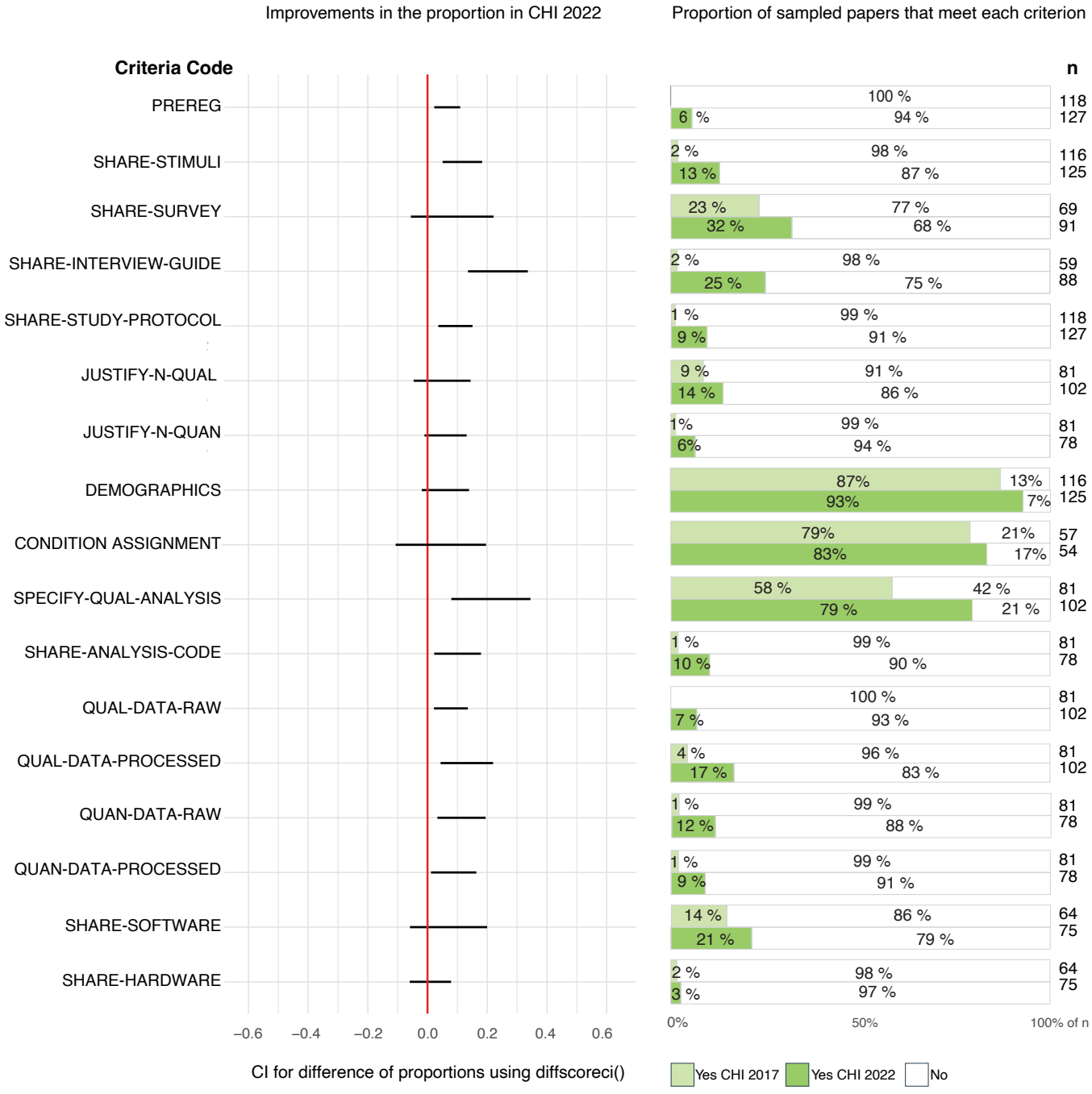

Figure 7: Transparency practices: (Left) The difference in CI of the proportions between CHI'17 and CHI'22. CI on the right of the red line indicates improvements in CHI'22. (Right) Proportion of sampled papers meeting each of the transparency-related criterion. n represents the number of papers applicable to each criterion.

As a preliminary investigation, we used the vulnerability of study participants as a proxy for ethical concerns. We divided all sampled papers into two groups: those with vs. without study participants from a vulnerable population (71 vs. 174 papers). We compared the availability of *any* type of data. We show the results in a mosaic plot in Figure 8 to emphasize a difference in the number of papers for the two groups. We found that 17% of the papers with non-vulnerable participants shared at least one type of data: either raw or processed and qualitative or quantitative. Only 8% of the papers with vulnerable populations shared their data. However, the

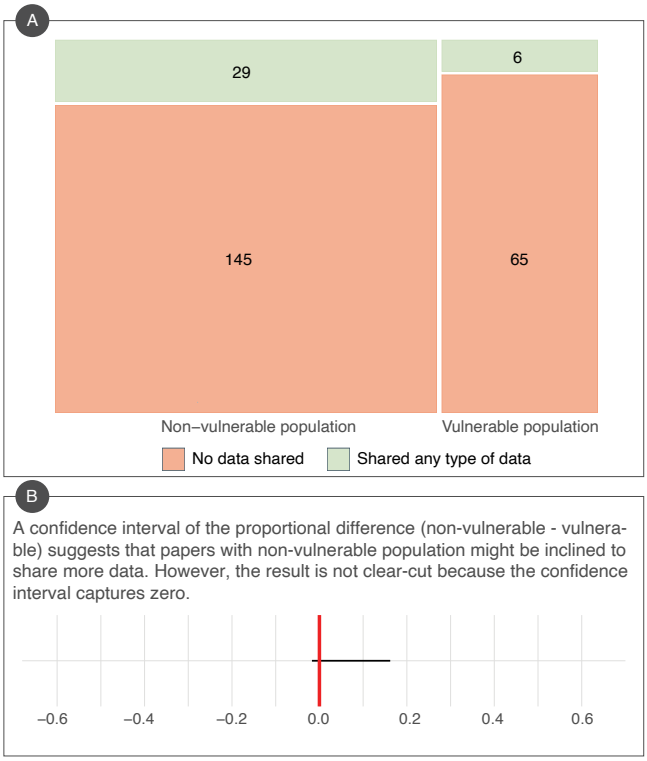

Figure 8: Ethics–Transparency Trade-off: Visualization of data sharing practices based on the involvement of participants from vulnerable populations.

confidence interval (95% CI [-0.017, 0.162]) indicates that, at best, the difference is negligible. This preliminary result only slightly supports the ethics-transparency trade-off and the result should be taken with caution due to the big difference between the number of papers in the two groups.

5.4 Lack of Change in Reporting Practices

Overall, the transparency practices related to reporting quantitative findings did not change between the two CHIs (see Figure 9). The findings showed that the ratios for some unchanged practices were high enough, such as reporting central tendency (STAT-DESCRIPTIVE), clarity of statistical tests (STAT-CLEAR-PROCEDURE), and reporting main statistical values (STAT-PARAMETERS) like p-value and F-value (79–100%). Surprisingly, more CHI'17 papers reported their degree of freedom. Regarding statistical assumptions, while more CHI'22 papers reported their normality assumption (STAT-NORMALITY), the use of other statistical assumptions (STAT-OTHER-ASSUMPTIONS) did not improve (95% CI [-0.087, 0.215]). We also checked the report on effect size (STAT-EFFECT-SIZE) and confidence interval (STAT-CI). These numbers were reported slightly more in CHI'22, while more than half of the quantitative papers in CHI'22 reported effect size, only around one-fifth reported confidence intervals for reporting data variability. However, the differences are inconclusive given the confidence intervals capture zero (STAT-EFFECT-SIZE: 95% CI [-0.034, 0.0.323], STAT-CI: 95% CI [-0.050, 0.226]).

We found only one paper with estimation analysis, where it properly reported data using interval estimates and visualized confidence intervals (ESTIMATES-INTERVAL & ESTIMATES-VIS-UNCERTAINTY).

Finally, reporting practices for qualitative results improved (QUAL-INTERVIEW-REPORT). While only 64% of CHI'17 papers properly reported their qualitative data, the rate was 90% in CHI'22.

6 A PROOF-OF-CONCEPT SCREENING TOOL

We proposed 45 criteria for research ethics, openness, and transparency. This sheer number of criteria could be prohibitive for authors and reviewers to keep in mind. We envision a future where a tool for research ethics, openness, and transparency is integrated into a writing environment, similar to spelling and grammar checkers. As the authors finish drafting each section of their paper, the tool assesses their text and reminds them to consider relevant criteria. The user can then (1) add information, (2) tell the tool to remind them later, or (3) decide that the suggestion is incorrect or irrelevant to their research method or domain. Reviewers will be assisted by a different tool: After reading the paper, the reviewer can go over the list of criteria and click on relevant criteria that the reviewer forgot to pay attention to during their first read. The tool will point to the locations in the text that satisfy the criterion or indicate that it could not find the text. The reviewer can use this feedback to selectively read the paper to verify. During the discussion phase, the lists of criteria from all reviewers are tabulated to provide a basis for discussion. In this vision, human authors and reviewers play an active role in making judgments. Their roles are necessary because of their knowledge about the research method, the domain, and the research settings.

To enable these tools, we need a system that can detect whether the text meets a criterion. Below, we describe design considerations, a proof-of-concept system, and a preliminary evaluation on eight criteria. The `Python` code for this proof-of-concept system is open-source at GitHub[12] for future research.

6.1 Design considerations

Some criteria apply to a subset of papers, for example, statistical reporting criteria do not apply to qualitative papers. Additionally, fulfilling one criterion may require sacrificing others. Combining a set of criteria into one score might inhibit nuanced discussion. Finally, one paper may present a combination of multiple studies that use different methods. Therefore, **each criterion should be evaluated independently (D1) at the level of the sentence or group of sentences (D2).**

An ideal system should be accurate in both (1) giving a positive response for a paper that satisfies the criterion (true-positive), and (2) giving a negative response for a paper that does not meet the criterion (true-negative). In reality, there is a trade-off between these goals. For example, a system could achieve a perfect true-negative rate by simply labeling that no sentences satisfy the criterion. However, this approach would incur false-negatives: sentences that actually satisfy the criterion are left undetected. This approach

[12]See https://github.com/petlab-unil/replica

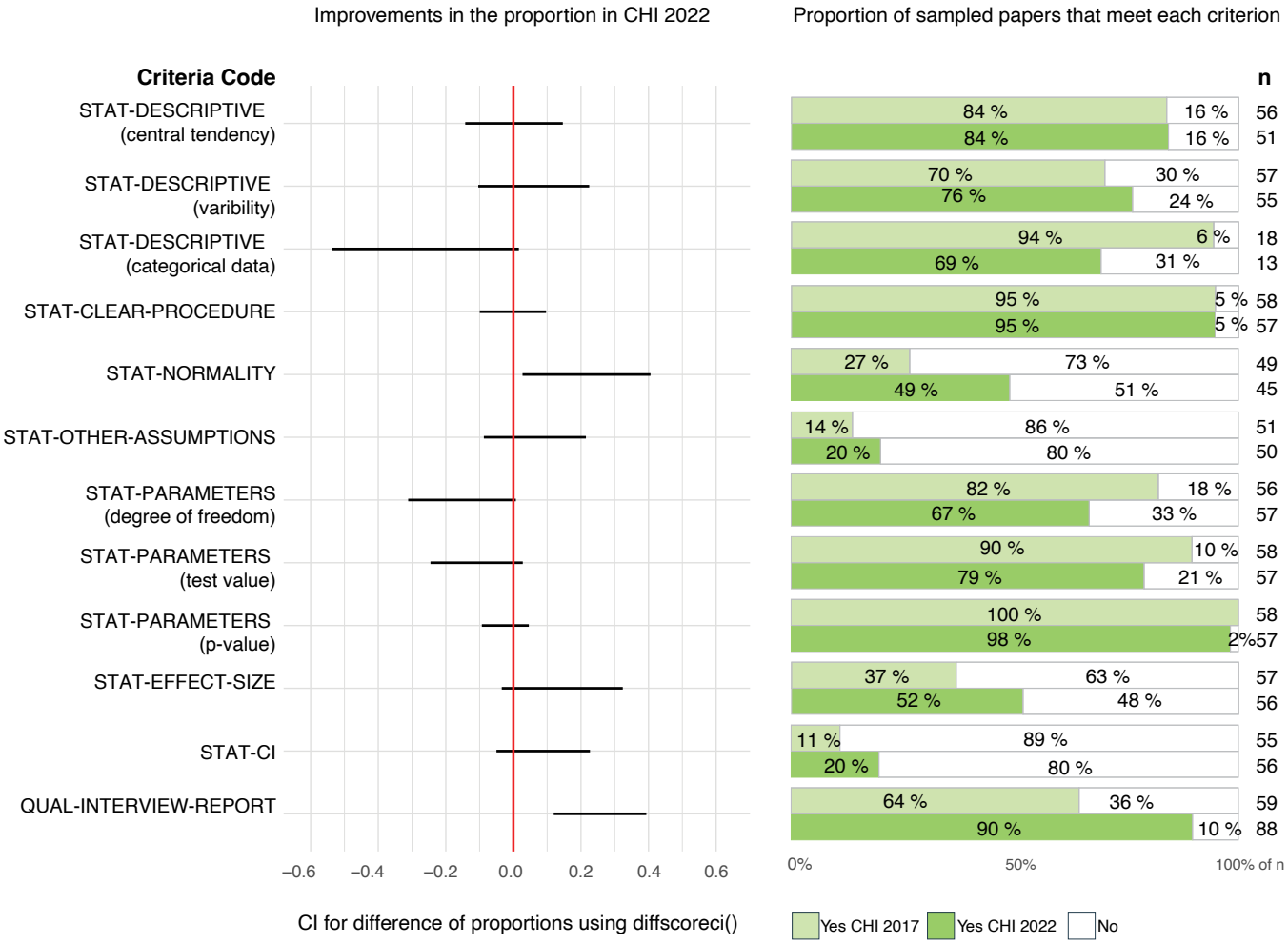

Figure 9: Reporting practices: (Left) The difference in CI of the proportions between CHI'17 and CHI'22. CI on the right of the red line indicates improvements in CHI'22. (Right) Proportions of sampled papers meeting each of the reporting criterion. n represents the number of papers applicable to each criterion. For the degrees of freedom criterion, we exclude two papers because they used path analysis and Cox Regression. To our knowledge, degrees of freedom are not conventionally reported for each test in these models—perhaps to retain readability. The implementation of these models in R also did not output degrees of freedom per test.

is also unhelpful because the whole paper needs to be manually checked. As mentioned in our vision at the beginning of Section 6, both the authors and the reviewers who use such tools will have already been familiar with the paper's content. For the text that is likely to fulfill the criterion, the system should bias towards highlighting the text rather than missing it. In other words, the system should **prioritize reducing false-negatives (D3)**. However, too many false-positives could be distracting for the users. Therefore, the system should provide a possibility for the user to **narrow down the positive results to the most confident ones (D4)**.

6.2 Implementation

We implemented a proof-of-concept system that detects how each sentence satisfies a criterion. The system architecture is shown in Figure 10. After preprocessing the PDF into a set of individual sentences, each sentence is independently analyzed in two steps: First, the system determines how similar the input sentence is with any reference sentences. Second, for each sentence that is adequately similar, the system assigns a probability that the sentence could be labeled by each of the criterion's keywords. The input sentences that pass both tests are positive results. Any paper with a positive sentence is classified as satisfying the criterion. Below are the implementation details.

6.2.1 Preprocessing PDF into sentences. The paper PDF files were processed with the `pdfminer.six` library,[13] resulting in the text with information such as hierarchical structure, font style, and blank spaces. We use this information to distinguish the body text from

[13]See https://github.com/pdfminer/pdfminer.six, last accessed January 2023.

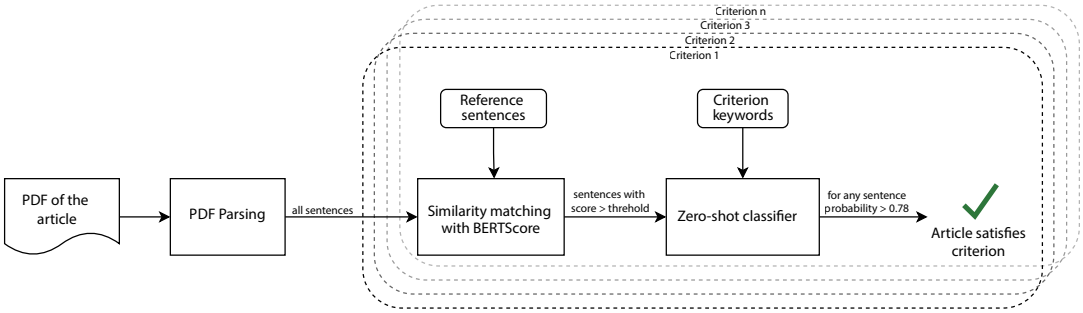

Figure 10: Architecture of the proof-of-concept screening tool

section titles. The body text was then segmented into sentences. Each sentence was individually used as input for the next two steps (D2).

6.2.2 Filtering based on sentence-similarity. For each criterion (D1), we manually extracted 5–10 *reference sentences* that make the papers fulfill the criterion. Here are two examples of the reference sentences for the IRB criterion: *"The study had institutional research ethics approval.", "The University of [...] institutional review board ([...]) approved our study."*. Input sentences that are adequately similar to any reference sentence are positive results.

The similarity is scored with the BERTScore method [112] with contextual embedding from a pre-trained language model DistilBERT [82]. To compare an input sentence to a reference sentence, the system used the language model to convert each word into a vector that encodes its contextual information. The cosine similarity is computed with the vectors of the input and the reference sentence word tokens. This approach is superior to exact or approximate pattern-matching because it does not restrict the matching to specific grammatical roles. For example, the two reference sentences above would have a high similarity score despite having their subject and object reversed.

There are three types of BERTScore: precision, recall, and F1. Precision is calculated based on greedy-matching the input to the reference, whereas recall is calculated in the opposite direction. The F1 score is a harmonic mean of precision and recall—eliminating the emphasis on any direction of the comparison. For this reason, we chose BERTScore's F1 as the similarity score.

When the score exceeds a threshold, the input sentence is a hit. This threshold hyperparameter is empirically derived for each criterion by testing the system with a small set of random samples. We set the threshold relatively low to reduce the false negatives (D3). The output of this step could be used for screening purposes, where the system highlights the hits and human authors or reviewers look at the hits to confirm.

6.2.3 Further narrowing down the hits with a text classifier. The number of hits can be further narrowed down to reduce the false positives (D4). This step is formulated as *Definition-Wild Zero-shot Text Classification* task [111]. The classifier infers the probability that a keyword *entails*, that is, logically follows, the input sentence. For example, the input sentence *"The study had institutional research ethics approval"* can be logically followed by *"This example is about IRB."*, which is created from the `IRB` as the keyword. The keywords are drawn from the list in Sup. 5.

A sentence with adequately high entailment probability to any of the keywords is a hit. Any paper with a hit is considered to satisfy the criterion. The entailment probability threshold was empirically determined to be 0.78 for all criteria. This approach requires no other training data. It is scalable for the large set of criteria we presented in this paper and possibly additional criteria in the future.

We deviated from Yin et al. [111]'s work by using the BART-large language model pre-trained on the MNLI dataset [102] because it was found to perform better than the BERT model.[14] Both this and the previous step were implemented with `Pytorch` [70] and `HuggingFace` [109] libraries.

6.3 Evaluation

We assessed the system on eight criteria, selected to strike a good balance between elements that are simpler to identify (e.g., IRB) and those that are difficult to assess (e.g., JUSTIFY-N-QUANT). The results are shown in Table 4. Most of the criteria have an imbalanced class distribution: There are many more papers that do not satisfy the criteria than those that do. To account for this imbalance, we report the precision, recall, and F1 score along with the accuracy to better understand the tool's true performance [48, 77]. A higher recall indicates a higher number of true positives and a lesser number of false negatives identified by the tool. Higher precision means lesser false positives indicating that if a higher number of positives were identified, then they were really positive in human coding. The F1 score is the harmonic mean of precision and recall indicating the balance between the two.

As explained in Section 6.2, each criterion was evaluated independently (D1) in two steps of filtering based on similarity and narrowing down with a text classifier. However, the definition of CONDITION-ASSIGNMENT and DEMOGRAPHICS are more granular in nature with each having three and four sub-criteria, respectively (see Sup. 2 for the sub-criteria under CONDITION-ASSIGNMENT and DEMOGRAPHICS). Hence, the evaluation is done in two steps for each sub-criteria under CONDITION-ASSIGNMENT and DEMOGRAPHICS. Finally, we performed a logical operation between their respective sub-criteria – logical AND for CONDITION-ASSIGNMENT and a logical OR for DEMOGRAPHICS to determine the outcome.

[14] See https://joeddav.github.io/blog/2020/05/29/ZSL.html, last accessed January 2023.

Table 4: Evaluation of our tool using Accuracy, F1 scores, Precision and Recall for eight example criteria. The SciScore paper [60] tested three criteria in common with our work. We provided the results from their paper for comparison.

	Our tool					Results from SciScore [60]		
Criterion	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**# of papers that meet a criterion**	**F1 Score**	**Precision**	**Recall**
IRB	0.89	0.85	0.85	0.85	87	0.81	0.85	0.80
CONSENT (reported)	0.81	0.78	0.73	0.84	98	0.95	0.96	0.93
STUDY-COMPENSATION	0.83	0.79	0.70	0.89	85	-	-	-
ANON	0.68	0.37	0.24	0.77	30	-	-	-
PREREG	0.99	0.80	0.75	0.86	7	-	-	-
JUSTIFY-N-QUANT	0.99	0.83	0.83	0.83	6	0.65	0.74	0.60
DEMOGRAPHICS	0.87	0.92	1	0.85	217	-	-	-
CONDITION-ASSIGNMENT	0.81	0.74	0.74	0.75	90	-	-	-

The criteria IRB, PREREG, JUSTIFY-N-QUANT, and DEMOGRAPHICS perform well in both accuracy and F1. The criteria CONSENT and STUDY-COMPENSATION have a very high accuracy and reasonable F1 score. They have a higher recall than precision, indicating that a very small number of the articles satisfying the criteria were missed, but had more false positives. These results indicate that the system satisfies design consideration D3 as expected. The criterion CONDITION-ASSIGNMENT was challenging for the tool because some independent variables are implicit. For example, a longitudinal study may not explicitly state that *time* is its independent variable; therefore, this paper could be misclassified as unmet the criterion. We observed high recall for ANON. However, we also found many false positives since the model could not distinguish between the lines referring to anonymization, participant codes, and data exclusion.

We compared our results with `SciScore`[15] to further validate our tool. `SciScore` is a proprietary automated tool which assesses research articles based on their adherence to criteria on rigor and transparency in biomedical science [12, 60]. There were three criteria that SciScore and our work have in common: IRB, CONSENT and JUSTIFY-N-QUANT. Our system yielded higher F1 on IRB and JUSTIFY-N-QUANT while SciScore was better on CONSENT. Since SciScore was trained on a large number of labeled sentences, whereas our approach does not involve any training, these preliminary results indicate that our approach is highly promising.

Based on these findings, we label each criterion in Table 2 on their potential for being identified with a screener tool. For instance, IRB and CONSENT, which have a conventional format of reporting, can definitely be identified with a screener tool. For criteria with high recall like ANON, the tool can be used to screen potential sections in the articles. Criteria like FACE-PHOTO and FREE-PDF-EXTERN might require a combination of natural language processing and computer vision.

Our approach could still be improved in several ways. The performance of the system depends on the efficiency of the PDF parser, which is prone to errors due to the various PDF styles. Using a more reliable format, for example, source files or HTML format could improve the performance. Our system assessed individual sentences. Incorporating the information about the section of the paper where the text is located might help in increasing the confidence of the prediction. Furthermore, the information about the research methods might help to further rule out more false positives, for example, by checking JUSTIFY-N-QUANT only for quantitative papers. Such information could be obtained at the submission time, for example, from PCS keyword checkboxes or subcommittee choices. Further, the availability of labeled data and using a model trained only on scientific articles like SciBERT [13] might also improve the results.

[15]See https://www.sciscore.com/, last accessed January 2023.

7 DISCUSSION

Overall, our findings showed positive changes in CHI 2022, where the authors of the CHI 2022 papers adhered to research ethics, openness, and transparency more than those of CHI 2017. In terms of the main practices, there are more improvements in research ethics and transparency and fewer in openness and reporting. However, despite such improvements, the overall rates are still low. For example, among the criteria related to research ethics, the highest rate was for consent forms with 57% adherence. This is indeed alarming and shows that almost half of the user studies in the most recent CHI proceedings either did not use a consent form at all or they did not report the consent collection in their papers. The report on consent collection for using photos was much lower for papers that used participants' facial photos in their figures.

The rates for transparency practices are even lower than for research ethics. For example, despite a great improvement in sharing interview protocols, 75% of the qualitative studies do not share their interview guides. Therefore, there is room for improvement in transparency practices in CHI. One of the areas that requires more improvement is artifact sharing where authors should be more mindful about sharing software and hardware designed or tested in the studies. Similarly, sample size justification for both quantitative and qualitative studies is still not a common practice in HCI.

In terms of transparency in reporting results, most of the practices for reporting quantitative statistical tests are at acceptable levels. However, the use and report of statistical assumptions, such as normality and reporting additional information including effect size and confidence intervals, should be improved to provide more insights into the results. Such practices can be further improved if CHI or other HCI outlets mandate existing reporting guidelines such as APA guidelines [10]. We find an interesting improvement in reporting qualitative findings. We observed a considerable number of CHI 2017 papers that did not systematically report their qualitative findings. Even some reported conducting interviews and

did not report the findings. In contrast, in CHI 2022, most of the qualitative studies followed standard reporting practices.

Our findings showed that around 29% of the selected CHI papers conducted research with vulnerable populations and thus they deal with more sensitive data and should apply stricter ethical constraints. These papers shared relatively fewer data compared with papers with non-vulnerable populations showing that ethical constraints may play against transparency. However, other reasons might exist such as lack of knowledge of software and techniques to anonymize the dataset. More in-depth studies (e.g., interviewing researchers) are required to shed light on the reasons for the lack of transparency and on how to systematically enhance transparency practices in ethically constrained studies.

What do these results suggest, in general terms? We distilled four implications ranging from measurement, creating awareness, and checking adherence to research ethics, openness, and transparency.

7.1 Self-Report Surveys vs. Actual Practices

We noticed an interesting discrepancy between our findings and the results of Wacharamanotham et al. [95], where CHI 2018 and CHI 2019 authors self-reported their transparency practices. As discussed above, the success rates of criteria for transparency practices were relatively low. For instance, the average data sharing and artifact sharing rates in CHI 2022 were 11% and 12%, respectively. However, in the study by Wacharamanotham et al. [95], the rates of similar practices were higher (17% and 40%). This difference may indicate that when researchers self-report their practices, they can be optimistic and truly believe they share what is needed to replicate their study. However, the *factual* data indicated that in practice they adhere less. This detail could also be related to our fine-grained criteria. Instead of considering data and artifact sharing as a general practice, we searched for specific practices for specific data types and artifacts (e.g., quantitative raw data).

On a different note, participants in Wacharamanotham et al. [95]'s study might indicate that they would share data upon request. In our study, we assessed the actual materials or the absence thereof. It is worth mentioning that earlier studies showed that the response rate and compliance rate for supplementary material requests are not high among the authors of papers promising to provide "data available upon request" [52]. Additionally, the chance of data availability rapidly declines when papers become older [92].

7.2 Raising Awareness

Among the three practices of transparency, openness, and ethics, the Guide to a Successful Submission on the CHI 2022 website[16] provides clearer instructions for transparency practices. This guide encourages authors of quantitative studies to ensure that their studies are reproducible. It also encourages the authors to do beta-testing to check steps taken for data collection and data analysis. Moreover, the authors of the qualitative studies are encouraged to be transparent with their study procedure and data analysis. Finally, sharing study materials and using FAIR-compatible repositories such as OSF are advised in the guide. Some of the changes we observed, such as sharing more data and data analysis procedures, might be due to the instructions in the CHI submission guideline. Surprisingly, some of the criteria (e.g., sharing via FAIR-compatible repositories) were not improved, despite being mentioned in the submission guideline. Two criteria that remain almost similar between CHI 2017 and CHI 2022 were justifying sample sizes for qualitative and quantitative studies. Interestingly, these criteria do not appear in the submission guide. We recommend that HCI venues provide specific guidelines with detailed instructions on how to meet each practice. Some of these practices require special skills and training, such as transparency practices for quantitative studies [96]. The current CHI guidelines somewhat support transparency, but they should also raise awareness about research ethics and openness. For instance, the fact that many studies did not report consent collection is somewhat worrisome. Thus, it is crucial to increase awareness and knowledge of the community to move forward in *all* aspects related to the research design, execution, and reporting.

7.3 How to Make Further Progress

A promising approach for improvement in research ethics, openness, and transparency would be for HCI journal editors or program chairs of the HCI conferences to define sharp and measurable criteria in the submission guidelines. One might think that using checklists in the submission platforms could help improve these practices where authors could skim through different practices and self-report their practices. However, the limitation of such checklists is that most authors might be optimistic while filling those forms, and they answer differently than their actual practices [95].

Another approach to improve adherence is to instruct associate chairs and reviewers about these criteria and provide specific instructions to check these upon inspection of the paper. The transparency instructions given to reviewers on the CHI 2023 website[17] are identical to those of the guide given to authors. We believe CHI should also provide specific *guidelines for reviewers* on how to assess these practices.

More recently, some venues in computer science[18] have a separate review process for the research artifacts of the accepted papers. Such a practice can support replication and reproducibility. It can also ensure that all promised data are available and adequately prepared to avoid the problem of missing data such as 8% of the papers in our samples.

However, we should acknowledge that applying a separate review process in CHI might not be feasible given the larger volume of submissions and the extra workload added to reviewers in a limited period. To reduce this workload, ideally, each aspect of criteria could be reviewed by one reviewer, either by assignment or volunteering. At a minimum, we suggest checking these criteria for the papers nominated for the "best paper award." This step ensures that at least the distinguished papers can meet the highest standards and become examples for future research.

[16] See https://chi2022.acm.org/for-authors/presenting/papers/guide-to-a-successful-submission/, last accessed January 2023.

[17] See the Transparency paragraph in Guide to Reviewing Papers at https://chi2023.acm.org/submission-guides/guide-to-reviewing-papers, last accessed January 2023.

[18] See, for example, PoPETs Artifact Review at https://petsymposium.org/artifacts.php, last accessed January 2023.

Ideally, the review process should include all submitted papers. From a futuristic perspective, this creates an opportunity for intelligent screening systems as scalable solutions to play an essential role in the review process. Ideally, paper submission platforms such as Precision Conference Solutions (PCS) can encourage authors to pass their submissions (i.e., paper and supplementary materials) over a screening system before the submission deadlines. Even if such systems are not entirely reliable, as an output, they can produce an evaluation list including approvals and warnings where the authors could go over the warnings and further clarify their practices for specific criteria not approved by the systems. PCS can be used to submit the system's output and the authors' clarification. This practice can assist reviewers in the review process by reducing their effort. Future studies on machine learning and natural language processing (NLP) should concentrate their efforts on developing reliable screening systems for assessing research ethics, openness, and transparency.

Abuse of explicit evaluation criteria and screening tools is a possibility. Authors who lack integrity can add keywords to make their manuscript satisfy the screening without actually having satisfied the criteria. These actions should be rare because they require more work (i.e., to game the system) than just complying with the requirements. However, the situation would be more precarious from a reviewer's role. For example, when a screening tool reports that some criteria are unmet, an uncaring reviewer could misuse this result to quickly dismiss the research without a proper (and fair) evaluation. Therefore, it is essential to educate the reviewers about the usage and limitations of this approach, shall such tools be used to support reviewers. Authors and reviewers should use the screening tools the same way as we use spell or grammar checker tools. These tools should assist human users in focusing their limited attention and time on areas requiring more in-depth evaluation. Additionally, the final decisions still require humans to be in the loop precisely because of trade-offs between transparent practices and compliance with research ethics. The criteria described in this work provide a concrete starting point for HCI sub-communities, whether by methodological or application domain, to discuss and develop guidelines to help authors and reviewers navigate these trade-offs. We also hope that educators will use these criteria and their subsequent refinements to educate young researchers to make their future contributions more ethical, transparent, and open.

7.4 Extra Care is Needed With Students

Among the vulnerable populations identified in our samples, the first two most frequent groups are participants with disabilities and students (see Figure 5C). Research involving people with disabilities has dedicated research communities (e.g., ASSETS conference, SIGACCESS, a dedicated CHI subcommittee) that could promote the appropriate treatment of participants through the discourses or peer review. Similarly, the student population will also benefit from a dedicated research community that ensures their equitable treatment as participants. University students are frequently used in HCI research. According to Linxen et al. [56], almost 70% of CHI 2016–2020 papers involved study participants who are university students or graduates.

We consider students a vulnerable population because, in some situations, they might be unable to protect their interests fully [84, p. 35][47]. Specifically, students might be subject to power dynamics because the evaluation of their learning progress might be conducted by the same institution recruiting them [81]. This power dynamic is particularly potent if the researchers are *directly* involved in the student's chosen courses. Without an appropriate informed consent process, coercion to participate in the study could occur directly or through an indirect assumption that it can lead to higher grades, for example. These situations could also threaten the study's internal validity by biasing students' responses in favor of the study condition they perceived as their instructors' work.

Therefore, we call for researchers to (1) avoid recruiting students in their courses or department as study participants unless strictly required by the research goal of the study or method,[19] (2) disclose power-relationship or the lack thereof explicitly, and (3) discuss ethical implications and safeguards in their paper. Reviewers should also be vigilant and demand authors to address these points. We also believe that the CHI community should examine the ethical issues of using students as study participants.

7.5 Limitations

Our study findings may be susceptible to limitations. First, the difference in the page-limit constraint between the two proceedings could be a confounding variable. CHI 2017 has a page limit per article, whereas, in CHI 2022 the authors were encouraged to adjust the length of their papers based on their contribution (i.e., not a strict guideline). The median of the page count (i.e., not including references and appendix section) for the sampled CHI 2017 papers was 10 pages, whereas it was 14 pages for CHI 2022. Therefore, this might have forced some authors to sacrifice some information or relegate it to supplementary materials. To mitigate this limitation, we thoroughly inspected the appendices and supplementary materials of the papers.

Second, our research focused on "good" research practices that can support transparency, openness, and being ethically sound. Nevertheless a good practice is not equal to "correct" practice. It was out of our scope to assess the correctness of the research practices (e.g., if a paper used the correct statistical test or if its degree of freedom matched the sample size). Although some of our criteria could be more in-depth in that respect, given that we focused on *changes* between two years, our assessment of the goodness of the practices was consistent across two years and should be reliable.

Third, our criteria list might not be exhaustive. For transparency, we could also consider reporting pre-processing steps such as data transformation and exclusion of outliers [44] (e.g., data blinding; a helpful step before analyzing data, particularly in randomized controlled studies, to reduce experimenter bias [76]). Additionally, our work only touched upon transparency criteria for qualitative research: Our criteria only cover the interview method and the generic description of analysis methods. Unlike quantitative analysis, where data analysis code details the analysis process, qualitative research has more diverse data-analytic artifacts. Not all artifacts

[19] For instance, specific research might require the researcher to take active roles in the research (e.g., *participant observation*).

are generated across all research methods. Also, research methods could differ in how research artifacts connect to transparency. For example, sharing codebooks shows transparency for coding-reliability methods such as Framework Analysis [85]. In contrast, codebooks may reveal little about the analysis process for interpretive methods. These differences call for more nuanced criteria specific to each qualitative analysis method. Lastly, our results with regard to CONDITION-ASSIGNMENT should be interpreted with caution. Although we excluded papers without user studies and with non-experiment studies, we noticed different types of studies (e.g., exploratory vs. confirmatory and basic design vs. factorial design) that may impose the authors to follow different reporting styles.

8 CONCLUSION

Within HCI and across scientific disciplines, there have been many initiatives to improve research ethics, openness, and transparency within empirical research in recent years. In this study, we show the current status quo of adoption of research ethics, openness, and transparency in HCI by assessing the changes in CHI literature between CHI 2017 and CHI 2022. This work makes the following contributions: We gathered pertinent criteria for research ethics, openness, and transparency, and operationalized them for evaluation based on published papers and research materials. We present the current state of practices in these issues and evaluate any developments between CHI 2017 and CHI 2022. Furthermore, we propose a proof-of-concept screening system to assess a certain subset of criteria. This study shows that adherence to these practices is improving overall. However, the HCI community still needs to become more mature by setting the highest standards in terms of research ethics, openness, and transparency. We hope that studies like this one will contribute to raise awareness and standards.

ACKNOWLEDGMENTS

We would like to thank Vincent Vandersluis for proofreading this article and the ACM DL team for providing a clear answer to our question about the ACM publication policy. Lastly, we sincerely thank the anonymous reviewers for their very constructive feedback and encouragement.

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 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300692

[2] ACM. 2019. ACM Policy on Submission, Hosting, Access, and Ownership of Digital Artifacts. https://www.acm.org/publications/policies/digital-artifacts

[3] ACM. 2019. Permanent Access. https://www.acm.org/publications/policies/permanent-access

[4] ACM. 2022. Open Access Publication & ACM. https://www.acm.org/publications/openaccess

[5] Balazs Aczel, Barnabas Szaszi, Alexandra Sarafoglou, Zoltan Kekecs, Šimon Kucharský, Daniel Benjamin, Christopher D. Chambers, Agneta Fisher, Andrew Gelman, Morton A. Gernsbacher, John P. Ioannidis, Eric Johnson, Kai Jonas, Stavroula Kousta, Scott O. Lilienfeld, D. Stephen Lindsay, Candice C. Morey, Marcus Munafò, Benjamin R. Newell, Harold Pashler, David R. Shanks, Daniel J. Simons, Jelte M. Wicherts, Dolores Albarracin, Nicole D. Anderson, John Antonakis, Hal R. Arkes, Mitja D. Back, George C. Banks, Christopher Beevers, Andrew A. Bennett, Wiebke Bleidorn, Ty W. Boyer, Cristina Cacciari, Alice S. Carter, Joseph Cesario, Charles Clifton, Ronán M. Conroy, Mike Cortese, Fiammetta Cosci, Nelson Cowan, Jarret Crawford, Eveline A. Crone, John Curtin, Randall Engle, Simon Farrell, Pasco Fearon, Mark Fichman, Willem Frankenhuis, Alexandra M. Freund, M. Gareth Gaskell, Roger Giner-Sorolla, Don P. Green, Robert L. Greene, Lisa L. Harlow, Fernando Hoces de la Guardia, Derek Isaacowitz, Janet Kolodner, Debra Lieberman, Gordon D. Logan, Wendy B. Mendes, Lea Moersdorf, Brendan Nyhan, Jeffrey Pollack, Christopher Sullivan, Simine Vazire, and Eric-Jan Wagenmakers. 2020. A Consensus-Based Transparency Checklist. *Nature Human Behaviour* 4, 1 (Jan. 2020), 4–6. https://doi.org/10.1038/s41562-019-0772-6

[6] Lena Fanya Aeschbach, Sebastian A.C. Perrig, Lorena Weder, Klaus Opwis, and Florian Brühlmann. 2021. Transparency in Measurement Reporting: A Systematic Literature Review of CHI PLAY. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (Oct. 2021), 233:1–233:21. https://doi.org/10.1145/3474660

[7] Alan Agresti. 2011. Score and Pseudo-Score Confidence Intervals for Categorical Data Analysis. *Statistics in Biopharmaceutical Research* 3, 2 (May 2011), 163–172. https://doi.org/10.1198/sbr.2010.09053

[8] Herman Aguinis and Angelo M. Solarino. 2019. Transparency and Replicability in Qualitative Research: The Case of Interviews with Elite Informants. *Strategic Management Journal* 40, 8 (2019), 1291–1315. https://doi.org/10.1002/smj.3015

[9] Alissa N. Antle. 2017. The Ethics of Doing Research with Vulnerable Populations. *Interactions* 24, 6 (Oct. 2017), 74–77. https://doi.org/10.1145/3137107

[10] APA. 2020. *Publication Manual of the American Psychological Association, Seventh Edition*. American Psychological Association (APA), IL, USA. https://apastyle.apa.org/products/publication-manual-7th-edition

[11] 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)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3411764.3445584

[12] Anita Bandrowski and Martijn Roelandse. 2022. SciScore, a Tool That Can Measure Rigor Criteria Presence or Absence in a Biomedical Study. In *The 1st International Conference on Drug Repurposing*. ScienceOpen, Maastricht, Netherlands, 2. https://doi.org/10.14293/S2199-1006.1.SOR-.PPPXBQN6.v1

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[14] Pernille Bjorn, Casey Fiesler, Michael Muller, Jessica Pater, and Pamela Wisniewski. 2018. Research Ethics Town Hall Meeting. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork (GROUP '18)*. Association for Computing Machinery, New York, NY, USA, 393–396. https://doi.org/10.1145/3148330.3154523

[15] Ángel Borrego and Francesc Garcia. 2013. Provision of Supplementary Materials in Library and Information Science Scholarly Journals. *Aslib Proceedings: New Information Perspectives* 65, 5 (Jan. 2013), 503–514. https://doi.org/10.1108/AP-10-2012-0083

[16] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[17] Barry Brown, Alexandra Weilenmann, Donald McMillan, and Airi Lampinen. 2016. Five Provocations for Ethical HCI Research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 852–863. https://doi.org/10.1145/2858036.2858313

[18] Amy S. Bruckman, Casey Fiesler, Jeff Hancock, and Cosmin Munteanu. 2017. CSCW Research Ethics Town Hall: Working Towards Community Norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 113–115. https://doi.org/10.1145/3022198.3022199

[19] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[20] Paul Cairns. 2007. HCI... Not as It Should Be: Inferential Statistics in HCI Research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but Not as We Know It - Volume 1 (BCS-HCI '07)*. BCS Learning & Development Ltd., Swindon, GBR, 195–201.

[21] Paul Cairns. 2019. *Doing Better Statistics in Human-Computer Interaction*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108685139

[22] Peter Celec. 2004. Open Access and Those Lacking Funds. *Science* 303, 5663 (March 2004), 1467–1467. https://doi.org/10.1126/science.303.5663.1467c

[23] Christopher D. Chambers. 2013. Registered Reports: A New Publishing Initiative at Cortex. *Cortex* 49, 3 (March 2013), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

[24] Mauro Cherubini, Kavous Salehzadeh Niksirat, Marc-Olivier Boldi, Henri Keopraseuth, Jose M. Such, and Kévin Huguenin. 2021. When Forcing Collaboration Is the Most Sensible Choice: Desirability of Precautionary and Dissuasive Mechanisms to Manage Multiparty Privacy Conflicts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 53:1–53:36. https://doi.org/10.1145/3449127

[25] Lewis L. Chuang and Ulrike Pfeil. 2018. Transparency and Openness Promotion Guidelines for HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3170427.3185377

[26] James Clifford (Ed.). 1990. *Notes on (Field) Notes.* Cornell University Press, NY, USA. https://www.jstor.org/stable/10.7591/j.ctvv4124m

[27] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173715

[28] Douglas Curran-Everett and Dale J. Benos. 2004. Guidelines for Reporting Statistics in Journals Published by the American Physiological Society. *Physiological Genomics* 18, 3 (Aug. 2004), 249–251. https://doi.org/10.1152/physiolgenomics.00155.2004

[29] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[30] Nathalie Percie du Sert, Amrita Ahluwalia, Sabina Alam, Marc T. Avey, Monya Baker, William J. Browne, Alejandra Clark, Innes C. Cuthill, Ulrich Dirnagl, Michael Emerson, Paul Garner, Stephen T. Holgate, David W. Howells, Viki Hurst, Natasha A. Karp, Stanley E. Lazic, Katie Lidster, Catriona J. MacCallum, Malcolm Macleod, Esther J. Pearl, Ole H. Petersen, Frances Rawle, Penny Reynolds, Kieron Rooney, Emily S. Sena, Shai D. Silberberg, Thomas Steckler, and Hanno Würbel. 2020. Reporting Animal Research: Explanation and Elaboration for the ARRIVE Guidelines 2.0. *PLOS Biology* 18, 7 (July 2020), e3000411. https://doi.org/10.1371/journal.pbio.3000411

[31] 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 (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395

[32] Morten W Fagerland, Stian Lydersen, and Petter Laake. 2015. Recommended Confidence Intervals for Two Independent Binomial Proportions. *Statistical Methods in Medical Research* 24, 2 (April 2015), 224–254. https://doi.org/10.1177/0962280211415469

[33] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. https://doi.org/10.3758/bf03193146

[34] Sebastian S. Feger, Cininta Pertiwi, and Enrico Bonaiuti. 2022. Research Data Management Commitment Drivers: An Analysis of Practices, Training, Policies, Infrastructure, and Motivation in Global Agricultural Science. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 322:1–322:36. https://doi.org/10.1145/3555213

[35] Sebastian S. Feger, Paweł W. Wozniak, Lars Lischke, and Albrecht Schmidt. 2020. 'Yes, I Comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 141:1–141:26. https://doi.org/10.1145/3415212

[36] Casey Fiesler, Christopher Frauenberger, Michael Muller, Jessica Vitak, and Michael Zimmer. 2022. Research Ethics in HCI: A SIGCHI Community Discussion. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3491101.3516400

[37] Casey Fiesler, Jeff Hancock, Amy Bruckman, Michael Muller, Cosmin Munteanu, and Melissa Densmore. 2018. Research Ethics for HCI: A Roundtable Discussion. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3170427.3186321

[38] Christopher Frauenberger, Amy S. Bruckman, Cosmin Munteanu, Melissa Densmore, and Jenny Waycott. 2017. Research Ethics in HCI: A Town Hall Meeting. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1295–1299. https://doi.org/10.1145/3027063.3051135

[39] John Furler, Parker Magin, Marie Pirotta, and Mieke van Driel. 2012. Participant Demographics Reported in "Table 1" of Randomised Controlled Trials: A Case of "Inverse Evidence"? *International Journal for Equity in Health* 11, 1 (March 2012), 14. https://doi.org/10.1186/1475-9276-11-14

[40] Aakash Gautam, Chandani Shrestha, Andrew Kulak, Steve Harrison, and Deborah Tatar. 2018. Participatory Tensions in Working with a Vulnerable Population. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2 (PDC '18)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3210604.3210629

[41] Clifford Geertz. 1976. *The Religion of Java.* University of Chicago Press, Chicago, IL. https://press.uchicago.edu/ucp/books/book/chicago/R/bo3627129.html

[42] {CHI guideline contributors in alphabetical order}, Pernille Bjørn, Fanny Chevalier, Pierre Dragicevic, Shion Guha, Steve Haroz, Helen Ai He, Elaine M. Huang, Matthew Kay, Ulrik Lyngs, Joanna McGrenere, Christian Remy, Poorna Talkad Sukumar, and Chat Wacharamanotham. 2019. *Proposal for Changes to the CHI Reviewing Guidelines.* Technical Report. Zenodo. https://doi.org/10.5281/zenodo.5566172

[43] Tamarinde L. Haven, Timothy M. Errington, Kristian Skrede Gleditsch, Leonie van Grootel, Alan M. Jacobs, Florian G. Kern, Rafael Piñeiro, Fernando Rosenblatt, and Lidwine B. Mokkink. 2020. Preregistering Qualitative Research: A Delphi Study. *International Journal of Qualitative Methods* 19 (Jan. 2020), 1–13. https://doi.org/10.1177/1609406920976417

[44] Constance Holman, Sophie K. Piper, Ulrike Grittner, Andreas Antonios Diamantaras, Jonathan Kimmelman, Bob Siegerink, and Ulrich Dirnagl. 2016. Where Have All the Rodents Gone? The Effects of Attrition in Experimental Research on Cancer and Stroke. *PLOS Biology* 14, 1 (Jan. 2016), 1–12. https://doi.org/10.1371/journal.pbio.1002331

[45] 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 (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[46] Hamid R. Jamali. 2017. Copyright Compliance and Infringement in ResearchGate Full-Text Journal Articles. *Scientometrics* 112, 1 (July 2017), 241–254. https://doi.org/10.1007/s11192-017-2291-4

[47] David W. Jamieson and Kenneth W. Thomas. 1974. Power and Conflict in the Student-Teacher Relationship. *The Journal of Applied Behavioral Science* 10, 3 (July 1974), 321–336. https://doi.org/10.1177/002188637401000304

[48] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge University Press, Cambridge, England.

[49] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1081–1084. https://doi.org/10.1145/2851581.2886442

[50] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving Transparent Statistics Forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 534–541. https://doi.org/10.1145/3027063.3027084

[51] Juho Kim, Haoqi Zhang, Paul André, Lydia B. Chilton, Wendy Mackay, Michel Beaudouin-Lafon, Robert C. Miller, and Steven P. Dow. 2013. Cobi: A Community-Informed Conference Scheduling Tool. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2501988.2502034

[52] Michal Krawczyk and Ernesto Reuben. 2012. (Un)Available upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials. *Accountability in Research* 19, 3 (May 2012), 175–186. https://doi.org/10.1080/08989621.2012.678688

[53] Daniël Lakens. 2022. Sample Size Justification. *Collabra: Psychology* 8, 1 (March 2022), 33267. https://doi.org/10.1525/collabra.33267

[54] Tom Lang and Douglas Altman. 2016. Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines. *Medical Writing* 25 (Sept. 2016), 31–36. https://journal.emwa.org/statistics/statistical-analyses-and-methods-in-the-published-literature-the-sampl-guidelines/

[55] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction - 2nd Edition.* Morgan Kaufmann, MA, USA. https://www.elsevier.com/books/research-methods-in-human-computer-interaction/lazar/978-0-12-805390-4

[56] 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)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445488

[57] Florian Mann, Benedikt von Walter, Thomas Hess, and Rolf T. Wigand. 2009. Open Access Publishing in Science. *Commun. ACM* 52, 3 (March 2009), 135–139. https://doi.org/10.1145/1467247.1467279

[58] Nora McDonald, Karla Badillo-Urquiola, Morgan G. Ames, Nicola Dell, Elizabeth Keneski, Manya Sleeper, and Pamela J. Wisniewski. 2020. Privacy and Power: Acknowledging the Importance of Privacy Research and Design for Vulnerable Populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375174

[59] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3,

CSCW (Nov. 2019), 72:1–72:23. https://doi.org/10.1145/3359174

[60] Joe Menke, Martijn Roelandse, Burak Ozyurt, Maryann Martone, and Anita Bandrowski. 2020. The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods. *iScience* 23, 11 (Nov. 2020), 101698. https://doi.org/10.1016/j.isci.2020.101698

[61] Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2022. *Qualitative Data Analysis: A Methods Sourcebook.* SAGE, CA, USA. https://us.sagepub.com/en-us/nam/qualitative-data-analysis/book246128

[62] Andrew Moravcsik. 2014. Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics* 47, 1 (Jan. 2014), 48–53. https://doi.org/10.1017/S1049096513001789

[63] Gaia Mosconi, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. 2022. Designing a Data Story: A Storytelling Approach to Curation, Sharing and Data Reuse in Support of Ethnographically-driven Research. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 289:1–289:23. https://doi.org/10.1145/3555180

[64] Cosmin Munteanu, Heather Molyneaux, Wendy Moncur, Mario Romero, Susan O'Donnell, and John Vines. 2015. Situational Ethics: Re-thinking Approaches to Formal Ethics Requirements for Human-Computer Interaction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 105–114. https://doi.org/10.1145/2702123.2702481

[65] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an Open Research Culture. *Science* 348, 6242 (June 2015), 1422–1425. https://doi.org/10.1126/science.aab2374

[66] Giovanna Nunes Vilaza, Kevin Doherty, Darragh McCashin, David Coyle, Jakob Bardram, and Marguerite Barry. 2022. A Scoping Review of Ethics Across SIGCHI. In *Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 137–154. https://doi.org/10.1145/3532106.3533511

[67] Bridget C. O'Brien, Ilene B. Harris, Thomas J. Beckman, Darcy A. Reed, and David A. Cook. 2014. Standards for Reporting Qualitative Research: A Synthesis of Recommendations. *Academic Medicine* 89, 9 (Sept. 2014), 1245–1251. https://doi.org/10.1097/ACM.0000000000000388

[68] Yuren Pang, Katharina Reinecke, and René Just. 2022. Apéritif: Scaffolding Preregistrations to Automatically Generate Analysis Code and Methods Descriptions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3517707

[69] Van L. Parsons. 2017. Stratified Sampling. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, NY, USA, 1–11. https://doi.org/10.1002/9781118445112.stat05999.pub2

[70] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/arXiv.1912.01703 arXiv:1912.01703 [cs, stat]

[71] 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)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445734

[72] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. 2016. *A Statistical Definition for Reproducibility and Replicability.* Preprint. bioRxiv. https://doi.org/10.1101/066803

[73] George Peat, Richard D. Riley, Peter Croft, Katherine I. Morley, Panayiotis A. Kyzas, Karel G. M. Moons, Pablo Perel, Ewout W. Steyerberg, Sara Schroter, Douglas G. Altman, Harry Hemingway, and for the PROGRESS Group. 2014. Improving the Transparency of Prognosis Research: The Role of Reporting, Data Sharing, Registration, and Protocols. *PLOS Medicine* 11, 7 (July 2014), e1001671. https://doi.org/10.1371/journal.pmed.1001671

[74] Marco Perugini, Marcello Gallucci, and Giulio Costantini. 2018. A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology* 31, 1 (July 2018), 20. https://doi.org/10.5334/irsp.181

[75] Kenneth D. Pimple. 2002. Six Domains of Research Ethics. *Science and Engineering Ethics* 8, 2 (June 2002), 191–205. https://doi.org/10.1007/s11948-002-0018-1

[76] Denise F. Polit. 2011. Blinding during the Analysis of Research Data. *International Journal of Nursing Studies* 48, 5 (May 2011), 636–641. https://doi.org/10.1016/j.ijnurstu.2011.02.010

[77] David .M.W Powers. 2011. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies* 2, 1 (Dec. 2011), 37–63. https://doi.org/10.48550/arXiv.2010.16061

[78] Lumpapun Punchoojit and Nuttanont Hongwarittorrn. 2015. Research Ethics in Human-Computer Interaction: A Review of Ethical Concerns in the Past Five Years. In *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*. IEEE, Ho Chi Minh City, Vietnam, 180–185. https://doi.org/10.1109/NICS.2015.7302187

[79] Camille R. Quinn. 2015. General Considerations for Research with Vulnerable Populations: Ten Lessons for Success. *Health & Justice* 3, 1 (Jan. 2015), 1. https://doi.org/10.1186/s40352-014-0013-z

[80] Judy Robertson and Maurits Kaptein (Eds.). 2016. *Modern Statistical Methods for HCI.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-26633-6

[81] Susan L. Rose and Charles E. Pietri. 2002. Workers as Research Subjects: A Vulnerable Population. *Journal of Occupational and Environmental Medicine* 44, 9 (2002), 801–805. https://doi.org/10.1097/00043764-200209000-00001

[82] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. https://doi.org/10.48550/arXiv.1910.01108 arXiv:1910.01108 [cs]

[83] Ralph Scherer. 2018. PropCIs: Various Confidence Interval Methods for Proportions. https://CRAN.R-project.org/package=PropCIs

[84] Council for International Organizations of Medical Sciences. 2017. *International Ethical Guidelines for Health-Related Research Involving Humans.* World Health Organization, Geneva, CH. https://www.cabdirect.org/cabdirect/abstract/20173377536

[85] Joanna Smith and Jill Firth. 2011. Qualitative Data Analysis: The Framework Approach. *Nurse Researcher* 18, 2 (2011), 52–62. https://doi.org/10.7748/nr2011.01.18.2.52.c8284

[86] Martin Spann, Lucas Stich, and Klaus M. Schmidt. 2017. Pay What You Want as a Pricing Model for Open Access Publishing? *Commun. ACM* 60, 11 (Oct. 2017), 29–31. https://doi.org/10.1145/3140822

[87] Jose M. Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3821–3832. https://doi.org/10.1145/3025453.3025668

[88] Poorna Talkad Sukumar, Ignacio Avellino, Christian Remy, Michael A. DeVito, Tawanna R. Dillahunt, Joanna McGrenere, and Max L. Wilson. 2020. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381066

[89] Allison Tong, Peter Sainsbury, and Jonathan 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 (Dec. 2007), 349–357. https://doi.org/10.1093/intqhc/mzm042

[90] UNESCO. 2021. UNESCO Recommendation on Open Science. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

[91] Alicia VandeVusse, Jennifer Mueller, and Sebastian Karcher. 2022. Qualitative Data Sharing: Participant Understanding, Motivation, and Consent. *Qualitative Health Research* 32, 1 (Jan. 2022), 182–191. https://doi.org/10.1177/10497323211054058

[92] Timothy H. Vines, Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24, 1 (Jan. 2014), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

[93] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 941–953. https://doi.org/10.1145/2818048.2820078

[94] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Association for Computing Machinery, New York, NY, USA, 4–18. https://doi.org/10.1145/3410404.3414229

[95] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448

[96] Chat Wacharamanotham, Fumeng Yang, Xiaoying Pu, Abhraneel Sarma, and Lace Padilla. 2022. Transparent Practices for Quantitative Empirical Research. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3491101.3503760

[97] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond 'One Size Fits All': Research Considerations for Working with Vulnerable Populations. *Interactions* 26, 6 (Oct. 2019), 34–39. https://doi.org/10.1145/3358904

[98] Dan S. Wallach. 2011. Rebooting the CS Publication Process. *Commun. ACM* 54, 10 (Oct. 2011), 32–35. https://doi.org/10.1145/2001269.2001283

[99] Shirley Wheeler. 2003. Comparing Three IS Codes of Ethics - ACM, ACS and BCS. *PACIS 2003 Proceedings* 107 (Dec. 2003), 1576–1589. https://aisel.aisnet.org/pacis2003/107

[100] Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology* 7 (2016), 12. https://doi.org/10.3389/fpsyg.2016.01832

[101] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[102] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[103] Günter Wilms. 2019. Guide on Good Data Protection Practice in Research. https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf

[104] 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 (CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 1185–1188. https://doi.org/10.1145/2212776.2212419

[105] Max L. Wilson, Ed H. Chi, Stuart Reeves, and David Coyle. 2014. RepliCHI: The Workshop II. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 33–36. https://doi.org/10.1145/2559206.2559233

[106] 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 (CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 463–466. https://doi.org/10.1145/1979742.1979491

[107] Max L. L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. 2013. RepliCHI: The Workshop. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 3159–3162. https://doi.org/10.1145/2468356.2479636

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

[109] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://doi.org/10.48550/arXiv.1910.03771

[110] Koji Yatani. 2016. Effect Sizes and Power Analysis in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 87–110. https://doi.org/10.1007/978-3-319-26633-6_5

[111] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. https://doi.org/10.48550/arXiv.1909.00161

[112] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. https://doi.org/10.48550/arXiv.1904.09675

[113] Kelly H. Zou, Julia R. Fielding, Stuart G. Silverman, and Clare M. C. Tempany. 2003. Hypothesis Testing I: Proportions. *Radiology* 226, 3 (March 2003), 609–613. https://doi.org/10.1148/radiol.2263011500