Transparency of CHI Research Artifacts: Results of a Self-Reported Survey

Chat Wacharamanotham[1] **Lukas Eisenring**[1] **Steve Haroz**[2] **Florian Echtler**[3]

[1]University of Zurich
Zurich, Switzerland

[2]Université Paris-Saclay, Inria
Saclay, France

[3]Bauhaus-Universität Weimar
Weimar, Germany

chat@ifi.uzh.ch, lukas.eisenring@uzh.ch, open_chi@steveharoz.com, florian.echtler@uni-weimar.de

ABSTRACT

Several fields of science are experiencing a "replication crisis" that has negatively impacted their credibility. Assessing the validity of a contribution via replicability of its experimental evidence and reproducibility of its analyses requires access to relevant study materials, data, and code. Failing to share them limits the ability to scrutinize or build-upon the research, ultimately hindering scientific progress.

Understanding how the diverse research artifacts in HCI impact sharing can help produce informed recommendations for individual researchers and policy-makers in HCI. Therefore, we surveyed authors of CHI 2018–2019 papers, asking if they share their papers' research materials and data, how they share them, and why they do not. The results (34% response rate) show that sharing is uncommon, partly due to misunderstandings about the purpose of sharing and reliable hosting. We conclude with recommendations for fostering open research practices.

This paper and all data and materials are freely available at https://osf.io/3bu6t.

Author Keywords

Open Science, public data sharing, open data, data availability

CCS Concepts

•Human-centered computing → Human computer interaction (HCI);

INTRODUCTION

Recently, scientific disciplines such as psychology [9], medicine [3], economics [6], and political science [16] have had many published research articles fail to replicate, i.e. rerunning the experiment with more statistical power does not yield an effect as strong as or in the same direction as the original experiment. Coding mistakes, inflated statistical statements, or decisions made after data collection began could have inflated the original effect. When many replications in a field find substantially weaker or even negative effects, the general public may question the credibility of published claims in the entire field [13, 62] and, possibly, science in general.

When discussing the validity of research, people use two similar sounding, but often confused terms: replicability and reproducibility.[1] A **replication** reruns a study to produce new data, which may or may not be analyzed using the original approach. If the effect found in the original study were true, a sufficiently powered replication should yield *similar* results. **Reproducibility** describes the ability to rerun the computational analyses on the original data. If the original analysis were done as described in the paper, the newly computed results (based on the original or a new piece of code) should be *identical* to the original. Both activities demonstrate the robustness and validity of the research claims, and both require researchers to be open and transparent by sharing research artifacts. However, the artifacts they need are different, with replication needing access to experiment materials, while reproducibility requires access to the original raw data and analysis code. When research artifacts are not public, requesting them from the original authors is unlikely to be successful (e.g., 27% success rate in psychology [54]). Such difficulty increases over time (e.g., in zoology, the odds of the original authors confirming the existence of their data reduces by 17% per year after publication [49]).

In a community as diverse as CHI, not all research contributions fit into the framework of hypothesis, experiment, and statistical validation. **Qualitative research** such as ethnographic studies may not lend themselves well to *reproduction*. Nevertheless, sharing research artifacts, e.g., interview guidelines, coding manuals, or transcripts, would facilitate evaluation and even future *replications* [32, 52].

Moreover, **software reproducibility and reusability** are particularly relevant in a computing-related field like HCI [44]. Only very few CHI papers (less than 5% in 2016–17, see [10]) provide any kind of source code or software underlying their research contribution is based on, making it more difficult to (1) assess the internal and external validity of the developed software, and also to (2) build on existing research [29, 31]. The second part, in particular, is noteworthy: without software being provided along with the paper, any work seeking to

[1] In this paper, we use the Claerbout terminology, which is widely used in science and has been given statistical definition in [40]. Unfortunately, the ACM's terminology is inverted. Plesser describes the origin of this confusion [41].

CHI '20, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376448

expand on and/or compare with this research will first have to re-implement an approximation of the prior system [44]. This approximation requires extra effort to build and may not directly comparable to the original system.

Despite being beneficial to the field, several concerns impede research artifacts from being shared e.g., the resources needed for sharing, the privacy of study participants, and data protection regulations. The multidisciplinary nature of HCI means that we generate diverse types of research artifacts, each with different level of pertinence to the knowledge being contributed in each publication. **The plurality of research artifact types and the difference in their concerns in sharing potentially confound the discussion about sharing practices in HCI.**

To provide a basis for bottom-up practices and top-down policy in the field of HCI, we present a preregistered online survey among authors of papers presented at CHI 2018–19. We highlight issues and possible directions to address common concerns in research artifact sharing.

Initiatives related to research artifact sharing in HCI

To date, no HCI publication venue has officially adopted guidelines on research transparency, such as the Transparency and Openness Promotion (TOP) Guidelines [35]. Yet their applicability for HCI research has been extensively discussed in recent years. The RepliCHI series of panels/SIGs/workshops during CHI 2012–2014 focused on the question of replication [58, 59, 60, 61]. More recently, Cockburn et al. discussed how the practice of preregistration, i.e. of publishing the study protocol and hypotheses ahead of actually conducting the study, might be adapted for the HCI context [8]. In a Special Interest Group (SIG) meeting at CHI 2018, Chuang & Pfeil lead a discussion on how the TOP guidelines match current practices at CHI [7]. Another SIG meeting, focusing on transparent statistics, was also held at CHI 2018 [50], following related workshops/SIGs in previous years [25, 26]. Echtler & Häußler analyzed papers from CHI 2016–17 and found that less than 5% provided any kind of publicly available source code [10]. From researchers practicing qualitative methods, there was a SIG at CHI 2018 focused on structured evaluation methods [43] as well as a workshop at CSCW 2019 to discuss transparent practices in qualitative research [14]. Most recently a community-led effort to improve the CHI Guides to a Successful Paper Submission also recommends sharing research artifacts. Additionally, ACM as parent organization of SIGCHI has created an Artifact Review and Badging scheme. However, as the badges do not differentiate types of artifacts such as preregistration, do not have clear criteria for being rewarded, and consist of differently colored ACM logos, their utility is questionable. Despite these efforts, a recent study shows that in the a subarea of CHI, research data are rarely published or only claimed to be shared upon request [1]. Such rarity had been criticized as a factor hindering replications [1, 23].

Concerns in research artifact sharing

Concerns about sharing research artifacts have been documented in many fields of research, but the prominence of these concerns vary with the prominence of different research artifacts (see tabulated differences in supplementary S1). Several surveys with natural scientists were conducted to understand barriers in sharing research data [47, 48]. One that is close to HCI was a survey in psychology conducted by Houtkoop et al. in 2016, which received responses from 600 researchers [24]. Their questionnaire (which was refined based on prior work: [47, 48]) classified barriers in sharing into three categories: (1) legal constraints (e.g., difficulties in anonymizing data); (2) fear-related (e.g., afraid of misinterpretation); and (3) non-fear-related (e.g., sharing is an uncommon practice in their field). Their results show that fear of misinterpretation or exposure of invalid conclusions are the most common reasons (Figure 3 ibid.). The convention reason: "sharing is not a common practice" was also agreed to by the majority of their respondents (Figure 2 ibid.). However, legal constraints are not rated as a major barrier (Figure S5 ibid.).

Concerns in sharing **qualitative data** have a slightly different focus. McGrath & Nilsonne [32] identified four areas of concerns: (1) anonymization difficulties; (2) fear of reprimands from workplace authorities; *(3) an opinion that the data has specific value in the study context*; and *(4) fear that study participants will not willingly offer their information if they know the data will be shared in the public domain*. The last two differ from any items in Houtkoop's classification.

Concerns in sharing **software code** have again different emphases. Based on the arguments by LeVeque (for small pieces of code for computational models [29]) and Barnes (for software developed in research, in general [2]), the concerns that could be mapped to Houtkoop's classification are the amount of work to polish the code, uncommon practice in the field, and intellectual property concerns. A unique concern for this type is *the fear that the code may not be runnable in the future*.

These findings suggest that concerns, or at least their severity, distribute differently by the type of research materials. In the field of HCI, these distributions could be further complicated by the diversity of research methods and the fact that each research paper could include a mixture, each with different degrees of merit [63]. Lastly, to our knowledge, no prior work looked into concerns in sharing research artifacts that are the products of design-oriented methods, e.g., prototypes. Therefore, to further the discussion of research material sharing, we need knowledge of the types of research artifacts, practices of sharing, and reasons that hinder sharing in the field of HCI.

METHOD

To better understand of research artifact sharing in HCI, we conducted two online surveys on the first authors of papers from CHI 2018–19. The surveys asked *what* research artifacts are generated, *where* are they shared, and *reasons* for not sharing them.

Overall survey design and testing

Three of the authors created and refined initial drafts. Then, between May–June 2018, we reach out to HCI researchers for comments via personal contacts and the official ACM SIGCHI and the CHI Meta Facebook groups. We received and incorporated feedback from 8 experienced HCI researchers

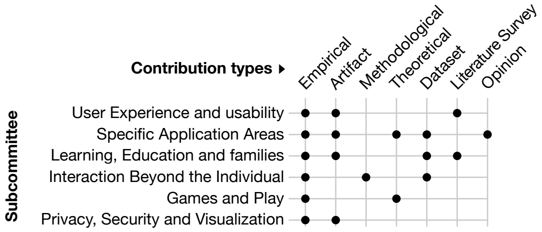


Figure 1. The distribution of the papers used to test the 2019 taxonomy. For details, see supplementary S2.

(all hold a doctoral degree and have published papers at CHI) and one person with extensive experience from the Center for Open Science (COS). The survey was then tested with two junior PhD students in HCI from two labs situated in different countries. Finally, in 2018, we rolled out the survey in two batches. The first batch included authors we know personally to identify any remaining flaws in the question structure or the user interface of the survey system. We wound up making no further changes between the two batches.

After analyzing the data in 2018, some responses indicated that our taxonomy of research artifacts did not permit qualitative data to be converted to quantitative (e.g., by counting term frequency in a transcript) and vice versa. Therefore we revised the taxonomy to explicitly enable this analysis while retaining a one-to-one mapping to the 2018 taxonomy.

To test the new taxonomy, one of the co-authors selected papers from the examples provided in the CHI 2019 subcommittee page to cover as many contribution types [63] across subcommittees as possible. This selection results in 19 papers (Figure 1). Then, he manually read and identified research artifacts generated in each paper. The results were checked by a different co-author and disagreements resolved by discussion. (For details in supplementary S2.) Then, they refined the taxonomy, which was approved by all co-authors.

A taxonomy of research artifacts in HCI

We define **research artifacts** as everything produced during a research project. The COS's open science badges lump research artifacts into either data (https://osf.io/g6u5k) or study materials (https://osf.io/gc2g8). These broad categories may be helpful as a simplified reward, but they do not adequately represent the broad variety of research methods and contributions in the field of HCI. Therefore, we developed a taxonomy of research artifacts with finer granularity, which is informed by the literature described in the section *Concerns in research artifact sharing*. For practicality, we limited the scope of our taxonomy to digitally-shareable artifacts.

The 2018 version of the taxonomy was created as a part of the survey design, and it was refined for 2019 as described below. The final taxonomy in the rest of this paper is shown in Figure 2, and the evolution of the data-related types is shown in supplementary S9.

One challenge in developing the taxonomy is that *quantitative* and *qualitative data* can be processed into either type. For example, an audio recording could be manually transcribed into text (qualitative), which could then be counted for word or event frequencies (quantitative). In contrast, the same recording could undergo spectral analysis (quantitative), and the results could be inspected and annotated by multiple researchers (qualitative). The annotations could then be used to calculate an agreement score (quantitative).

A. Study materials are produced by researchers and presented to participants to elicit their responses (e.g., visual stimuli used during experiment or questionnaires).

Raw data
- **B. Selective:** Data collected at researchers' discretion (e.g., field notes during ethnographic study)
- **C. Nonselective:** Data collected without researcher discretion at the time of collection, (e.g., task completion times logged by software)

Data processing procedure
- **D. Qualitative** (e.g., coding manual)
- **E. Quantitative** (e.g., statistics analysis script)

Processed data
- **F. Output from qualitative processing:** human involved in interpretation (e.g., transcription, annotations, and categorization)
- **G. Output from quantitative processing:** human may involve in defining the rules but not making judgements at the time of processing (e.g., error rate and outliers)

Prototypes
- **H. Software:** Executables and/or source code, excluding those in E.
- **I. Hardware:** (e.g., 3D designs, circuit diagrams)

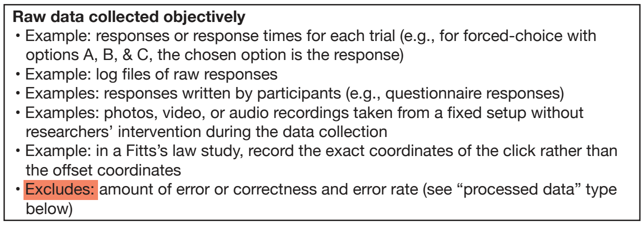
Figure 2. The taxonomy of research artifacts. See the full questionnaire in the preregistrations for details and examples.

To avoid this confusion, in the 2019 version, we shifted the focus away from the qualitative/quantitative dichotomy by (1) organizing the top-level hierarchy to focus on the maturity of data, and (2) using the terms "objective" and "subjective" to clarify whether recorded data was filtered by humans. One respondent commented on the 2019 survey that the raw data collection in quantitative research also involves researchers' subjective decisions on, e.g., sample sizes, operationalization of the variables, or measurement instruments. Therefore, in the final version of the taxonomy in Figure 2, we renamed the two types of raw data based on how much researchers' discretion are involved at the time of collection.

The second challenge is the fact that research artifacts in the same *format* may face different challenges in sharing and have different future use. For example, a software prototype of novel technology may face intellectual property concerns more severely than a piece of software code for statistical analysis. The former would be necessary for empirical *replication*, but only the latter (together with its input data) is necessary for *reproducing* the analysis.

Survey questions

The survey had four sections. The first section determines which types of research artifacts were generated. To avoid overwhelming participants, the survey progressively disclosed relevant types by asking two screening questions: whether the researchers (1) collected data from human participants and (2) produced products or prototypes. Based on these answers, a subset of nine artifact types was presented in a fixed sequence with concrete examples that are included or excluded in each type (Figure 3). We also provided an additional *others* type which was always available for providing a free-text description of materials outside the taxonomy. At the end of this section, the respondents indicated which types of materials are pertinent to the claims made in the paper.



Raw data collected objectively
- Example: responses or response times for each trial (e.g., for forced-choice with options A, B, & C, the chosen option is the response)
- Example: log files of raw responses
- Examples: responses written by participants (e.g., questionnaire responses)
- Examples: photos, video, or audio recordings taken from a fixed setup without researchers' intervention during the data collection
- Example: in a Fitts's law study, record the exact coordinates of the click rather than the offset coordinates
- Excludes: amount of error or correctness and error rate (see "processed data" type below)

Figure 3. In the survey questions, examples of inclusions and exclusions are provided for each artifact type.

The second section comprised a set of questions for each artifact type selected in the first section. These questions were the same in both years. To facilitate recall, a recap of the type (Figure 3) was shown at the top of each page. The survey asked if the artifacts were currently available for external researchers (as of 1 July 2018 or 1 May 2019 in the respective version). A **yes** answer led to location selection (e.g., ACM Digital Library, a public repository, or a university repository). The respondents could optionally provide the URLs. If the respondent indicated that the artifacts are available *upon request*, we asked about the conditions to share them (e.g., having proof of ethical or human subjects training). A **no** answer led to a question about reasons (e.g., contain sensitive data of study participants). For these questions about the locations of shared artifacts and reasons for not sharing them, the survey accepted multiple predefined answers as well as a free-text comment.

The third section asked for demographic information and provided a unique anonymized response ID. To ensure anonymity, the fourth section was collected on a separate survey URL. Here, participants could voluntarily associate the DOI of their paper with their responses via the response ID. The surveys were run on a LimeSurvey[30] server installed at the University of Zurich. The complete snapshot of the questions and the survey logic are in the preregistration.

Participant recruitment

Paper authors of CHI 2018 (655 invitations) and CHI 2019 (701 invitations[2]) were invited to participate in the survey. CHI neither imposes a specific order of author names nor has any declaration of the responsibilities of each author. However, based on our personal experience, the last author tends to appear on multiple papers. Therefore, to minimize confusion that may arise from associating each invitation to the paper, the first authors were invited. We also asked them to forward the invitation to an appropriate co-author if they believed that they were not responsible for the majority of generated artifacts.

In 2018, the email addresses were retrieved from a public conference planning system (MIT Confer). In 2019, they were extracted from the PDF files by a script (supplementary S8) and were manually checked by one of the co-authors who filled the missing entries. When emails were not in the paper, they were retrieved by searching the author's name on the Internet. For each year, around 20 emailed invitations bounced (e.g., because the author graduated). In those cases, we also searched for their current email address on the Internet. If we could not reach the first author, we sent an invitation to the next author using the same process. Each paper received a unique link to the survey with a randomly-generated code to ensure that each paper could contribute at most one response to the survey regardless of which co-authors answered. This mechanism allowed us to determine which invitation had filled the survey, but as the code was not associated with the response entry, each respondent remains anonymous.

Each survey was run for one month, and we sent one reminder after 10 days unless opted-out via a provided link. In 2018, the survey was run in July, and we received multiple vacation auto-responses. In these cases, we sent one reminder after the period indicated in the auto-responses and another reminder one week before the survey closed. To avoid vacations, we ran the 2019 survey in May, which yielded few vacation response problems beyond the first reminder. To avoid confounding the motivation to participate in 2019, we did not release the results of or mention the first survey in the invitation. In both years, we provided a link to the corresponding preregistration with the invitation.

As a participation incentive, we raffled prizes of Amazon gift cards (in the country of each recipient's choice), each valued equivalent to €50. We offered two prizes in 2018 and six prizes in 2019 hoping for a higher response rate. To retain anonymity of the response, the contact data for prize draw was collected in a separate survey URL that could be reached from the last page of the main survey.

Data analysis

We discarded incomplete responses (that did not reach the last page of the survey). We then conducted an exploratory analysis with three preregistered research questions:

RQ1: How many artifacts of each type are shared?
RQ2: If shared, where are they located?
RQ3: If not, what are the reasons?

We converted the proportion of the responses into a probability (e.g., 57 out of 168 responses results in a 33% probability) for three reasons: (1) Several questions of our survey accept multiple answers. (2) The total number of responses differs across artifact type. (3) In each paper, one author may contribute an answer to multiple artifact types. The probability values allow comparison across types and is calculated within each type: for RQ1, across all responses; for RQ2, across all responses that share artifacts; for RQ3, across all responses that do *not* share artifacts. We calculated 95% confidence intervals via the Clopper-Pearson exact method and computed inferential statistics on proportions using functions from the PropCIs

[2] Post-survey check: The initial release of the proceedings contains two duplicates. For these papers, the invitations were sent twice, but the duplicated invitations were unused. The invitations were not sent to the authors of Paper No. 10 and 968 due to errors in manual email address processing.

	n	Invited	Response rate	No artifact	Respondents' degree		
					BSc	MSc	PhD
'18	222	655	34%	21	29	103	90
'19	238	701	34%	18	38	104	96
∑	**460**	**1356**		**39**	**67**	**207**	**186**

Table 1. Response rate and academic degrees of respondents

package [46]. The analysis script and data are provided in supplementary material S3.

As for responses to free-text questions (e.g., other reasons for not sharing or comments about the survey), one of the co-authors deduplicated the responses that were entered for multiple artifact types by the same respondent, and then he conducted a thematic analysis [5] to iteratively group similar responses and come up with themes that summarize the results. The thematic analysis was conducted independently and in parallel with the quantitative analysis. After the thematic analysis, another co-author read through the categories and quotes and discuss with the first analyst until reaching an agreement. In the results below, we quotes representative or exceptional quotes based on the identified themes. A full list of the themes and exemplar quotes for each are provided in supplementary S5.

Preregistration and ethics approval

The data collection is approved by an IRB and is preregistered.[3] We deviated from the 2018 version of the preregistration by eliminating minor research questions for which the survey made inadequate assumptions about the dichotomy of qualitative vs. quantitative studies, as some data could be transformed between methodologies. For the 2019 survey, we also eliminated the research questions (e.g., the frequency of ACM Digital Library usage) that do not seem useful from the 2018 results, though the analysis for these 2018 research questions are in the supplementary S3.

RESULTS

Table 1 shows characteristics of the survey responses. Both years of the survey received a similar response rate of 34%. Among all responses, the majority (393 out of 460) were filled by an author with a masters or a higher academic degree.[4] 39 responses indicated that no research artifacts were generated, which could be from theoretical or opinion papers. These responses were excluded from percentage calculations. To avoid confusion, below we use the term "respondents" for those who answered our survey, while "participants" refers to people who participated in the studies described the papers published by the respondents. Also, some respondents may be the first author of multiple papers. Due to the anonymization of the survey, we could not indicate how many responded for multiple papers. However, each paper yielded at most one response. The median duration of the survey was 6.04 minutes (IQR = 5.10). We received 88 free-text comments in 2018 and 72 in 2019 (after deduplication 121 unique comments in total).

Overall sharing by artifact types

A low percentage of research artifacts are shared (Figure 4 left). The proportion for each type seems to be consistent in both years (for statistical test results, see supplementary S4). This consistency indicates that the revised classification scheme did not influence the results. Therefore and for the sake of readability, we combined the responses from both years in the analyses as shown in Figure 4 (right) and in the sections below.

Study materials (Figure 4 A)

Study materials are any code, questionnaires, stimuli, and anything else involved in collecting data as part of a study. Sharing study materials enables future replications. Unfortunately, only around 27–37% of the study materials are shared. The top five reasons against sharing are shown in Figure 6.1 (see supplementary S6 for complete results). 74 responses mentioned that they do not see the benefits of sharing. Given that artifacts of this type are *"... produced by researchers and presented to participants"* and *"**Excludes**: results or data collected"*, it was surprising that two out of the top-five reasons for not sharing concern participants' data and their permission. We checked for potential misunderstandings between study materials and data within these 114 responses. For each response, we compared their answers for the study materials with those for other artifact types. 70% of them responded differently. Thus, any possible confusion did not appear to be common.

One respondent indicated in the free-text field that they stored subject responses in the same spreadsheet as the questionnaire description, while other respondents seemed to conflate study materials with study data. One respondent used sensitive medical images for their stimuli, which legitimately could not be shared.

Raw and processed output data (Figure 4 B, C, F, G)

Data sharing is necessary for subsequent researchers to computationally reproduce analysis results. This survey separates raw data that researchers' discretion are involved during collection (e.g., field notes) and those that did not (e.g., response time logged by a software). Once collected, the raw data could be processed qualitatively or quantitatively. The sharing rate is below 25% for raw data and 40% for processed data.

Papers that share raw data are more transparent than papers that share only processed data; both are more transparent than papers that share neither. As shown in Figure 5, out of all responses that generate any type of data, only around 17–26% shared raw data, and around 25–23% shared only processed but not raw data. More than half shared neither.

Reasons for not sharing are distributed similarly in (1) selective raw data and (2) non-selective raw data, and (3) qualitative output (Figure 6.2). The top two reasons are the sensitivity of data and the lack of permission from study participants. Several respondents also expressed a fear that *"even 'anonymised'*

[3] IRB: The Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich (OEC IRB # 2018-031). Preregistrations: https://osf.io/3nvjp (2018), https://osf.io/gk9em (2019)

[4] In the U.S., doctoral students may only have a bachelor's degree until they receive a PhD.

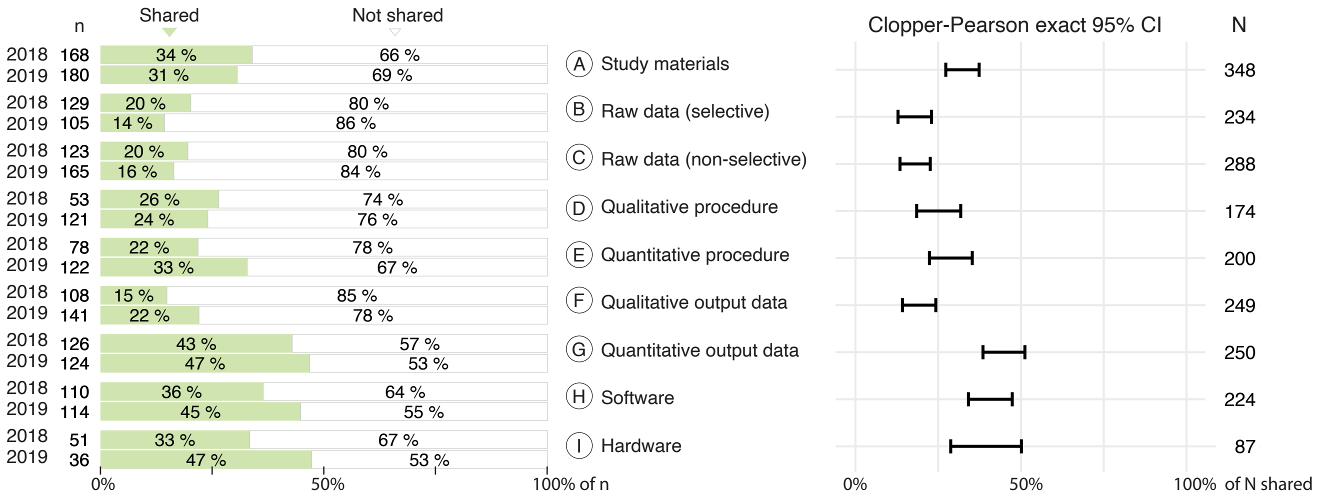


Figure 4. Percentage (left) and confidence interval (right) of research artifacts reported to be public.

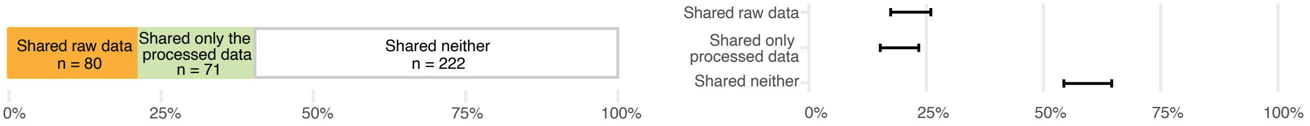


Figure 5. Proportion of shared data.

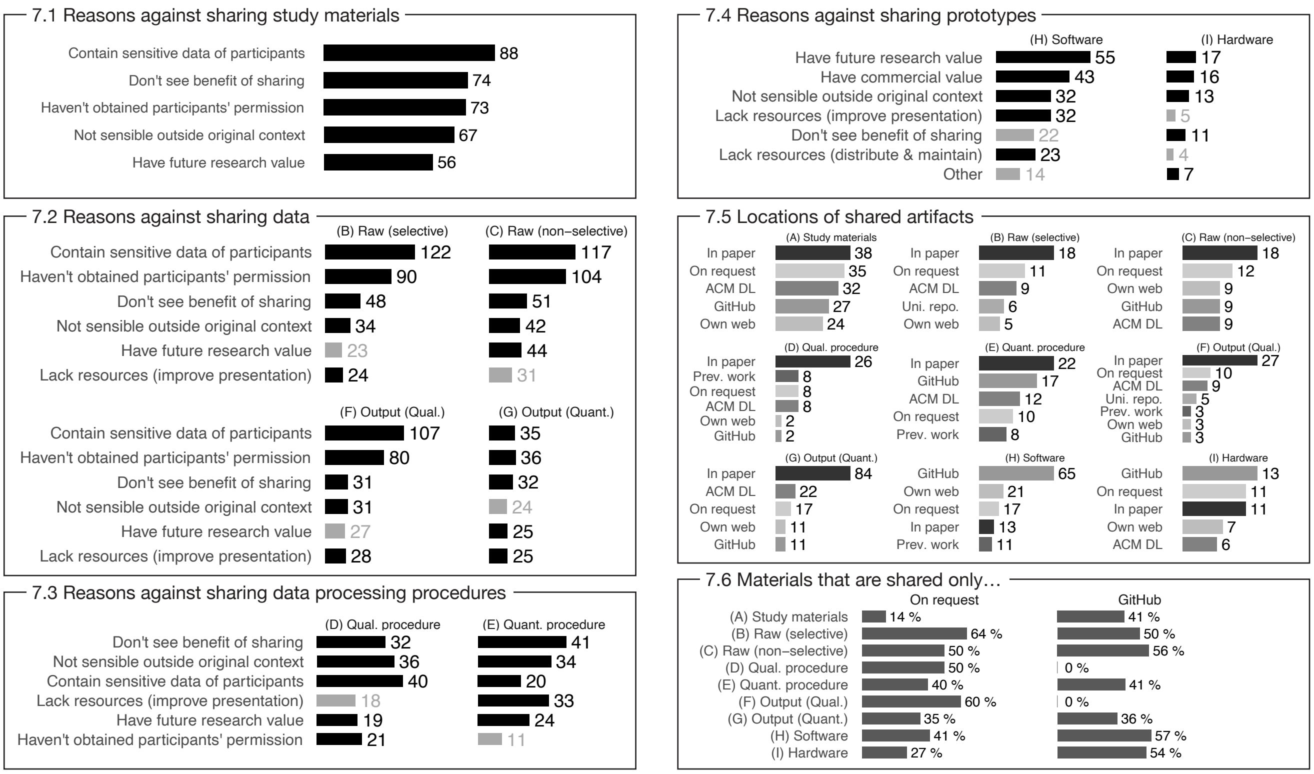


Figure 6. Top five reasons against sharing and the locations of shared artifacts. For full results, see supplementary S6.

(pseudonymised) data can be de-anonymised with enough effort; putting things online increases risk". Other respondents mentioned that sharing was prohibited by their IRB or ethics board. Some indicated that such restriction may be self-imposed: *"I don't think our ethics board makes it possible for others to access the data unless they are added to the ethics application (the addition needs to be approved)"*. These concerns seem to be dramatically reduced in the output from quantitative processing. Therefore, it is surprising that quantitative outputs are shared at only a slightly higher rate than qualitative ones. Similar to the study materials, many respondents (30–50 responses, depending on type) mentioned that they do not see the benefits of sharing.

Although the lack of resources to improve the presentation was ranked 5th or 6th, many free-text responses explained this concern in conjunction with the lack of incentives and unclear benefits to future research. The concern seems to be exacerbated in data that requires qualitative processing: *"Getting these materials ready to share will be another burden on [the junior co-authors]. Whilst sharing this information is clearly of value, how can [the junior co-authors] benefit from this process? Or, how can time be made for them to complete this extra task? "* The other response expressed a fear of being criticized *"it's stressful and difficult to get published as it is - giving everybody out there yet another avenue to tear my research apart sounds terrifying"*

A respondent indicated that the data in some domains (e.g., brain study) could be *"analyz[ed] with different metrics"* that the authors or the fields did not conceive. Lastly, two respondents did not share because their data is not in English.

Data processing procedures (Figure 4 D, E)

Procedures for analyzing data could be quantitative (e.g., data wrangling source code) or qualitative (e.g., codebook or transcription manual) and could be applied to transform any type of data to the other. Only around 18–35% of the data processing procedures are shared. Although the reasons against sharing are distributed differently between the two types (Figure 6.3), two of the reasons are consistently frequent: *"don't see benefits of making them public"* and that *"they do not make sense outside the original context"*. The first reason is consistent with the findings in study materials and data. It is surprising to see responses indicating that quantitative data processing procedures (i.e.*"source code involved in transforming raw data to final results"*) would *"not make sense outside the original context"*.

One response wished for an ability to scrutinize the analysis but also acknowledged the risks in sharing data: *"what I really would like to see is a platform for sharing user study data and the steps the authors took for analyzing those data. I understand that there are risks associated with sharing raw quantitative/qualitative data, but if done carefully this would be invaluable. Maybe we can encourage authors to create Jupyter notebooks allowing people to explore the data and submit those along with the paper?"*

Prototypes or products (Figure 4 H, I)

Prototypes could be created in the course of empirical research. Alternatively, without empirical studies, prototypes may be created as pure technology or design contributions. Sharing prototypes enables future research to build upon them. Although there are more responses in software than hardware, the shared proportions are similar, with the confidence intervals around 28–50%.

Top three reasons against sharing are consistent in hardware and software (Figure 6.4): future research value, commercial value, and that they do not make sense outside the original context. Notably, for the software prototypes, the lack of resources to (1) improve presentation and (2) to distribute and maintain them are two frequent reasons. One respondent elaborated: *"I see tremendous value in releasing the source code and compiled software associated with a paper. Yet doing so is non trivial- it requires extensive additional effort in cleaning and organizing the source, providing documentation, and supporting and maintaining the software itself."*

Locations of the public artifacts

Respondents who shared their research artifacts were also asked to specify where they are available (Figure 6.5). Except the software and hardware prototypes, most respondents stated that the artifacts are provided in the paper itself. This response is surprising due to the limited space and format of the paper. (There was a separate option for supplementary materials.)

Many responses indicate that the artifacts are available upon request to the authors (Figure 6.6, left). The literature suggests that this percentage is likely to decrease over time [49, 54]. Although in the past, sharing upon request was better than not sharing at all, at present online open- or protected-access repositories are more reliable and persistent. It is puzzling that 14% of the shared study materials and 50% of the shared quantitative data processing code are *only* shared upon request.

Across all categories, GitHub outranked science-oriented repositories that have a long-term availability plan, such as OSF. Among the responses that mention Github as one of the locations, around 36–57% of these responses shared *only* on Github (Figure 6.6, right). Some free-text responses even indicate a belief of GitHub being the right solution: *"I believe in sharing code and data. Most of my other recent papers share source-code and data on GitHub."*

Exploratory analysis of volunteered DOIs

In our survey, the respondents could voluntarily provide the DOI of their paper to be associated with their responses, allowing for a more detailed analysis. In the 2018 iteration of the survey, we received 18 volunteered DOIs (8% of overall responses). These were split among three of the co-authors to inspect the types of artifacts, whether they are shared, and whether they contain identifiable personal data. We consider that an artifact contains identifiable personal data if either: (1) removing participants' personal data reduces the ability to perform the analysis, or. (2) when de-anonymizing the data is possible even if substantial effort is required, participants' privacy could be harmed. Our inspections are then compared with the authors' responses in the survey.[5]

Regarding sharing, we found six DOIs where the authors responded that they shared the artifacts, but where we were unable to retrieve them due to no response to request, broken/nonexistent download link, or that the respondents only provided the data as an image in the paper.

With respect to identifiability, four DOIs indicated that their data are sensitive to participants' privacy. However, our inspection found no reason for participants' personal data to be in the dataset, or it could have been removed without losing in the ability to perform the analysis. Examples are anonymous timestamped logs of website events, screenshots from an experiment run on the researchers' computer, or logs of signals from custom input devices.

LIMITATIONS

Our study is firstly limited by the self-selection bias [22]: It is possible that the respondents could be more inclined to support research artifacts sharing than the non-respondents. This bias is exacerbated in the volunteered DOI by its small number of responses. The risk of self-selection bias could have been mitigated by (1) collecting information about attitude to artifact sharing (e.g., using the questions from [24, 17]), or by (2) analyzing non-participant data [27, Figure 1]. We decided against both because the survey is already long, and because non-respondents may have circumstantial reasons that prevented them from participating even if they wanted to (e.g. career transition or a parental leave). Nevertheless, we believe that the influence of self-selection bias is relatively low for two reasons: (1) Across all artifact types, the sharing percentage is less than half. (2) In free-text comments, several respondents were amusingly honest about their reasons for not sharing artifacts, e.g., *"No one has asked for them and I am lazy."*.

The second limitation is the granularity of response options. Some free-text responses suggest an option to mark artifacts as partially public or a different format of artifact (e.g., executable software instead of its source code). Our survey traded granularity for a reasonable length of the questionnaire. A future survey could be conducted with a sub-community, allowing them to narrow down to fewer artifact types and increase the granularity of the questions.

BARRIERS AND PROBLEMS IN SHARING

The desire to protect potentially personally-identifiable data: This concern is a major barrier in sharing research data. Nevertheless, the analysis of the volunteered DOIs suggests a lack of consensus of what considered as identifiable. Surprisingly this barrier also extended to study materials, which are produced by the researchers themselves. The results suggest that some study materials might have been stored together with the data, impeding both from being shared.

Lack of participants' permission: There are valid and respectable causes such as a prohibition from the IRB or studies involving identifiable personal information from participants. Research with vulnerable population may not be possible without the participants trusting that their data will not be shared [53]. In these cases authors could still be transparent about why data was not shared – see recommendation 11 below. Nevertheless, the analysis of volunteered DOIs indicates that even when the research data is not identifiable (such as trivially anonymizable speed and accuracy data), researchers might have neglected asking for appropriate permissions, whether from the IRB or the participants [33].

Lack of motivation, resources, and recognition of the benefits of sharing: The same concerns were also reported by an interview study with physicists [12] and a systematic literature review across various academic journals [11]. A likely cause is the lack of recognition of the benefits of sharing, which in turn could be caused by the rarity of HCI research work that leverages shared artifacts [1, 23]. But the reverse is also true, causing a chicken-and-egg problem.

Belief that the artifacts do not make sense outside the original context: We are surprised to see this reason mentioned relatively frequently in the study materials, quantitative data processing procedures, and software prototypes. At least sharing research materials enables the readers to better assess the quality and extent of the knowledge that the paper contributes, even when it authors claim it is unlikely to expect reproducibility or replication (e.g., research on a specific population of minorities [18])

Among the shared artifacts, the results indicated **misunderstandings about reliable methods to share**, e.g., choosing to share only upon request or only on GitHub. These barriers and problems suggest a need for knowledge on the benefits of sharing and on how to do so.

RECOMMENDATIONS

How can the field of HCI progress towards more frequent research artifact sharing or even universal open research practices? This challenge is unlikely to be solved by one single action or stakeholder [11], but instead many roles within the HCI community can take actions to improve transparency. Below, we list our recommendations about research artifact sharing in a broader context of open research practices. We believe that these recommendations are applicable across subdisciplines in HCI. Nevertheless, each sub-discipline could expand on them by providing more specific guidance for frequently-occurring artifacts and research contexts.

For authors

1. Be informed. Keep yourself up to date about data-related policies at various levels, especially about privileges they provide for scientific research. (Supplement S7 describes privileges provided in the EU's GDPR.)

2. Plan early. From the beginning of a project, determine what materials and data will be created or collected. Planning could be made concrete by writing a Data Management

[5] Since only one or very few papers about each specific topic exist, details about individual instances could be de-anonymized. Therefore, we deliberately describe these findings in a coarse granularity without sharing the supporting data. This analysis was not preregistered.

Plan (DMP).[6] Consider enhancing credibility with preregistration, an approach to clarify what facets of data collection and analysis planning occur before data collection.

Discuss with your regulatory or ethics board about what will be shared and how identifiable information will be protected. See Meyer (2018) for practical tips [33].

3. Preregister. A preregistration is timestamped evidence of when data collection and analysis decisions were made. Preregistration can help show how flexible a study's data collection was [55], i.e. where the study falls on the spectrum from exploratory to confirmatory [51]. Such benefits apply to both quantitative and qualitative research [21]. Templates for quantitative studies are available on https://osf.io and https://aspredicted.org, and a template for qualitative study is available on OSF. For more information about preregistration, see https://cos.io/prereg.

Note: Neither a DMP nor preregistration precludes changing plan. A preregistration is described as a plan, not a prison, as the goal is only to describe deviations from any initially planning rather than locking you into initial plans. Altered decisions from the preregistration need only be transparently stated and explained in the research paper. For example, the current paper includes deviations from the preregistrations. See section *Preregistration and ethics approval* on page .

4. Storing and sharing study materials. A description of methods in the paper is constrained by the page limit, so providing all artifacts can reduce the likelihood of any misinterpretations. Ensure that any code, stimuli, questionnaires, instructions, or other artifacts involved in data collection are available. Store them (or a copy of them) separate from the data. Digital study materials can even be included in the preregistration.

5. Collecting data. Collect as little identifying information as possible or collect them separately from data that is necessary for the main analysis [32]. For example, rather than collecting date of birth, only collect age in years. When obtaining participants' consent, honestly explain what will be shared and how identifiable information will be protected.

6. Storing data. Get into the habit of always storing raw data separately from study material and processed data. Data is rarely if ever self-explanatory, as interpreting other people's (or own old) datasets can be incredibly confusing and sometimes impossible. Therefore, create a *data dictionary*, a text file that briefly explains all column names, variables, values, file naming, and file structure. Always keep the data dictionary with the data. For tips on organizing data, see [57].

7. Sharing data. First, judging whether the data will be useful or fit the context should be deferred to the readers. Weigh the trade-off between (1) transparency, (2) privacy and data protection, (3) whether substantial effort is needed to prepare the data, and (4) rawness. Consider if dissociating the data completely from participants' information will still allow scrutiny and re-analysis. For example, quantitative performance (e.g., accuracy or task completion time) or data from Likert-type questions can be completely dissociated from detailed attributes about participants, (e.g., date of birth) if an interaction is not one of the research questions. Dissociated data should be shared in as raw a form as possible. The demographic data could still be obfuscated (e.g. replacing an age with an age range).

However, when transcripts or video of subjects are collected, subject privacy is an obvious concern. They could be anonymized (see [45] for a guide with examples.) Additionally, these three questions from the Data Protection Working Party of the European Parliament and the Council [38] could be useful for assessing the extent of anonymity:

- Is it still possible to single out an individual? (e.g., by name)
- Is it still possible to link records relating to an individual? (e.g., by a combining gender, age, postal code, and height)
- Is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes? (e.g., a birth year could be deduced from age)

These questions should be considered in light of additional information shared in the paper (e.g., institution name or the recruitment procedure). These concerns also apply to data obtained from publicly-available sources. Although it is publicly available, such data may be deleted later at the source. For further discussion on sharing public data, see [33, p. 142].

There are several additional methods to prevent privacy breaches such as only sharing summarized or aggregated data or putting the data in a protected access repository (for a list, see section 8 of [4]) that manages access to qualified researchers through a documented process. For quantitative data that requires more complex anonymization, consider creating a synthetic dataset that mimics the characteristics of the original dataset [42]. Lastly, data in languages other than English can be shared without needing a translation.

8. Sharing data processing procedures. Quantitative procedures (e.g., statistical analyses or simulations) and data processing code should be prepared such that a competent person in the field can reproduce any numerical results or figures. The code quality of each processing step could be determined proportionally to its role in the contributed knowledge. For example, each individual bar chart in Figure 6 could be reproduced by the analysis code, even though overall figure was manually assembled with graphics software.

9. Sharing software or hardware prototypes. While descriptions in a paper may seem clear to authors, sharing code or hardware schematics ensures that others can reliably implement the prototype on their own to retest it, reuse it, or develop the idea further. If intellectual properties or future research values are the concerns, researchers could consider sharing only relevant excerpts [29]. It is not expected that the code has to be perfectly organized [2]. Even code that is not runnable could be helpful in evaluating the research contribution [29].

[6] Funding agencies, including NSF, NIH, ERC, EPSRC, ANDS, SNSF, already instituted requirements for DMP in some funding schemes. (The link on each funders' name points to its DMP page, some contain templates.) See also an online tool: dmptool.org.

10. Selecting the location to share. The FAIR principles outline properties of a reliable location for sharing [56]:

Findable The location of the research artifacts should be easy to find. To enable findability, deposit the data in a searchable repository and place the URL in the paper.

Accessible The artifacts should not be locked behind any paywall. Moreover, to ensure persistent availability in the long-term, any repository that hosts these artifacts must have a clear plan or fund to remain available for decades.

Immutable It should not be possible to surreptitiously modify artifacts after being reviewed. They can be updated as long as the previous versions remain accessible.

Reusable While publishers may perform services related to a paper, the additional materials and any means of reading them cannot be owned or copyrighted by the publisher. Otherwise, future reuse could be prevented.

Example repositories that meet these criteria include OSF, Zenodo, or the repositories listed at re3data.org. **GitHub repositories are not immutable** (an owner can replace a repository with a different one, using the same name). **GitHub, as a company has no long-term persistence plan.** Therefore, in addition to using GitHub, we recommend depositing a snapshot to one of the mentioned repositories prior to submission. OSF and Zenodo have GitHub integration which can automatically retrieve a snapshot from GitHub. Zenodo even generates versioned DOIs for a specific or all version(s) of the code.

Asking readers to request research artifacts from the authors is not a reliable solution in the long run [54, 49]. If data privacy is a concern, consider depositing the artifacts to a protected access repository mentioned above.

It is also important to use repositories that are compatible with the submission process. For double-blind review, some repositories such as OSF allow repositories to be anonymized for peer review (see [20] for step-by-step instructions). Lastly, repositories such as Zenodo accept large datasets (50 GB).

11. If sharing is not possible, justify and describe the reason. If the artifacts cannot be shared (due to the privacy, ethical, or practical concerns), share as much as is reasonable and be transparent about why by providing a clear reason in the paper. Such explanations usually exist in the DMP (see recommendation 2.). Furthermore, describing the collected variables or features will allow other researchers to collaborate with the owner of the data to answer future research questions.

For reviewers

CHI reviewers are asked to judge if each paper provides *"a strong contribution to the field of HCI?"* [39]. The artifacts used to conduct the research or produced as part of it constitute evidence of the contribution and its validity. While reviewers cannot always be expected to carefully review every submitted artifact, they should should *at least* ensure that future readers can verify artifacts, or that the authors provided a clear justification against sharing. Reviewers should weigh the granularity and availability of artifacts against the ethical and data protection aspects [1], in relation to the artifacts' relevance to the contribution [15].

Many reviewers have formalized the practice by signing the Peer Reviewers' Openness (PRO) initiative [34]. The signatories pledge not to accept any submission unless it shares any data and material on an accessible repository or explicitly states why they cannot be shared.

For technical program chairs and steering committees

Besides nudging the authors (e.g., by adding a checkbox about data sharing on the PCS) and clarifying appropriate expectation to the reviewers (see above), they can also adopt policies that increase the transparency of research published at the conference. The Center for Open Science created badges to show the openness of multiple research components [37]. These badges are rewarded for Open Materials, Open Data, or Preregistration. They have been implemented by dozens of journals (see section 5 of [4]). Early evidence shows that a journal that implemented the open data badge saw substantial increase in data sharing compared with other journals in the same field [28]. See [4] for guidance and templates.

The Transparency and Openness Promotion (TOP) guidelines provide policy templates with multiple levels of strictness for three categories of research practices [36]. The different levels include (1) allowing authors to report if and where an artifact is available, (2) requiring open research artifacts or an explanation why they are not available, and (3) having a stage of review that checks research artifacts, such as rerunning analyses. The TOP guidelines are compatible with the open science badges and can use different levels for different artifacts. For example, a publication venue may use level 2 (requirement) for analysis code and level 1 (encouraged reporting for experiment code and stimuli. Over 1,000 publications venues have implemented TOP, and a study has shown a rapid increase in data sharing after the journal implemented the policy [19].

CONCLUSION

Sharing research artifacts is a prerequisite for replication, reproducibility, or at least a thorough assessment of research validity. To understand the landscape of research artifacts that are generated in HCI, how they are shared, and the reasons against sharing them, we conducted a survey with CHI 2018–19 paper authors. The results suggests four barriers: (1) concerns about participants' personally-identifiable data, (2) lack of participants' permission, (3) lack of motivation, resources, or recognition of the benefits of sharing, and (4) belief that the artifacts do not make sense outside the original context. We provided recommendations for authors who wish to consider sharing. We also discussed how reviewers can support sharing research artifacts and how chairs can advance open practices in their respective conferences or journals. We hope this paper encourages further conversations and improvements in research practices that will increase the credibility of HCI research in the future.

ACKNOWLEDGMENTS

We thank Kelly Caine, Lewis Chuang, Pierre Dragicevic, Alexander Eiselmayer, Eva Hornecker, Elaine Huang, Jofish Kaye, Wendy Mackay, David Mellor, and Antti Oulasvirta for their comments on an early version of the survey.

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)*. ACM, New York, NY, USA, Article 462, 14 pages. DOI: http://dx.doi.org/10.1145/3290605.3300692

[2] Nick Barnes. 2010. Publish your computer code: it is good enough. *Nature* 467, 7317 (2010), 753–753. DOI: http://dx.doi.org/10.1038/467753a

[3] C. Glenn Begley and John P.A. Ioannidis. 2015. Reproducibility in Science. *Circulation Research* 116, 1 (2015), 116–126. DOI: http://dx.doi.org/10.1161/CIRCRESAHA.114.303819

[4] Benjamin B Blohowiak, Johanna Cohoon, Lee de Wit, Eric Eich, Frank J Farach, Fred Hasselman, Alex O Holcombe, Macartan Humphreys, Melissa Lewis, Brian A Nosek, and et al. 2019. Badges to Acknowledge Open Practices (wiki). (Nov 2019). osf.io/tvyxz/wiki

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: http://dx.doi.org/10.1191/1478088706qp063oa

[6] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436. DOI: http://dx.doi.org/10.1126/science.aaf0918

[7] 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)*. ACM, New York, NY, USA, Article SIG04, 4 pages. DOI:http://dx.doi.org/10.1145/3170427.3185377

[8] 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)*. ACM, New York, NY, USA, Article 141, 12 pages. DOI:http://dx.doi.org/10.1145/3173574.3173715

[9] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). DOI: http://dx.doi.org/10.1126/science.aac4716

[10] 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)*. ACM, New York, NY, USA, Article alt02, 8 pages. DOI: http://dx.doi.org/10.1145/3170427.3188395

[11] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. 2015. What Drives Academic Data Sharing? *PLOS ONE* 10, 2 (02 2015), 1–25. DOI: http://dx.doi.org/10.1371/journal.pone.0118053

[12] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woźniak. 2019. Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 455, 14 pages. DOI: http://dx.doi.org/10.1145/3290605.3300685

[13] Christopher J Ferguson. 2015. "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist* 70, 6 (2015), 527.

[14] Casey Fiesler, Jed R. Brubaker, Andrea Forte, Shion Guha, Nora McDonald, and Michael Muller. 2019. Qualitative Methods for CSCW: Challenges and Opportunities. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. ACM, New York, NY, USA, 455–460. DOI: http://dx.doi.org/10.1145/3311957.3359428

[15] James Fogarty. 2017. Code and Contribution in Interactive Systems Research. In *Workshop HCITools: Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits at CHI 2017*.

[16] Annie Franco, Neil Malhotra, and Gabor Simonovits. 2015. Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results. *Political Analysis* 23, 2 (2015), 306–312. DOI: http://dx.doi.org/10.1093/pan/mpv006

[17] Carolin Haeussler. 2011. Information-sharing in academia and the industry: A comparative study. *Research Policy* 40, 1 (2011), 105 – 122. DOI: http://dx.doi.org/https://doi.org/10.1016/j.respol.2010.08.007 Special Section on Heterogeneity and University-Industry Relations.

[18] Anna Harding, Barbara Harper, Dave Stone, Catherine O'Neill, Patricia Berger, Stuart Harris, and Jamie Donatuto. 2012. Conducting Research with Tribal Communities: Sovereignty, Ethics, and Data-Sharing Issues. *Environmental Health Perspectives* 120, 1 (2012), 6–10. DOI:http://dx.doi.org/10.1289/ehp.1103904

[19] Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, Gustav Nilsonne, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J. Yoon, Michael Henry Tessler, Richie L. Lenne, Sara Altman, Bria Long, and Michael C. Frank. 2018. Data availability, reusability, and analytic reproducibility:

evaluating the impact of a mandatory open data policy at the journal <i>Cognition</i>. *Royal Society Open Science* 5, 8 (2018), 180448. DOI: http://dx.doi.org/10.1098/rsos.180448

[20] Steve Haroz. 2018. Open Practices in Visualization Research : Opinion Paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 46–52. DOI: http://dx.doi.org/10.1109/BELIV.2018.8634427

[21] Tamarinde L. Haven and Dr. Leonie Van Grootel. 2019. Preregistering qualitative research. *Accountability in Research* 26, 3 (2019), 229–244. DOI: http://dx.doi.org/10.1080/08989621.2019.1580147 PMID: 30741570.

[22] James J. Heckman. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1 (1979), 153–161. http://www.jstor.org/stable/1912352

[23] 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 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3523–3532. DOI: http://dx.doi.org/10.1145/2556288.2557004

[24] Bobby Lee Houtkoop, Chris Chambers, Malcolm Macleod, Dorothy V. M. Bishop, Thomas E. Nichols, and Eric-Jan Wagenmakers. 2018. Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018), 70–85. DOI: http://dx.doi.org/10.1177/2515245917751886

[25] 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)*. ACM, New York, NY, USA, 1081–1084. DOI: http://dx.doi.org/10.1145/2851581.2886442

[26] 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)*. ACM, New York, NY, USA, 534–541. DOI: http://dx.doi.org/10.1145/3027063.3027084

[27] Claire Keeble, Graham Richard Law, Stuart Barber, Paul D Baxter, and others. 2015. Choosing a method to reduce selection bias: a tool for researchers. *Open Journal of Epidemiology* 5, 3 (2015), 155–162.

[28] Mallory C. Kidwell, Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology* 14, 5 (05 2016), 1–15. DOI:http://dx.doi.org/10.1371/journal.pbio.1002456

[29] Randall J LeVeque. 2013. Top ten reasons to not share your code (and why you should anyway). *SIAM News* 46, 3 (2013). https://sinews.siam.org/Details-Page/top-ten-reasons-to-not-share-your-code-and-why-you-should-anyway

[30] LimeSurvey Project Team / Carsten Schmitz. 2012. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany. http://www.limesurvey.org

[31] Dennis McCafferty. 2010. Should Code Be Released? *Commun. ACM* 53, 10 (Oct. 2010), 16–17. DOI: http://dx.doi.org/10.1145/1831407.1831415

[32] Cormac McGrath and Gustav Nilsonne. 2018. Data sharing in qualitative research: opportunities and concerns. *MedEdPublish* 7, 4 (2018).

[33] Michelle N. Meyer. 2018. Practical Tips for Ethical Data Sharing. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018), 131–144. DOI: http://dx.doi.org/10.1177/2515245917747656

[34] Richard D. Morey, Christopher D. Chambers, Peter J. Etchells, Christine R. Harris, Rink Hoekstra, Daniël Lakens, Stephan Lewandowsky, Candice Coker Morey, Daniel P. Newman, Felix D. Schönbrodt, Wolf Vanpaemel, Eric-Jan Wagenmakers, and Rolf A. Zwaan. 2016. The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science* 3, 1 (2016), 150547. DOI:http://dx.doi.org/10.1098/rsos.150547

[35] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Chris Chambers, Gilbert Chin, Garret Christensen, and et al. 2016. Transparency and Openness Promotion (TOP) Guidelines. (Oct 2016). DOI: http://dx.doi.org/10.1126/science.aab2374

[36] 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 (2015), 1422–1425. DOI: http://dx.doi.org/10.1126/science.aab2374

[37] Center of Open Science. 2019. Open Science Badges. (2019). https://cos.io/our-services/open-science-badges/

[38] The Working Party on The Protection of Individuals with regard to the Processing of Personal Data. 2014. Opinion 05/2014 on Anonymisation Techniques. (2014). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/1999/wp26_en.pdf

[39] CHI 2020 organizers. 2010. CHI 2020 Guide to Reviewing Papers. (2010). https://chi2020.acm.org/guide-to-reviewing-papers/

[40] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. 2016. A statistical definition for reproducibility and replicability. *bioRxiv* (2016). DOI: http://dx.doi.org/10.1101/066803

[41] Hans E. Plesser. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics* 11 (2018), 76. DOI: http://dx.doi.org/10.3389/fninf.2017.00076

[42] Daniel Quintana. 2019. Synthetic datasets: A non-technical primer for the biobehavioral sciences. (Aug 2019). DOI: http://dx.doi.org/10.31234/osf.io/dmfb3

[43] Christian Remy, Oliver Bates, Jennifer Mankoff, and Adrian Friday. 2018. Evaluating HCI Research Beyond Usability. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article SIG13, 4 pages. DOI: http://dx.doi.org/10.1145/3170427.3185371

[44] Kristin Yvonne Rozier and Eric W. D. Rozier. 2014. Reproducibility, correctness, and buildability: The three principles for ethical public dissemination of computer science and engineering research. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*. 1–13. DOI: http://dx.doi.org/10.1109/ETHICS.2014.6893384

[45] Benjamin Saunders, Jenny Kitzinger, and Celia Kitzinger. 2015. Anonymising interview data: challenges and compromise in practice. *Qualitative Research* 15, 5 (2015), 616–632. DOI: http://dx.doi.org/10.1177/1468794114550439 PMID: 26457066.

[46] Ralph Scherer. 2018. *PropCIs: Various Confidence Interval Methods for Proportions*. https://CRAN.R-project.org/package=PropCIs R package version 0.3-0.

[47] Birgit Schmidt, Birgit Gemeinholzer, and Andrew Treloar. 2016. Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLOS ONE* 11, 1 (01 2016), 1–29. DOI: http://dx.doi.org/10.1371/journal.pone.0146695

[48] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE* 6, 6 (06 2011), 1–21. DOI: http://dx.doi.org/10.1371/journal.pone.0021101

[49] 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 (2019/09/07 2014), 94–97. DOI: http://dx.doi.org/10.1016/j.cub.2013.11.014

[50] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2018. Special Interest Group on Transparent Statistics Guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article SIG08, 4 pages. DOI:http://dx.doi.org/10.1145/3170427.3185374

[51] Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han LJ Van Der Maas. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). (2011).

[52] Howard Waitzkin. 1990. On Studying the Discourse of Medical Encounters: A Critique of Quantitative and Qualitative Methods and a Proposal for Reasonable Compromise. *Medical Care* 28, 6 (1990), 473–488. http://www.jstor.org/stable/3765672

[53] Jenny Waycott, Greg Wadley, Stefan Schutt, Arthur Stabolidis, and Reeva Lederman. 2015. The Challenge of Technology Research in Sensitive Settings: Case Studies in 'Ensitive HCI'. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (OzCHI '15)*. ACM, New York, NY, USA, 240–249. DOI: http://dx.doi.org/10.1145/2838739.2838773

[54] Jelte M. Wicherts, Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist* 61, 7 (2006), 726 – 728. http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2006-12925-016&site=ehost-live

[55] 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), 1832. DOI: http://dx.doi.org/10.3389/fpsyg.2016.01832

[56] 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 (15 March 2016). https://doi.org/10.1038/sdata.2016.18

[57] Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. Good enough practices in scientific computing. *PLOS Computational Biology* 13, 6 (06 2017), 1–20. DOI: http://dx.doi.org/10.1371/journal.pcbi.1005510

[58] 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*. ACM, 1185–1188.

[59] 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*. ACM, 33–36.

[60] 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*. ACM, 463–466.

[61] Max LL Wilson, Paul Resnick, David Coyle, and Ed H Chi. 2013. Replichi: the workshop. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 3159–3162.

[62] Tobias Wingen, Jana Berkessel, and Birte Englich. 2019. No Replication, no Trust? How Low Replicability Influences Trust in Psychology. (Feb 2019). DOI: http://dx.doi.org/10.31219/osf.io/4ukq5

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