# Diagnosing Bias in the Gender Representation of HCI Research Participants: How it Happens and Where We Are

by

Anna Maria Offenwanger

B.A., University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

**Master of Science**

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

**Diagnosing Bias in the Gender Representation of HCI Research Participants: How it Happens and Where We Are**

submitted by **Anna Maria Offenwanger** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

**Examining Committee:**

Dongwook Yoon, Computer Science, UBC
*Co-supervisor*

Julia Bullard, School of Information, UBC
*Co-supervisor*

Anne Condon, Computer Science, UBC
*Supervisory Committee Member*

# Abstract

In human-computer interaction (HCI) studies, bias in the gender representation of participants can jeopardize the generalizability of findings, perpetuate bias in data driven practices, and make new technologies dangerous for underrepresented groups. Key to progress towards inclusive and equitable gender practices is diagnosing the current status of bias and identifying where it comes from. In this mixed-methods study, we interviewed 13 HCI researchers to identify the potential bias factors, defined a systematic data collection procedure for meta-analysis of participant gender data, and created a participant gender dataset from 1147 CHI papers. Our analysis provided empirical evidence for the underrepresentation of women, the invisibility of non-binary participants, deteriorating representation of women in MTurk studies, and characteristics of research topics prone to bias. Based on these findings, we make concrete suggestions for promoting inclusive community culture and equitable research practices in HCI.

# Lay Summary

This research looks into the gender of who takes part in Human Computer Interaction (HCI) studies as research participants. Historically, more men have participated, which might make technology harder to use for people in other gender groups. We interview researchers about how they treat gender in research, and collect data about gender and participation from published research articles. In order collect enough data for statistical analysis, we developed a data schema to outline the gender data included in research papers and a tool to help us pull this data from papers quickly and accurately. We used this schema and tool to extract gender and recruitment data from 1147 research papers published at a prestigious venue in HCI research. Analysing this data, we found evidence that women are still underrepresented, non-binary people are nearly invisible among research participants, and identified characteristics of research topics that are prone to bias.

# Preface

This dissertation is an original intellectual product of the author, A. Offenwanger. The interviews reported in Chapter 3 were covered by UBC Ethics Certificate number H19-01245.

The analysis described in chapters 3 and 5 was done in close collaboration with supervisors D. Yoon and J. Bullard.

The software development on the MAGDA tool described in Chapter 4.3 was done partly by A. Offenwanger, and partly by A. Kobayashi.

The topic modeling work described in Appendix F was primarily done by A. Milligan, in collaboration with the rest of the research team.

# Table of Contents

# List of Tables

# List of Figures

# List of Supplementary Materials

1. Video Demonstration of MAGDA

   This submission includes a video demonstration of the Machine Assisted Gender Data Annotation (MAGDA) tool.

2. Dataset of Participant Gender Data in HCI

   The full dataset we generated as a core contribution is part of this submission, and is provided as set of CSV files, which are readable by most spreadsheet software, and README outlining the structure of the data.

   The dataset consists of the following files:

   (a) README.txt

   (b) auxiliaryDataset.csv

   (c) metadata.csv

   (d) primaryDataset.csv

   (e) primaryDatsetOversample.csv

# Glossary

**CS** Computer Science

**DER** Distance from Even Representation of Men and Women

**HCI** Human-Computer Interaction

**MAGDA** Machine Assisted Gender Data Annotation

# Acknowledgments

I wish to acknowledge and thank supervisors Dongwook Yoon and Julia Bullard, whose guidance made this project possible, and Anne Condon for her feedback as a thesis committee member. I would also like to thank Alan Milligan and Austin Kobayashi for their many hours on analysis and tool development, and Minsuk Chang for his input on the Machine Learning portion, as well as the interview participants for their crucial input.

I would also like to thank Sheldon Armstrong from the University of British Columbia Library and the Association for Computing Machinery for access to the CHI paper text.

Additionally, I would like to thank the following people for their timely feedback and review on various stages of the project: Ashish Chopra, Izabell Jansen, Soheil Kianzad, Yelim Kim, Ning Ma, Morgan Mo, Tamara Munzner, Angelika Offenwanger, Mohi Reza, Joice Tang, Mint Tanprasert, Brian Watson, and Marissa White.

Finally we would like to acknowledge the NSERC CGS-M, Discovery Grant, and Designing for People CREATE program, as well as the generous gift from Adobe that funded this research.

# Chapter 1

# Introduction

Bias in the gender distribution of research participants has long been of concern within the Human-Computer Interaction (HCI) community [38] because it can jeopardize the generalizability of research findings, perpetuate bias in data driven practices [45, 93], and make new technologies dangerous for underrepresented demographics [61, 116]. While gender bias has many dimensions [15, 39, 117], we focus on one aspect, *gender bias in research participation*, and define it as the incorrect assumption that knowledge produced is applicable to all genders when the data only justifies generalization to one gender group. We aim to investigate this gender bias within HCI research.

Previous data-driven surveys of HCI research participants showed that participant demographics are biased in favour of men [16, 36], but as these studies were focused on evaluation and sample sizes there remains a gap in our knowledge about where bias in research participant gender comes from. To address this, we did a quantitative analysis on how gender is treated across different variables using a gender dataset compiled from 40 years of CHI proceedings. This *data-driven* approach adds to the close reading approaches[1] taken in previous gender related HCI research studies [35, 74, 105] because distant empirical data can provide generalizable insights into the current state of research practice, including sources of bias, which can in turn suggest solutions for gender bias that go straight to the source.

[1] Close reading approaches involve fully reading and analysing individual works, as opposed to reviewing them via metadata or through visualization [69].

Our analysis focuses on women and non-binary individuals who have historically been problematically underrepresented [9, 16, 36, 61, 116], though it is important to remember that gender underrepresentation affects dominant groups (typically men) as well [33].

Our goal is to identify potential variables associated with gender bias, and investigate how those variables are related to bias. The term 'variables' describes the data we analyze better than 'factors', because 'factors' connotes a straightforward causal relationship which measurable values that correlate with gender bias do not necessarily have. Some variables can both cause and be a result of bias. For example, a "male default can be both a cause and a consequence of the gender data gap bias" [93, p. 17]. This can be seen in people reading "gender neutral" words as referencing men [29], which could be both an effect of women's exclusion from research [44, 64] and a cause for women to self-exclude [119, 120, 125]. There are only a few pointers to what variables might be possible, so our first aim is to determine what the potential variables might be.

The core methodological challenge in our data-driven approach towards analysing variables in relation to participant gender bias is how to systematically and robustly collect data from the large volume of published manuscripts. Gender terms used in HCI publications are flexible and nuanced, especially with the important consideration of incorporating non-binary gender [61, 110], and there is no available data schema for participant gender representation, let alone connected variables, to structure this data. Gender reporting varies widely in published research which leads to many mistakes and a lot of time needed for extraction. When papers contain multiple studies, for example, gender reporting may differ in both terminology and format (total numbers or participant tables) across a single paper (e.g., Hornof et al. [65]). Barkhuus and Rode [16] extracted data from 358 papers published in 1983 to 2006, and Caine [36] from 465 papers published in 2014 through manual analysis. The rate of gender reporting is somewhere between 50% [16] and 70% [36], so if we want gender data out of an average of even 10 papers per year over the 40 years of CHI, we will need to extract data from close to double the number of papers that previous studies did, and will therefore require systematic methods and tool supports to allow gender data to be collected efficiently.

In this paper, we posit and address the following research questions.

- RQ1: How are women and non-binary individuals (under)represented as participants in HCI research?
- RQ2: What are the variables that are associated with the gender demographics of HCI research participants?
- RQ3: How can we systematically collect gender data from published research for a data driven analysis?

We take a mixed methods approach that first qualitatively identifies variables that are potentially connected to gender bias and then triangulates the qualitative interview results with quantitative data extracted from published research (N = 1,147). In the qualitative phase we identified potential variables by interviewing 13 HCI researchers about the way they treated gender in their studies, building an understanding of issues and patterns underlying gender bias. The patterns informed the design of the data schema and tools to structure and extract a robust dataset of participant gender and connected variables, thereby tackling RQ3. We conclude with a statistical analysis of the dataset to answer RQ1 and RQ2 by exploring and triangulating patterns of gender bias.

Our study provides three main contributions.

- Empirical Contribution: We identified three key variables that can impact and/or indicate gender bias: recruitment, gender reporting, and research area. We report the trends of these variables and the current state of gender bias in HCI research.
- Methods Contribution: To extract gender data and connected variables, we iteratively developed a set of guidelines to support accurate and efficient data extraction, and structured a data schema for participant gender data coding, applicable to human subject based scientific studies generally[2].
- Dataset Contribution: We provide the resulting gender dataset of study participants in the CHI proceedings, extracted from 1147 papers published be-

[2]The data schema and guidelines were instantiated in a tool for extracting gender data from the papers. However, design of the tool is beyond the scope of this paper and we do not claim any contribution about the tool.

tween 1981 and 2020, which includes gender reporting and reporting language, participant counts, and data on where participants were recruited from.

# Chapter 2

# Related work

### Gender in HCI

Gender has long been a concern for HCI researchers for a variety of reasons, including the potential benefits that can be gained from inclusive research [38, 50, 71, 104], as well as supporting social justice [31, 73], and ensuring that the resulting HCI body of knowledge is equally usable by all persons. Plenty of studies have detected relationships between gender and differences in behaviour and in technology use [15, 17, 18, 89]. Different kinds of gender bias crop up in different areas of HCI, including hiring [87, 121], retention [39, 117], publishing [24, 126], and willingness and ability to engage in computing fields [54, 86, 120, 123]. Stereotypes of technology users being predominantly men pervade the field [29], and it is possible for HCI researchers to intervene in these patterns with intentionally designed technology [15, 55]. Lack of attention to gender can cause harm through gender reductionism, especially to people with gender non-normative identities [70, 75], and through unintentional exclusion [93, 122], which can lead to role stereotyping [26, 107] and new technologies being difficult to use for underrepresented groups [9, 34, 85, 88, 102].

### Gender bias in research participants

In order to firmly ground our analysis of gender bias, we explored alternatives definitions and converged on a definition of gender bias applicable to HCI research

participant demographics. We follow the Canadian Institutes of Health Research definition of gender as "the socially constructed roles, behaviours, expressions and identities of girls, women, boys, men, and gender diverse people", acknowledging both that "[g]ender identity is not confined to a binary (girl/woman, boy/man) nor is it static", and that gender is different from sex [6]. Gender Bias has been defined as "any set of attitudes and/or behaviors which favors one sex over the other" [23, p. 83], which we generalize to attitudes and behaviours which favor one gender group over others. The medical definition that gender bias is "a systematically erroneous gender dependent approach related to social construct, which incorrectly regards women and men as similar/different" [101, p. ii46] is foundational to our definition, in that gender bias in HCI research is also systematic (occurring across multiple studies, researches, and institutions), and erroneous (the data produced is incorrect).

Sampling and gender bias is generally of concern to researchers as it impacts the soundness of research methodology [14, 16, 58, 63]. The need to consider gender bias in subject sampling is acknowledged in other fields [43, 64, 95], especially medicine [19, 101], where it has been linked to poor health outcomes and lower quality of life [28, 99]. Biased gender data can easily affect many kinds of HCI research because of gender differences in, for example, posture [127], finger size [82, 106], cognitive performance [59], learning [67], body image [49], and social behaviors [11, 48, 91] including technology acceptance [119, 131]. "More men than women participated in user studies" [36], but technologies are considered to be equally usable by all genders, which has led to problems [9, 116]. In addition to the problems that arise, researchers may miss opportunities, as gender analysis can lead to more accurate/statistically significant findings [116] and increases the audience for potential devices [114].

## Potential variables connected to gender bias in HCI

There is little literature available on potential variables connected to gender bias, but it does provide some pointers. Behaviours connected to participant gender may lead participants to self-exclude [119, 120, 125]. Researchers can purposefully exclude a gender for reasons connected to the research [41, 44]. There is also a link

between author gender and attention to gender and sex analysis [90]. Since women are underrepresented as authors for Computer Science (CS) research publications [39, 80], this could lead to a lack of gender and sex analysis. Gender bias could also be field specific: "[t]he under-representation of females might be related to other gender/technology issues, and seems common in studies of GPS users" [81, p. 1678].

### Data collection

Extracting gender data from publications is difficult because gender is complex, reporting is inconsistent [105], and often absent altogether. While there are good reasons for data to be absent, such as a concern for participant privacy [77, p. 4], absent data contributes to data silences where problems problems are impossible to detect let alone solve because of a lack of information [93]. Data silences are important to identify, further complicating the task. "While participant demographic information was reported more frequently [than other contextual information], the type of information provided varied greatly" [p. 6] [105], which makes it difficult to do any kind of automated extraction. Mixed initiative systems have sought to improve human performance on document and text classification [40, 42, 78] and correction [83] using machine assistance, and might prove effective in this task, which is essentially text classification. In addition, the nuance of how researchers collect and report gender demographics is important [70]: for example, do "20 participants, (5 female)" and "5 of our 20 participants reported themselves as female" mean the same thing? Although there exists gender reporting guidelines for HCI studies [103], not all papers, especially the older ones, comply with the guidelines. To handle this methodological problem, we propose a data schema and gender collection guidelines, which we provide for the use of future researchers in this area. We also present a machine-assisted data collection tool, MAGDA as detailed in Chapter 4.3.

### Bias beyond gender in research participants

Gender is only one aspect of diversity, but is a good first step to studying research participant demographics due to its high ratio of reporting [105]. There are many

ways in which different dimensions of participant diversity can affect research. Such dimensions include ability [21, 27, 53, 98], gender modality [60], displacement [72], job stigma [113], homelessness [112], race and class, [128] and race and gender [32]. Previous research on "intersectionality, a framework that focuses on how various dimensions of identity (e.g., gender, race, and class) coalesce inseparably and relate to the conditions of one's surroundings" [105, p. 1], looked into how well intersectionality was reported within CHI. 85% of publications provided gender, but only a third provided data for socioeconomic class, and less than a third provided data for race [105]. Gender and other identity categories are inseparable, but it would be difficult to study them in concert due to lack of data, so we focus on gender as a starting point, with the hope that this will build towards comprehensive analyses in future.

# Chapter 3

# Identifying potential patterns of gender bias

We qualitatively investigated gendered practices in HCI studies through the perspective of researchers to identify potential patterns in the underlying culture and practices which could be relevant to gender bias in research participant demographics and could impact the participation of historically underrepresented gender groups. Based on our findings, we identified participant recruitment, research area, and time as potential variables.

## 3.1 Methods

We interviewed HCI researchers about their research practices around participants, focusing on gender, research design, and decisions involved in recruiting and reporting. Semi-structured interviews with researchers were most appropriate because researchers have an end to end perspective of the processes and decisions in research, and the semi-structure of the interview would allow us to collect comparative data from researchers in different areas, while still allowing for unexpected data to appear.

We interview 13 participants in total. Our inclusion criteria was that the researcher had to have recently published one or more full papers involving participants at HCI venues. We randomly sampled publications from CHI'19 [1] and

UbiComp'18 [3] as they are generally regarded as top-venues in HCI and for their scale and diversity. We then emailed the contact author or the last author. We emailed 82 authors and received 35 replies, of which 13 agreed to participate. We aimed to get a diverse sample of researchers across gender identities (6 women, 7 men, 1 non-binary), researcher role (5 principal investigators, 6 research assistants, 2 supervisors), research experience (5 less than 5 years, 8 more than 5 years), country (4 Canada, 3 Germany, 2 Japan, 2 US, 1 Sweden, 1 Australia), and department (10 CS, 1 Electrical Engineering, 1 Communications, 1 Sociology/Digital Technology).

We prepared two interview protocols, one for primary investigators, one for principal investigators, both of which were aimed at understanding research practices and decisions around research participants. We separated the two protocols because we expected the primary investigators to give us more insight into participant recruitment on the ground, whereas principle investigators would be more in touch with what motivates decisions about study design (full interview protocols can be viewed in Appendix A). We cut the interviews off at the one hour marker, occasionally cutting off the discussion about motivations for research participation.

To analyse the data, we used theoretical thematic analysis [30, p. 12]. Thematic analysis is flexible enough to allow for goals as to results of analysis, but the process of coding still allows for themes to emerge organically and be deeply connected to the data. The thesis author conducted interviews and coded the transcripts, with supervisors reading through six of them and discussing possible themes identified from the codes in multiple iterations. Analysis was performed in parallel with data collection. As each interview was transcribed, an initial coding pass was performed digitally. A second pass on the data collected the low levels codes into themes, then integrated those themes with existing themes. When a theme started to be well developed, it was transferred from digital to paper. This process facilitated reflection on the themes, and prompted physically rearranging the codes into different themes. When the arrangement was complete, it was presented to the rest of the research team for further discussion and analysis. When the resultant changes had been incorporated, the arrangement was presented to the research lab to elicit a second round of review. All the changes would be reflected in digital version, which served as a final check and reflection on the theme. This

process was followed for all major themes, and took approximately a week for each major theme. We used NVivo for digital analysis.

We achieved early data saturation at 13 participants since the objective was to establish an exploratory basis for the data driven investigation, rather than to theorize or conceptualize the patterns solely grounded on the qualitative data.

## 3.2 Potential patterns of gender bias

The results (R1-R4) from our qualitative interviews outline how bias can be produced by sources of recruitment, affects the strength of research claims, and impacts the feasibility of research.

*R1: Research feasibility is improved when a researcher can get participants easily, but easy to access participant networks can introduce bias into the participant populations of research studies.* The primary cause of gender bias in research comes from the bias in the participant pools that researchers recruit from. P1 put it very plainly, "If that place has a gender distribution that is even, then that's what will happen, but because it's all about that place." P10 usually winds up with more women in their study, because they draw from a participant pool supplied from the media communications program, where there is a skew towards women. P5 recruited haptics design experts, and the haptics design field "is originally from the mechanical engineering field, which is [...] very dominated by male [researchers]", their participant sample only has around 20% women. P11 recruited dancers, and only had a couple of men participate. P6 mentioned how personal networks can produce this effect, "most of the students and those people are male, but we [...] try to get some diversity. So that's kind of our target, but at the same time, in reality, it's sometimes very hard to get at those people" (P6).

Access to participants directly impacts the success of the research. Having lots of participants means the research can recover if something goes wrong. Both P10 and P5 had some participants not complete the study, but P10's university recruitment pool meant they had enough participants to simply drop the incomplete data and carry on. P5, on the other hand, struggled to find participants so the incomplete data had to be incorporated and was a challenge. Five researchers (P3, P6, P7, P8, P11) all had issues getting participants for a study and were forced to

make study design modifications. P6 called this a "very kind of last choice, for us," so it makes sense that researchers would gravitate towards practices that make it easy to get participants.

The ease of access to participants depends on the access to networks which can be used for recruitment. Specialized participants can be nearly impossible to get without some kind of network, P4 described the recruiting process for blind participants as "walking my feet off, like going to [an association for the blind], trying them, having them send it out, [...] calling people, like, hey do you know somebody?" P5 reached out to their professional network to get haptics designers, P6 contacted people at companies to recruit engineers, and P11 contacted their dance school for dancers. One network almost every researcher has access to is a university, but recruiting from university networks tends to result in recruiting students, like in P10s case. In some cases, researchers are even restricted to using people from their university. Both P6 and P11 created equipment that required participants to come to the lab, in P6's case, multiple times a day. This restricted the available participants to those who spent their day on campus.

The homogeneity of students make them less desirable as recruitment pools, but they are frequently used in spite of this. P1 described emailing the CS grad students as "typical", and P5 said that if "you're doing more general research, [you recruit] from your own department." If researchers "use the students in the same department, they pretty much have the same background [...] So it's too homogeneous" (P6), and researchers who "wanted diversity [...] did not want 23 grad students from down the hall" (P8). There are risks associated with homogenous demographics, as P9 put it, "if you want to make claims about all adults being able to do a particular task, [...] but you only have like, thirty-year-olds, right? Then suddenly you're age biased in a particular way that would misrepresent the performance characteristics of your widget". In addition to weakened claims, some claims might be missed altogether. P6 did a statistical analysis on gender, but "in the end [...] we are able only to recruit two or three people, the female participants. So we cannot claim any useful things."

Bias being introduced by participant source is a key finding for gender bias in HCI. Since researchers perceive the use of students to be a potential contributor to gender bias, and it is known that gender representations of student populations in

different disciplines, such as CS [79] and psychology [52], are biased, this is a suggestive avenue for investigation. Our findings also show other recruitment sources, such as haptics engineers or dancers, can also produce bias, so we will investigate a wide range of recruiting sources and methods in our quantitative phase.

*R2: Rigorous recruiting strategies can be hampered by resources, notably time, and lack of time can cause diversity criteria to be dropped.* Thorough recruiting often requires the one thing researchers are chronically short on: time. P11, a master's student, described a point in the research where it came down to "I need to graduate, so I need participants." Time pressure that is caused by master's students needing to complete their degrees is felt by all the researchers, P12 found themselves asking "what's a method where we can get participants without spending a lot of time on it, and with master's it's like you jump on a project and already it's like very quick." More thorough recruiting strategies can take more time than is available in a standard two year master's degree. P13 mentioned "interviewing over a period of about one and a half years". The sad reality for a lot of researchers is "student[s] have to graduate [...] so, I feel like, the realpolitik of the research often times pushes us to take shortcuts" (P7). P7 felt the pain of this when they found a correlation with gender in their results, but realised fully investigating it would have required "more money, more time, more student hours" (P7), so "it's a result for that study, but it's kind of, in some sense, limited" (P7). These pressures could prevent researchers from countering existing population biases in recruiting, which doubles the need to ask what areas of recruitment, including methods, are at risk of producing biased demographics.

*R3: Gender inclusion can be driven by previous literature in the research area.* There are two reasons for prior literature to influence researcher's inclusion of gender. First, sound academic practice requires researchers to respond to previous literature; and second, researchers are happy to borrow methods from previous studies. P8 described having gender related literature pointed out to them as "lucky", because "it's the kind of thing that I could have missed, and then run the study, [...] and then like, oh crap, I should have done that, and then didn't". P10 ported an application from another study to VR, and then replicated the analysis from the previous study. The previous study "found out, okay, female participants had better decision making than male participants, when we carry it over to virtual

reality, we do the same thing" (P10). The practice of citing related work can cause common recruitment methods and reporting practices to be passed around specific areas of research. This in addition to other factors, like restrictions on recruitment methods imposed by method apparatus (e.g. requiring participants to come to a lab, R1), or how much the research relates to bodily experience of users (e.g., haptics, wearables, virtual reality) can impact the way gender is treated in that field, making research area another promising variable in our investigation.

*R4: There is a fairly universal consciousness among* HCI *researchers that gender norms are changing, but researchers do not have standard ways to handle gender beyond binary categories.* Nearly all the researchers we interviewed mentioned some notion of non-binary gender (P1, P3, P4, P6, P7, P8, P9, P10, P11, P13), or described gender as a personal choice (P2, P5, P12), however, none of the researchers had a means of handling non-binary gender in research. Gender being fluid, a range, or a unique property of an individual was perceived to be incompatible with categorical gender. P8 said "a more fluid understanding of gender does make me question our binary categorization in papers", but common ways of handling gender rely on binary categories, for example, "[gender] balancing just means having as equal a number as possible, and here we're using binary gender" (P8).

Theoretically, the idea of gender balancing can be extended to include an equal statistical proportion of non-binary genders, but "[w]hat is the right number of categories? Is it just male female, non-binary? Is it some other set of things?" (P9). HCI researchers have started reporting "x female, y male, z, you know, preferred not to disclose, or non-binary, or whatever it is" (P8), but beyond binary gender researchers "don't really have a standard way to handle those things" (P6). Researchers draw back from incorporating non-binary gender because "if you want to get into all the variations [in your gender survey options], it becomes very long" (P5).

Recently, new guidelines have been provided both from the HCI community [103], and from the style guides [12, 92]. Caine [36] observed that gender representation appears to be changing over time, and developments like these might be responsible. We have compiled a list of the guidelines (Appendix B), though it is too early to expect much adoption in the HCI community. We found no cor-

relation between these guidelines and gender representation, however, since researchers change their treatment of gender based on previous work (R3), we can expect changes to trickle through the field as researchers change their practices. It will be worthwhile to know how gender representation and gender practices have changed over time, so we will investigate this in our quantitative phase.

# Chapter 4

# Establishing a gender dataset for analysing patterns of bias

Investigating the potential patterns of bias we identified in section 3.2 calls for creating a dataset of HCI research participant gender representation. This chapter presents our approach to gender data collection and analysis that answers the methodological question: "How can we systematically collect gender data from published research for a data driven analysis?" (RQ3) Through two rounds of manual, iterative data extraction and preliminary analysis we established guidelines for gender data collection in tandem with a data schema for structuring the dataset.

## 4.1 Data collection guidelines

The gender data collection guidelines governed the procedure for recording instances of gender reporting in HCI research. This was necessary to handle ambiguous cases that stemmed from complex concepts and nuanced languages regarding gender. The guidelines also shaped the structure of the resultant dataset as presented in section 4.2. We outline the final guidelines (G1-4) here.

*G1: Only count data entities which are explicitly reported in the paper.* Some studies often have their gendered practices partially reported or entirely unreported. For these papers, making assumptions can lead to the interpreter/annotator introducing their own biases into the data, which can replicate problems we are trying

to diagnose, such as misgendering [61] and stereotyping [29]. In order to generate claims that are strongly justifiable, we ground them only on data entities which are explicitly written in the paper, and minimize speculations about what the author did beyond what's reported.

*G2: Keep the data representation flexible enough to encompass unexpected and nuanced data, especially gender terms.* As an interdisciplinary field, HCI includes a wide range of research methods and reporting styles that the data collection needs to encompass. Gender language is also extensive and evolving, so we capture expressions used by the authors as they are, and classify those expressions post hoc to analyse the data. Recruitment reporting also has very little consistency, so it is difficult to know how much of it to collect (e.g. "students", or "students from our department with 20-20 vision"), so we collect all text which talks about characteristics of participants and classify this data post hoc.

*G3: Do not assume gender is binary.* Using binary gender excludes non-binary persons and over-simplifies the complexity of gender, so avoiding binary gender is considered best practice in HCI [103]. However, binary gender is baked into the common text reduction strategy of reporting only one gender, e.g. "20 participants (10 women)". This is meant to report that 10 women and 10 men participated, but to conclude this we must join the authors in assuming gender is binary. We resolve this issue by recording only what was reported, however, when the authors make an apparent binary assumption we also collect that as data.

*G4: Carefully read sections of the paper that are likely to contain data.* While we aim to collect a complete data sample, some compromises are necessary to make collecting sufficient high quality data feasible. We therefore assume bits of participant information will be in proximity to each other. While this is not the case 100% of the time, it is often the case and simplifies searching for the data by reducing the amount of text that must be carefully read.

## 4.2 Data schema for research participant gender

With these guidelines, we developed a data schema to record reported participant gender data. We opted for using gender categories (Fig. 4.1, 1.2 - 1.4), and captured the words used to describe the genders (Fig. 4.1, 1.2.1), which allows us to

| Schema Classification | Definition and Examples |
|---|---|
| Paper | Main container for each paper. |
| 1 Participant set | Container for each set of participants in the paper. |
| 1.1 Participant total count | Total count of participants in the set. E.g. 21 |
| 1.2 Participants Reported as ⚧ | Container for reporting non-binary, gender-fluid, etc. |
| 1.2.1 Text Indicator ⚧ | Text for the gender category. E.g. "non-binary", "gender-fluid" |
| 1.2.2 Number Classified As ⚧ | Number of participants associated with this classification. E.g. 3 |
| ... | |
| 1.3 Participants Reported as ♂ | Container for reporting men, male, boys, etc. |
| ... | |
| 1.4 Participants Reported as ♀ | Container for reporting women, female, girls, etc. |
| ... | |
| 1.6 Binary Assumption | Container for data indicating gender was reported with a binary assumption. E.g. "20 participants (10 women)." |
| ... | |
| 1.7 Participant Source | Indicator for where participants came from. E.g. "CS students" |
| ... | |
| ... | |

**Figure 4.1:** Sample of the data schema developed for participant gender data in research publications.

interrogate our own choice of classification and the nuances of how that data was reported. This schema encodes only data which is reported (G1), but also captures non-reporting; if a data category is not recorded, this means that this information was not reported. To keep our data representations flexible (G2), the data entities in our schema are mostly semi-structured text entries (Fig. 4.1, 1.2.1, 1.7). We aimed

to strictly avoid assuming binary gender (G3), but had to balance that against gender assumptions used by researchers. A reasonable trade-off was to add a data field to specify whether the author reported gender with a binary assumption (Fig. 4.1, 1.6). See Appendix C for the full dictionary.

## 4.3 Robust data collection with the Machine Assisted Gender Data Annotation (MAGDA) tool

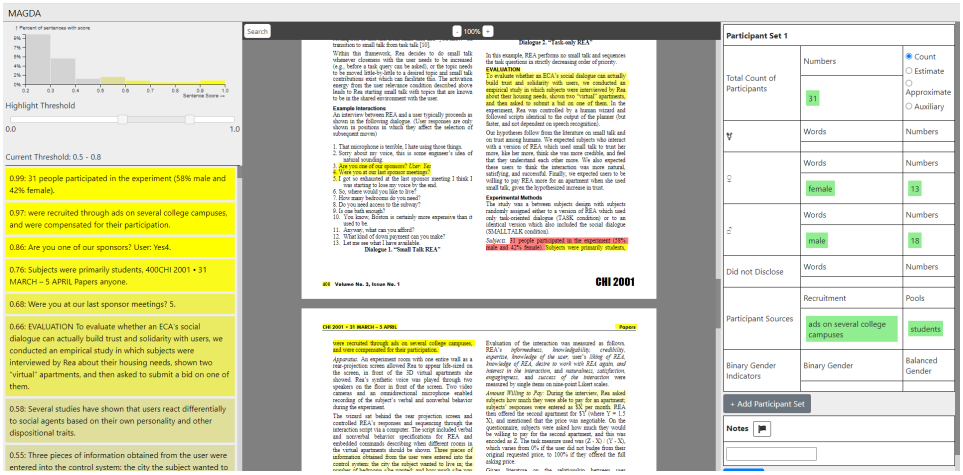


**Figure 4.2:** The MAGDA interface. An informal calculation of our average time to extract data from a paper showed a 4x extraction speed increase.

The manual collection of data (reading papers on a PDF reader and inputting entries into a spreadsheet) was slow, taking approximately 10 min per paper for our annotators, because finding data in PDFs has high cognitive load. The process was additionally error prone as manually entering the data often led to mistakes. For faster and more robust data collection, we developed and used the Machine Assisted Gender Data Annotation (MAGDA) tool, which is an instantiation of our guidelines and data schema (Fig. 4.2).

MAGDA employs an annotation metaphor to force all data to come from text explicitly reported in the paper (G1). Annotators select text in the paper, and input it into the appropriate data entry field (Fig. 4.2, right side), which links back to the body text. This method allows for flexible free-form text data entry (G2),

which handles unexpected data (e.g., "one housewife" [76]) and categories (e.g., "Gendered relationships").

To direct the annotator's attention to where gender data is likely to be found (G4), we developed a machine learning system to identify sentences that might contain data. We trained a logistic regression classifier on previously annotated sentences embedded by averaging the vectors of words in the sentence (vectors obtained from a gender neutral word embedding set [129]), which then predicted the probability of a new sentence containing data. As this model was not 100% accurate, we wished annotators to be aware that there is always uncertainty in the sentence classifications. Therefore, we applied variegating highlighting to likely sentences to visually represent the uncertainty inherent in the model (Fig. 4.2, central panel). To further improve the annotators' understanding of how the model was performing, we included a histogram of the classification probabilities. We also included controls for the highlighting thresholds to conceptually shift the task of evaluating the classification off the machine and onto the human intelligence.

To ensure data was not missed, thereby improving the robustness of the dataset, we also instructed annotators to search the immediate area around found gender data, as the model was trained only on sentences that contained gender data. Training the model strictly on the gender data was determined to improve classification accuracy, and did not impede the efficiency of the text search as gender and recruitment data tended to be in proximity to each other (We modeled the distance between gender and recruiting data as a Laplace distribution with a 95% confidence interval of 400 words away from gender data). This remaining need to search the text helps ensure that data is not missed due to false negatives in the machine classification. As a final check, we included a gender keyword search feature that allowed annotators to be reasonably sure they didn't miss any gender words (keyword dictionary used for this feature included in Appendix D).

The annotators found the MAGDA tool to be highly effective. Previously using spreadsheets and PDF readers, each paper took approximately 10 minutes to process, whereas the analytics data from our tool showed that this time dropped to 2.5 minutes with MAGDA. Using MAGDA helped us embrace the two core contributions during our data collection, the data schema and guidelines, and enabled the dataset contribution. See supplementary materials for a video demonstration of

MAGDA.

## 4.4 Dataset

We chose to gather data from ACM SIGCHI [1]. We selected papers via a random sample so that our data would generalize to the rest of CHI. Our goal was to collect data from at least 1000 papers.

We collected data from 1,147 papers (147 annotated by the author for initial data exploration and 1,000 annotated half-and-half by the author and a research assistant) published in all the different years of CHI (1981 to 2020). Our sample had more papers from the later years, reflecting the trend of increased publishing in CHI. To ensure the trends we observed did indeed apply to the earlier years, we extracted data from an additional 144 papers to bring the total number for each year to 16. This additional data showed no change in the found patterns reported in section 5, so we excluded it from analysis. For each paper, we extracted the fields outlined in the data schema (Fig. 4.1, Appendix C). See supplementary materials for the complete dataset.

To analyse types of gender reporting, we classified papers by what portion of the participants had gender reported (gender language coverage, Table 4.1). We had two categories of full gender coverage, one for papers in which all participants had gender reported, and one where the author reported participant number and only men or only women, leaving the rest to be assumed to be the other. For papers with multiple studies, we summed up participant totals and gender reporting to assign a single value for each paper (pilots were dropped following previous studies [36, p. 984]). We also classified papers by the gender words used in the paper (Gender language categorization, Table 4.2). For the small amount of data (less than 10%) which have multiple gender language reporting categories (i.e. males/females and non-binary used in the same paper) we applied the first classification from Table 4.2. Table 4.2 contains a comprehensive list of the gender words encountered in the papers we analysed. We collected recruitment data verbosely, and used affinity diagrams [94] to find trends in reported recruiting practices. From this we created a classification code book, and categorized the papers into the recruitment classifications (Appendix E).

**Table 4.1:** Gender language coverage

| Categories of Gender Reporting Coverage Types | Papers Classified | Criteria |
|---|---|---|
| Full coverage | 332 | Every participant has their gender reported or reported as unspecified |
| Assumed full coverage | 179 | In these papers, every participant has their gender reported if we assume that all participants not reported as male are female or vice versa. Includes only papers which have a binary assumption or report 'balanced' gender |
| Partial coverage | 73 | Gender is reported but not for every participant |
| No or insufficient coverage | 341 | No participant gender reported, or insufficient data reported to determine coverage (i.e. some gender is reported but we don't know how many participants there were in total). |
| No participants | 222 | Paper did not include research participants |

**Table 4.2:** Gender language categorization

| Category | Papers classified | Words included |
|---|---|---|
| Non-binary | 12 | gender queer, nonbinary, non-gender-identifying individual, gender-fluid, trans-men, other, trans, transgender |
| women/men | 52 | women, woman, men, man |
| females/males | 115 | females, males [noun] |
| female/male | 367 | female, male [adjective] |
| f/m | 20 | f, m |
| gendered relationship | 13 | mother[s], father[s], grandmother[s], grandfathers, daughter, son[s], girlfriend, boyfriend, sisters, husband, wife |
| boy/girl | 17 | boy[s], girl[s] |
| Balance | 7 | balanced, equally represented |

In order to compare the representation between men and women across widely varying study sizes, we developed a metric, Distance from Even Representation of Men and Women (DER):

$$DER = \frac{women - men}{women + men}$$

This metric was calculated for papers which have full, assumed full, or partial gender data coverage (584/1147 papers, Table 4.1). This metric is bounded between -1 and 1, and is directional, 0 being even representation, positive meaning more women participated, and negative meaning more men. In this formula we count the number of men and women actually reported, and those reported under a binary assumption. For example, if an author reports "22 participants (10 women)", *women* would be 10, and *men* would be 12, since this style of reporting clearly means us to assume that the remaining 12 participants were men. This is to accept that the paper under investigation had binary assumption rather than to accept the binary assumption per se. We opted to use this data in an analysis of the disparity of representation between women and men, but acknowledge that this is only one of many kinds of gender underrepresentation, and we believe that investigating this issue does not mean that we are giving in to binary assumptions made by the authors of the publications ourselves. DER is confined to comparing the representation of men and women and cannot be used to analyse more complex gender representation.

## 4.5 Analysing patterns of gender bias based on the dataset

We analyzed the data by examining the potential patterns found in the qualitative study using statistical testing and plotting. A Shapiro-Wilk test on DER showed that the data is not normal ($W = 0.99, p < .001$), so we use non-parametric tests such as the Mann-Whiteney U, Wilcoxon Signed-rank, and Kruskal-Wallis for statistical significance of trends. To investigate the relation between research area and gender bias, as was suggested by R3, we examined how research topics were related to both gender statistics and recruitment sources. To classify the papers by research area, we applied probabilistic topic modeling [25] to the majority of CHI papers from 1981 to 2020 (N = 7,456) to assign a selection of 25 topics to each

paper generated using the MALLET library [57]. We chose topic modeling as it captures the broad content of the publication, taking in a paper's full text, and has been previously used in meta-analysis [124]. Table F.2 provides a list of the topics and the number of the 1,147 papers in each topic. Details about this process are included in Appendix F. 10% of the extracted data was annotated by both of the coders for calculating inter-rater reliability. The Cohen's Kappa statistic was over .6 for all data categories, with one exception, which is outlined in Appendix G.

# Chapter 5

# Findings

The overarching pattern of data indicated that *gender bias of human subjects is a persistent and extensive problem in HCI studies*. Specific analysis revealed reasons for both optimism and concern.

## 5.1 Women are underrepresented and non-binary people are invisible

A chronological analysis of gender reporting data reveals a trend of stagnant representation of women, despite of increase in gender reporting. Also, there are few non-binary people reported as participants in HCI research.

*Non-binary individuals are invisible in aggregated small size studies and are still being "othered".* As can be seen in Fig. 5.1, the reported number of non-binary participants is practically invisible. Only 12 of 548 studies that report gender mention non-binary, and those studies tend to be large studies (median participant count of 161). The proportion of studies reporting non-binary participants is increasing (Fig. 5.2), but there is no apparent trend in the language use; even in 2020 papers are still 'othering'. Six of the 12 non-binary papers exclusively use 'other', two from 2020, which deviates from best practices [103, G-5]. Of the other six studies, two reported transgender participants, but did so in such a way as to indicate that the trans people were a separate category from men and women. For many trans people this is completely inaccurate, but as we strictly adhere to

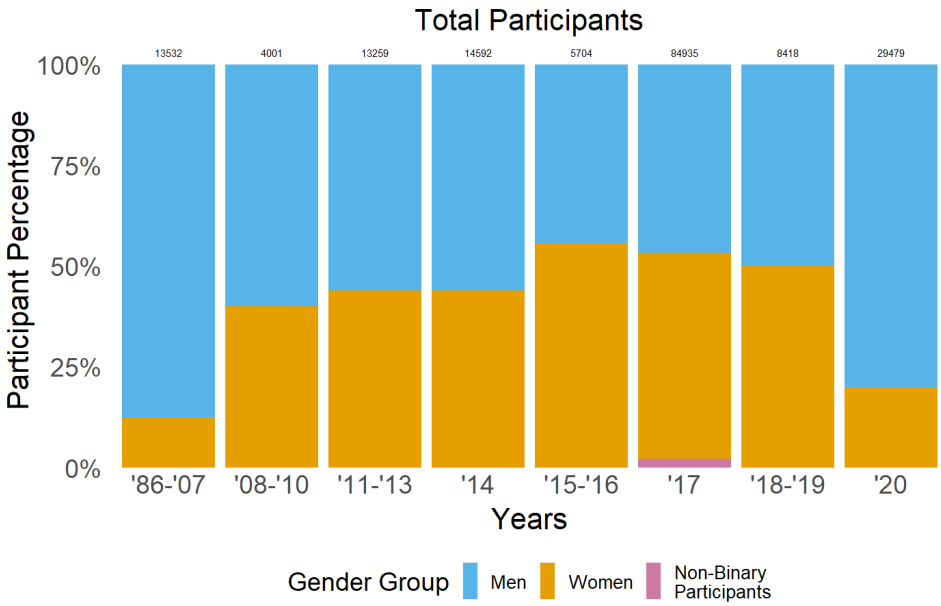


**Figure 5.1:** The total count of participants for the different year groups presented in proportion to each other. Non-binary participants are hardly visible.

what the authors report, we include them in our non-binary statistics. In another study, the author reported all their participants as trans men, while indicating that gender identity of some of the participants varied. While this indicates that some of the participants might have been non-binary, we do not know how many and do not include them in our non-binary statistics. Of studies that report gender (584 of 1,147), the percentage of participants reported as non-binary compared with the total number of participants is 0.9% (1858 of 210,575). This is largely due to a single 2017 study of 81,131 participants [66] (representing 39% of all participants in studies we analysed) where 2.2% of participants were reported as non-binary. If we analyse the central 95% of the studies (studies with 6-900 participants), we find that the percent of non-binary participants is 0.07% (22 of 32,838). The observed 0.07% reporting of non-binary participants will be useful for future comparison, as current demographic data does not reliably differentiate between binary and non-binary transgender populations, so there are no reliable population demographics

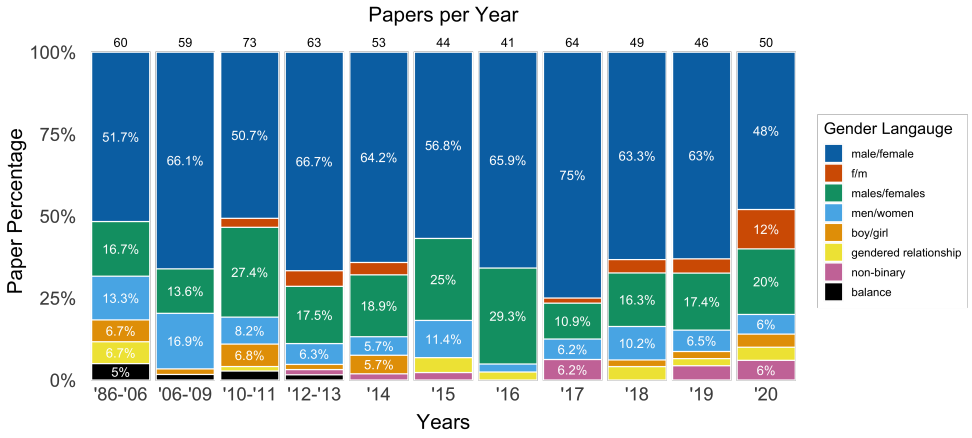


**Figure 5.2:** Percent of studies that fall into particular gender language categorization over the years of CHI. The use of "balance" language drops off as "non-binary" language appears, though the majority of reporting is still male(s)/female(s). A full breakdown of each categorization can been found in Table 4.2

for non-binary participants [7, 51, 68, 84].

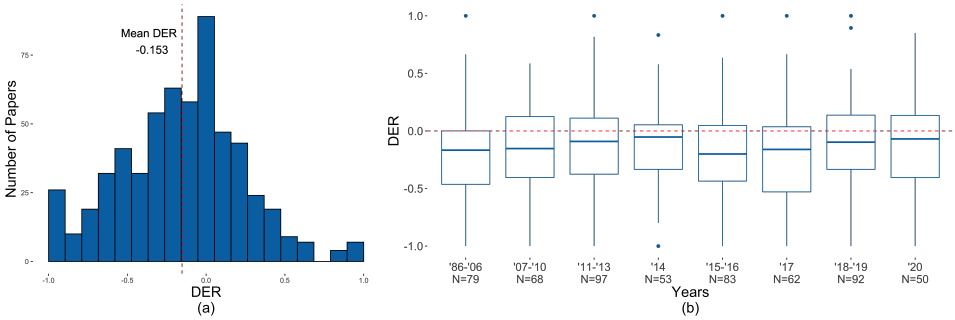


**Figure 5.3:** (a) The histogram of paper's DER, 584 papers have sufficient data to calculate DER. DER mean is -.153, meaning more studies were biased in favour of men. (b) The chronological trend of DER. More studies have more men participating, and this trend is not changing.

*The underrepresentation of women is persistent.* We compared the number of participants reported variously as women, female, etc, with the number reported as men, male, etc. The median number of women participating in studies was 10, and

men was 13. A Wilcoxon signed-rank test shows that there is a significant effect of these two groups ($W = 43032, Z = -7.60, p < .001, r = 0.22$), so there is still a bias in favour of men, as found in the previous studies [16, 36]. We then summed the DERs for all studies and divided by total number of studies with gender data, which gave us an average DER of -.15 (Fig. 5.3 (a)). To investigate the trend of women's representation over time, we grouped the various years of CHI to create a set of chronological buckets with as even a number of studies in each bucket as possible (Fig. 5.3 (b)). Unlike previous research [36, p. 989], our investigation shows that while the median DER of studies in each bucket fluctuates around the overall average of -.15, the proportion of women participating in research does not appear to be increasing. We applied linear regression to the data, and did not find a significant correlation between year and the DER of studies ($\beta = 0.002, t(582) = 0.41, p = 0.41$), providing no support for the participation of women increasing.

*Studies recruit all women intentionally and all men by coincidence.* Looking at the extremes of DER, we find a difference in the gender treatment of men and women which is linked to gender language. Twenty studies recruited all men, and seven studies recruited all women. Three of the all men studies and five of the all women studies used 'men/women' language. 15 of the remaining all men studies used 'male(s)' language, whereas none of the all women studies used 'female'. Looking at the studies that use "men/women" language from both extremes, we find gender is situated in context. Four of the five all women studies look at how specific groups of women interacted with technology. Two of the three all men studies were also highly contextual; one looked at domestic violence [20], the other at trans experiences of medical crowdfunding [56]. Looking at the all men studies that use "male/female" language, we find that the majority (13 of 15) had all men by coincidence, only two reported recruiting all men on purpose.

*For studies that report only one gender, but have both men and women, reporting men is correlated with bias in favour of men.* When reporting only one gender and leaving the other portion of the participants to be assumed (e.g. "We recruited 16 participants (8 female.)"), 5 studies report women for every one that reports men (150 to 30). The studies that report men have a lower DER; in other words have a bias in favour of men. A Mann-Whitney U test shows the difference between the DER of studies that report only the number of men (median -.33) and

the DER of studies that report both men and women (median -.13) to be significant ($U = 3801, p = 0.016, r = 0.11$). The DER of studies that report women does not significantly differ from the DER of studies that report both men and women (medians -0.14 and -0.13 respectively, Mann-Whitney U test does not show significance, $p = 0.24$).

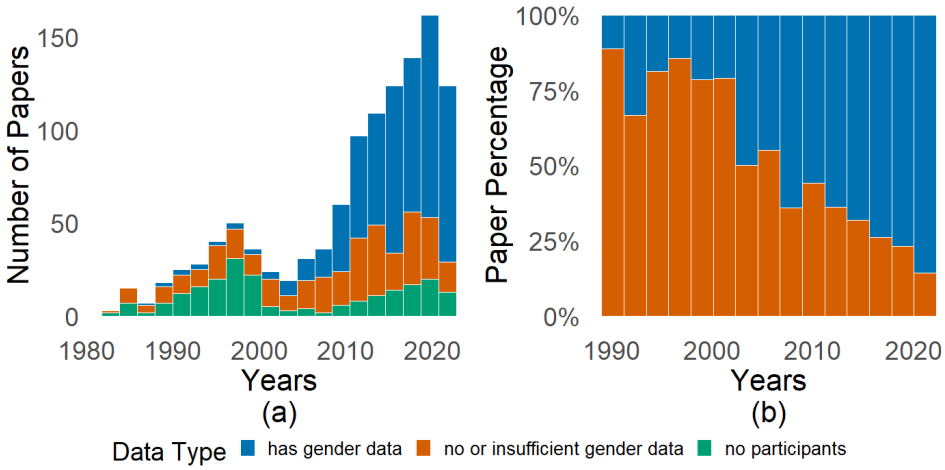


**Figure 5.4:** Graph of papers by year. (a) Number of papers sampled over the years of CHI, the number is higher after 2000. (b) The percent of papers with participants and gender reporting. Gender reporting is increasing. To smooth a jagged trend, 1980-1990 are grouped in first bar.

*Gender reporting is becoming prevalent.* To examine how gender reporting practices have evolved over time, we plot the percent of papers that have participants, report gender, and report numbers for those participants (Fig. 5.4 (b)). The proportion of studies reporting gender data has been steadily increasing. In 2020, fully 80% of papers reported some gender information. When we examine the gender language used in published papers (Fig. 5.2), reporting of participants with non-binary gender identities began to appear in the early 2010s. Reporting of 'gender was balanced' disappeared around the same time. This transition might be indicative of the cultural shift in the field's gender reporting practices where the notion of binary gender classification is being challenged (R4) [111].

## 5.2 The gender bias in the participant source skews DER

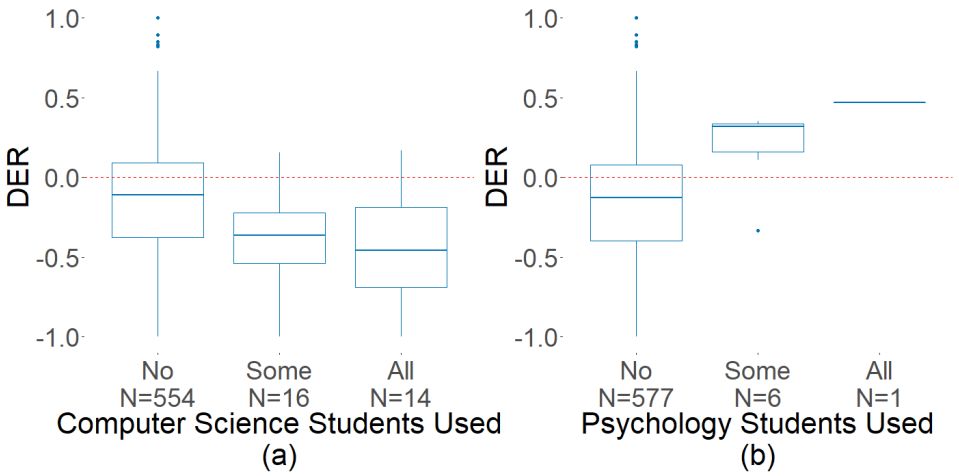


**Figure 5.5:** Participant recruitment. "All" means all participants belong to that classification, "Some" means only some of the participants belong to that classification. (a) Mean DER for studies that use CS students. The more CS students, the more men participate. (b) Mean DER for studies that use and psychology stduents. The more psychology students, the more women.

*Studies that recruit* CS *students are biased in favour of men and those that recruit psychology students in favour of women.* Figure 5.5 shows a clear trend in the mean DER in both cases. Studies that include at least some CS students (median DER -.41, $N = 30$) are more biased in favour of men than studies that do not report including CS students ($-.11, N = 554$). A Mann-Whitney U test showed this to be significant ($U = 4950.5, p < .001, r = .15$). Studies that include at least some psychology students (median DER .33, $N = 7$) are more biased in favour of women than studies that do not report including psychology students ($-.13, N = 577$). A Mann-Whitney U test showed this to be significant ($U = 3278, P = .002, r = .12$).

*Amazon's Mechanical Turk (MTurk) is becoming increasingly biased towards men.* MTurk is a crowdsourcing system commonly used to recruit research participants [37]. We applied linear regression to the data, and found year significantly predicted the DER of studies that used MTurk ($\beta = -0.03, t(32) = -3.01, p < .001$). The overall model of year predicted the DER of studies that use MTurk

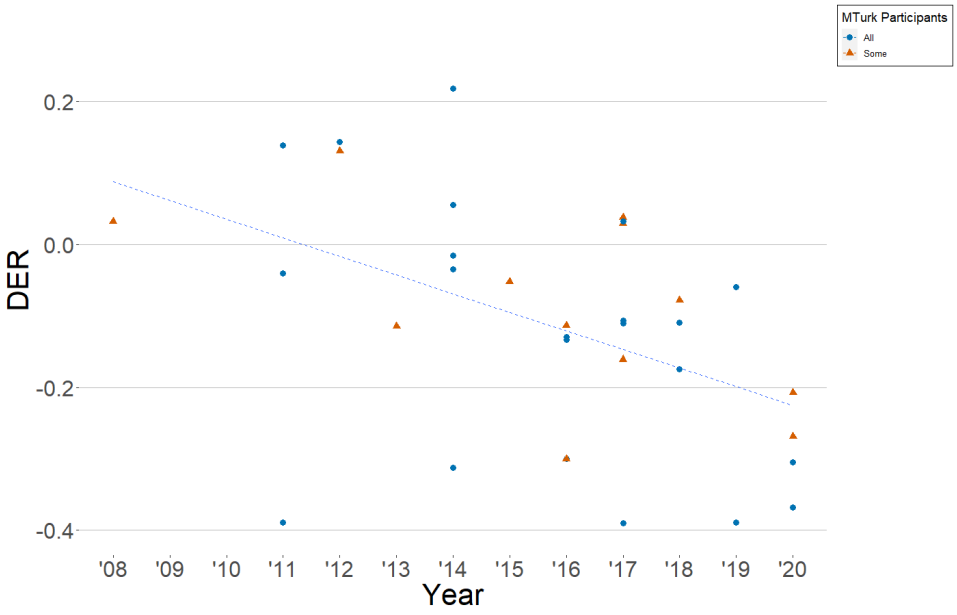


**Figure 5.6:** DER in studies that use Amazon's Mechanical Turk (MTurk), each point is a study. There is a statistical decrease in the number of women participating from MTurk over the last 10 years of CHI.

sufficiently ($adjustedR2 = .20, F(1,32) = 9.03, p = .0051$). Study DER decreases as year increases (Fig. 5.6). We discuss the implications of the deterioration in chapter 6.4.

*There are sources that appear to bias studies in favour of women.* Studies that include both children and adults show more women participating than studies that do not include children (median DER -.05, $N = 21$, and -.14, $N = 541$, respectively); a Mann-Whitney U test showed this to be significant ($U = 7669.5, p = .010, r = .10$). Studies where at least some of the participants were reported as having an illness or being in hospital showed more women participating than those that did not (median DER .235, $N = 14$, and -.13, $N = 570$, respectively); a Mann-Whitney U test showed this to be significant ($U = 6134.5, p < .001, r = .14$). Finally, studies that report using research pools, which are sets of people assembled through a mailing list or system specifically for the purpose of recruiting research participants, also show more women participants than those that did not (median DERs

of .288, $N = 11$, and -.136, $N = 573$, respectively); a Mann-Whitney U test showed this also to be significant ($U = 5018.5, p < .001, r = .14$). It is possible that many participant pools are hosted by psychology departments, but only 3 of 15 studies indicated the recruit pool was psychology, and only one of those provided gender data.

## 5.3 Gender bias patterns differ between studies in different topics

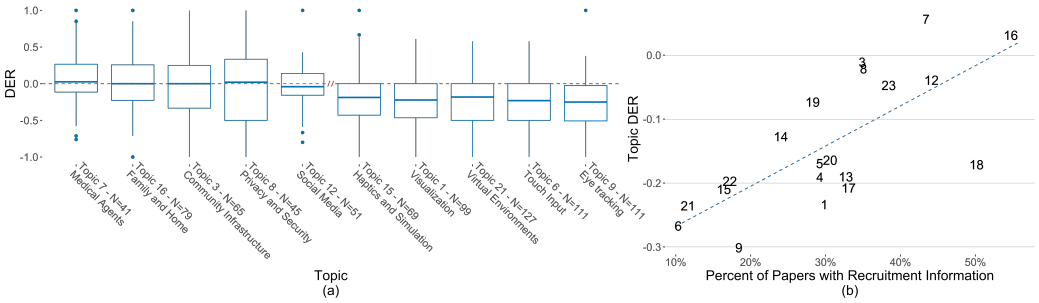


**Figure 5.7:** Comparison of topic to the DER. (a) The mean DER of papers classified in each topic, N = the number of papers classified. Showing the five topics with the highest mean DER, and the five with the lowest mean DER. Topics with high DER appear to be related to social interaction, topics with low DER to physical interaction. (b) The mean DER of paper classified as the given topic plotted against the percentage of those papers which have at least one recruitment classification. The higher the percent of studies in a topic that report participant recruitment information, the more women or fewer men participate in studies from that topic.

*Some research topics are more biased towards men than others.* In investigating the relation between topic and the gender demographics of men and women, a Kruskal-Wallis test to showed that there is a significant effect of topic on study DER ($\chi^2 = 118.56, p < .001$). Examining closer, we found that the studies in topics involved with physical interaction tended to have lower DER (the right side of Fig. 5.7 (a)), and topics that involve social interaction tended to have higher DER (the left side of Fig. 5.7 (a), Appendix Table F.2). For example, topics found on the

lower end of the scale include *Virtual Environments* (mean DER -.23), *Touch Input* (-.26), and *Eye Tracking* (-.29). Topics on the other end of the scale include *Family and Home* (mean DER .04), *Community Infrastructure* (-.01), and *Social Media* (-.04). Other topics on the higher end of the mean DER scale tended to recruit from sources that result in more women, discussed in the previous section. For example 11% of *Medical Agents* (mean DER .06) and 10% of *Health Metrics* studies recruit participants who are ill or patients.

*A topic's representation of men and women is correlated with the rate at which that topic reports participant recruitment.* The percent of papers in a topic that reported participant recruitment information correlates with the proportion of women who participated in studies from that topic (Fig. 5.7 (a)). We applied linear regression to the data, and found the percent of papers that reported recruitment information significantly predicted the mean DER for the topic ($\beta = 0.62, t(2352) = 48.52, p < .001$). The overall model predicted the mean DER fairly well ($adjusted R^2 = .50, F(1, 2355) = 2355, p < .001$).

# Chapter 6

# Discussion

## 6.1 The promises and perils of a data-driven approach to gender meta-analysis

Taken together, our results demonstrate three different ways that a data-driven approach can shed a new light on the problem of inequity in gender representation. First, there are specific, high-resolution patterns that a large set of data can reveal, such as the relationship between HCI research topics and bias. Second, putting the data points on a temporal scale reveals trends over time; we have identified the problems of deteriorating representations of MTurk participants and persistence of women's underrepresentation in the field. Third, examining group representation against the scale of HCI participant recruitment shows how most reporting renders non-binary participants invisible.

The two cornerstones of our data-driven approach are the data collection guidelines (Chapter 4.1) and the data schema (Chapter 4.2). The guidelines served as data collection principles that shape the characteristics of resultant data set. The schema implements the guidelines into a data structure that makes the data set functional for computational analysis. This gender data collection process is discipline agnostic, so our methodological approach is potentially applicable beyond HCI, to any field of science or engineering research that involves and reports groups of participants.

However, our study also found that data-driven analysis, as a distant reading ap-

proach, must be complemented by close, qualitative reading. The distant reading of the data can identify interesting patterns, but is not suitable for explaining those patterns, or for providing evidence as to causality in the patterns. For example, during our large scale analysis, we noted that studies using "men/women" language had a comparatively high number of studies with all women participants, but it was only when we individually read these papers that we discovered their trend of handling gender contextually. Similarly, while we previously suspected that recruiting CS students was in part responsible for gender bias in favour of men, it was not until we talked to the researchers themselves that we realised the need to embed flexibility in our data-schema to capture various potential recruitment sources. Despite our efforts to integrate nuanced analysis into our data-driven approach, we have found it impossible to perform this data-driven analysis with complex and non-binary gender, because our data is limited to what researchers report.

## 6.2 Beyond balancing, beyond binary

Our qualitative interviews with HCI researchers and background reading revealed two problematic aspects to gender balancing, which is a model of gender equity that has been used for many years in good faith [13, 22, 115]. The first problem is equating the proportional representation of the population with fairness. Gender statistics can be a good proxy for representation, but a simplistic representation like "balance" is not appropriate. For areas with known gender bias, such as software engineering, we expect and do see a majority of participants being men (the Programming Tools topic has a mean DER of -.20), but in this case the fact that most of the research participants are men can become a self-fulfilling prophecy. The low participation of women can allow problems which disproportionately affect women to go unnoticed [34], which has a detrimental effect on women's success in the field, leading to fewer women in the field to help test the new technology. In gender biased areas underrepresentative demographics are not only a consequence, but also a *cause* of gender bias and proportional representation can perpetuate the bias.

The second problem is that the language and concept of *balancing* inherently assumes gender binary and creates a false dichotomy. Close to a third of papers

which reported gender did so in a way which necessitated a binary assumption (179/587), while only 12 definitely did not assume binary gender. Even in these 12, the majority of them clearly considered binary gender to be the norm, as evidenced by half of them using "other" language to capture anything not binary.

Reflecting on our attempts to move away from binary gender in our own analysis reveals important future work that needs to be done in this area. Like previous work, we found "[t]he large majority of the work on gender with HCI implications has been from a binary perspective." [114, p. 3]. Fully 30% of papers reported gender with an explicit binary assumption (179/584, Table 4.1). Even in papers which did not have an explicit binary assumption, if they report only men and women, we have no way of knowing whether or not this data was collected with a binary assumption. Because of the unknown populations statistics of non-binary people, it is difficult to come up with an inclusive model of equitable representation. How researchers collect participant demographics feeds this lack of data; unless researchers ask for gender in a way that makes people comfortable disclosing, analysis of non-binary gender in the papers themselves will come up short. To break out of this mold, we require a model of population gender representation that encompasses gender diversity [7] and we need researchers to report gender in a way that is compatible with this model. With these, we could perform an inclusive analysis to track and improve gender representation in HCI research.

## 6.3 Where bias in gender representation comes from

Research expedients can drive researchers towards taking shortcuts (R2), which can result in gender considerations being dropped (R2) and in recruiting from easy sources, such as students (R1). Our data driven analysis has shown that students do introduce bias into participant samples, but also that other recruitment sources correlate with bias, and that bias is localized to certain research topics.

The CS student shortcut is easy, and easily overlooked. Students are a source of quick participants (R1), and CS students are especially so due to the number of HCI studies coming from CS researchers. As we have shown, the use of CS students biases studies in favour of men, and while psychology studies conversely bias studies in favour of women, there is a disproportionate amount of studies that use CS

students (49 CS to 15 psych). It is highly likely that the number of studies that use CS students is underreported. CS student use is so common that studies report when CS students are *not* used [8, 10]. Of the 1147 papers, only 49 reported some CS students participants, so it is highly likely that CS student use is underreported, and therefore a partially invisible source of bias.

The invisibility of men as men is another source of bias. Our analysis of the studies at the extreme ends of DER shows two problems. The first is that "man" as a gender, like whiteness or gender conformity, is invisible [46]. Very few studies focus on men as men. This could lead to addressing factors that affect men and technology only implicitly, never explicitly. The second problem is that studies that coincidentally include only men tend not to be questioned, resulting in results being only questionably applicable to women and non-binary participants. It is worth noting that only reporting the men in a study instead of only reporting the women, despite being correlated with a higher participation of men, helps to counter the invisibility of the "default male" [93], though it is better to report all genders and avoid assuming a gender binary. We cannot claim causality between language use and gender representation, but the correlation merits further investigation, which is left to future work.

We have highlighted several variables as potential causes for bias, and we have also shown that causes for bias change over time, but how these two factors interact is left to future work. Attitudes with respect to participants changed drastically over the three waves of HCI, with third wave HCI focusing much more on participants' lived experience [47], which could be the reason for the majority of studies shifting to including participants and reporting gender after 2000 (Figure 5.4). Methods also changed between the waves of HCI, shifting from quantitative to qualitative methods [47]. While our data has shown that research topic significantly impacts gender representation, it is possible that this could be partially explained by the methods preferred in different topics, and this could potentially have shifted with the waves of HCI. Research methods was beyond the scope of this analysis, so we recommend future research consider augmenting our provided dataset with research method to allow for an investigation of whether method impacts gender representation, and interacts with the other variables. Additionally, as the lack of gender reporting pre-2000 makes it difficult to do a data-driven analysis,

we recommend a further interview study, focusing on researchers who published before 2000 to investigate how gender identity was considered, and participant recruitment conducted.

## 6.4 Weak spots in HCI

Based on our analysis, we raise concerns about HCI research involving emerging technologies such as crowdsourcing, machine learning, wearables, virtual reality, and haptics, as they seem prone to bias in favor of men.

Topics with low DER should examine their recruiting practices for what sort of biases are introduced through recruitment populations. The shift of MTurk to uneven gender representation is concerning because of the lack of perception of this being the case, and because we can expect to see more researchers leaning on this source of participants due to COVID-19 restrictions. Previous studies on MTurk worker demographics had a roughly even representation of men and women [62, 100], but our data shows that there is a steady decrease in the proportion of women taking part in research via MTurk, which has been observed in the previous studies [100]. Previous research has extensively looked at how bias in machine learning algorithms can be traced back to biased datasets [32, 93], and as MTurk is often used to build datasets [96, 118], this could lead to the ML applications generated from these sets being biased, which can have serious negative consequences. For example, missing non-binary and trans people in facial datasets can lead to gender recognition systems misgendering, and if gender recognition systems are used to gate gender restricted areas, like washrooms, this can have a hugely negative impact on a vulnerable population [75].

Our analysis of different research topics shows that bias is localized to specific reasearch areas, and the causes of bias in those areas may differ. Studies in topics such as *Programming Tools* often require specialist participants (programmers), and are more likely to recruit from sources that have high proportions of men, such as CS students or software engineering companies. However, the same does not apply to topics such as *Eye Tracking*. Previous research has observed that "social acceptance of wearable devices differs between genders" [41], and in one study, "female users tended to report feeling uncomfortable with putting the device on

the chest" [41]. This could explain part of why studies that involve physical interaction with devices, like wearables, tend to have lower DER values. The source for bias could be research methods that make participants, women in particular, uncomfortable. Wearable devices are not the only emerging technologies which involve interacting with the bodies of participants. Virtual reality and haptics devices can occasionally involve full body interaction [108]. As the sources for bias likely differ between different research areas, so must the solutions.

## 6.5 Call to action

Comprehensive data is necessary for being able to conduct a gender inclusive analysis of representation, and also as a publication level reality check. Our data driven analysis moves towards supplying evidence for the extent of the problem, and we propose the following actions towards solving it. We propose that CHI collect participant demographic information, not just for gender, but for recruitment source, to track who is benefiting from the research published at CHI, and who is left out. The data schema and guidelines we provide, along with the recommendations for including gender collection in HCI [103, 109], can serve as a foundation for this effort. Subcommittees that handle topics most prone to bias can raise awareness about this issue by questioning participant sources in publications, and inquiring whether the participants actually represent the generalized population. Workshops targeted towards equitable recruiting, and standardizing methods for handling gender beyond binary can also go a long way to solving some of the issues we have encountered. Finally, in order to move beyond a binary model of gender equity, researchers need to collect and report gender data in a way that is compatible with nuanced models of gender, and so we recommend all researchers to check and incorporate the available guidelines [103, 109].

Researchers face competing constraints (R2) which can be obstacles to action for improving gender representation, and can push it to the side unless gender representation itself is considered to be a constraint. Reframing gender representation as a research constraint is a necessary attitude shift to achieve gender equity. The recommended practice of including gender in research design [116] can remove barriers to equitable recruiting by ensuring funds are allocated for accommoda-

tions (P9 mentioned childcare as one), and avoid invalid results due to gender differences in physiology, among other things. Several research bodies have started to include gender considerations in their application process [2, 4, 5], which is a positive sign that this attitude shift is taking place. We encourage researchers to make this inclusion standard practice.

# Chapter 7

# Conclusion and future work

We have provided empirical evidence for the underrepresentation of women and non-binary participants in HCI research (RQ1), and have shown that recruiting is a key factor in gender representation, with evidence that easy to access student networks might be introducing bias into participant populations, and studies using MTurk are becoming increasingly biased in favour of men (RQ2). In addition, we have shown that studies that are associated with physical interaction are more prone to an uneven representation of men and women, and research topics with more studies that report recruitment tend towards a more even representation of men and women (RQ2). Based on our analysis, we recommend a systematic survey of participant sources, especially for studies that involve any physical devices, as researchers should be aware that this could impact participant demographics.

Our gender data extraction guidelines and participant gender data schema, instantiated in the MAGDA tool, produced a structured and reliable set of gender data within HCI (RQ3), which we provide in the supplementary materials for the use of future researchers. The process of developing this schema and collecting this data set has necessitated resolving several conflicts and challenges in gender data coding, including the need to reconcile common binary reporting styles with non-binary conceptualizations of gender. Gender is complex and not static, and our data schema captures only gender as is currently reported in HCI research. As gender conceptualizations evolve, the data schema will need to be updated to adapt to the new ways in which gender will be reported.

Our study is an important first step in identifying sources of gender bias in research participation. We have identified trends in variables associated with gender bias, and the next step would be to do a factor analysis to determine whether there are causal relationships between these variables and gender bias. In addition to the factors suggested by the trends we have shown, the relation of topic to DER suggests that method might be a factor, which previous studies have examined [36]. Another variable to investigate in this future work is author gender. Author gender could impact participant recruiting in different ways, including impacting the willingness of participants to interact with researchers in different settings [130], and the effect of gender on a researchers connections [97], which can impact who reseachers have easy access to. We investigated the relation between institution and the representation of men and women, but we did not find any significant correlation. We suspect this might be because other factors, like recruiting and topic, are mixed up in institution, and a large scale factor analysis across different institutions could be illuminating. Identifying attitudes about sex and gender in publications could lead to a more nuanced understanding about how theories of gender affect representation, and might be possible through an analysis of how gender is analysed in published research, so this, along with method and author gender, is left to future work.

Finally, we would like to stress the iterative nature of this work. We expanded on previous quantitative literature [11, 31] with qualitative interviews, which informed our quantitative study. Our qualitative analysis did not have a large number of non-binary researchers, so there is opportunity here for further iteration. Factors which we identified in our quantitative analysis, such as topic and therefore possibly method as well, can affect the participation of men and women, so it is reasonable to think that they might affect non-binary people as well. Non-binary researchers would be able to provide insight into these issues, having knowledge of both research practices as well as the experiences of non-binary people, that could pin point factors that affect non-binary or transgender peoples' ability to participate in research. We therefore recommend a further qualitative investigation to prioritize non-binary perspectives.

# Bibliography


[1] Special interest group on computer-human interaction. URL https://sigchi.org/. Accessed: Sept 09, 2020. → pages 9, 21

[2] Fix the knowledge. URL https://www.fwf.ac.at/en/about-fwf/gender-issues/fix-the-knowledge. Accessed: Nov 26, 2020. → page 40

[3] Ubiquitous computing. URL https://ubicomp.org/ubicomp2020/. Accessed: Sept 09, 2020. → page 10

[4] Irish research council (2013) gender strategy and action plan 2013–20, 2013. URL http://research.ie/assets/uploads/2013/01/irish_research_council_gender_action_plan_2013_-2020.pdf. Accessed: Nov 26, 2020. → page 40

[5] Sex, gender and health research, Nov 2019. URL https://cihr-irsc.gc.ca/e/50833.html. Accessed: Nov 26, 2020. → page 40

[6] What is gender? what is sex?, Apr 2020. URL https://cihr-irsc.gc.ca/e/48642.html. Accessed: Aug 28, 2020. → page 6

[7] Sex at birth and gender: Technical report on changes for the 2021 census, July 2020. URL https://www12.statcan.gc.ca/census-recensement/2021/ref/98-20-0002/982000022020002-eng.cfm. Accessed: Sept 03, 2020. → pages 27, 36

[8] C. Ahlberg and B. Shneiderman. The alphaslider: a compact and rapid selector. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 365–371, 1994. → page 37

[9] M. Al Zayer, I. B. Adhanom, P. MacNeilage, and E. Folmer. The effect of field-of-view restriction on sex bias in VR sickness and spatial navigation

performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. → pages 2, 5, 6

[10] F. Alt, A. S. Shirazi, T. Kubitza, and A. Schmidt. Interaction techniques for creating and exchanging content with public displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1709–1718, 2013. → page 37

[11] A. Anzani and A. Prunas. Sexual fantasy of cisgender and nonbinary individuals: A quantitative study. *Journal of Sex & Marital Therapy*, pages 1–10, 2020. → page 6

[12] A. P. Association. *Publication manual of the American Psychological Association: the official guide to APA style*. American Psychological Association, Washington, DC, 7th edition, 2020. ISBN 9781433832161;143383216X;9781433832154;1433832151;. → page 14

[13] B. Bair and J. McGrath Cohoon. Special issue on gender-balancing computing education. *Journal on Educational Resources in Computing (JERIC)*, 4(1):1–es, 2004. → page 35

[14] S. Baltes and S. Diehl. Worse than spam: Issues in sampling software developers. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 1–6, 2016. → page 6

[15] S. Bardzell, S. Gross, J. Wain, A. Toombs, and J. Bardzell. The significant screwdriver: care, domestic masculinity, and interaction design. In *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction 25*, pages 371–377, 2011. → pages 1, 5

[16] L. Barkhuus and J. A. Rode. From mice to men-24 years of evaluation in chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 10, 2007. → pages 1, 2, 6, 28

[17] L. Beckwith, M. Burnett, S. Wiedenbeck, C. Cook, S. Sorte, and M. Hastings. Effectiveness of end-user debugging software features: Are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 869–878, 2005. → page 5

[18] L. Beckwith, M. Burnett, V. Grigoreanu, and S. Wiedenbeck. Gender HCI: What about the software? *Computer*, 39(11):97–101, 2006. → page 5

[19] A. K. Beery and I. Zucker. Sex bias in neuroscience and biomedical research. 35(3):565–572, 2011. → page 6

[20] R. Bellini, S. Forrest, N. Westmarland, and J. D. Smeddinck. Mechanisms of moral responsibility: Rethinking technologies for domestic violence prevention work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. → page 28

[21] C. L. Bennett and O. Keyes. What is the point of fairness? disability, ai and the complexity of justice. *arXiv preprint arXiv:1908.01024*, 2019. → page 8

[22] F. R. Bentley, N. Daskalova, and B. White. Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 1092–1099, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346566. doi:10.1145/3027063.3053335. URL https://doi.org/10.1145/3027063.3053335. → page 35

[23] N. E. Betz and L. F. Fitzgerald. *The career psychology of women.* Academic Press, 1987. → page 6

[24] V. S. Bhagat. Women authorship of scholarly publications in stemm: Authorship puzzle. *Feminist Research*, 2(2):66–76, 2018. → page 5

[25] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55 (4):77–84, 2012. → page 23

[26] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. → page 5

[27] D. Bora, H. Li, S. Salvi, and E. Brady. Actvirtual: Making public activism accessible. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 307–308, 2017. → page 8

[28] S. H. Bots, F. Groepenhoff, A. L. Eikendal, C. Tannenbaum, P. A. Rochon, V. Regitz-Zagrosek, V. M. Miller, D. Day, F. W. Asselbergs, and H. M. den Ruijter. Adverse drug reactions to guideline-recommended heart failure drugs in women: a systematic review of the literature. *JACC: Heart Failure*, 7(3):258–266, 2019. → page 6

[29] A. Bradley, C. MacArthur, M. Hancock, and S. Carpendale. Gendered or neutral? considering the language of HCI. In *Proceedings of the 41st graphics interface conference*, pages 163–170, 2015. → pages 2, 5, 17

[30] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006. → page 10

[31] S. Breslin and B. Wadhwa. Exploring nuanced gender perspectives within the HCI community. In *Proceedings of the india hci 2014 conference on human computer interaction*, pages 45–54, 2014. → page 5

[32] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. → pages 8, 38

[33] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan. Gendermag: A method for evaluating software's gender inclusiveness. *Interacting with Computers*, 28(6): 760–787, 2016. → page 2

[34] M. Burnett, R. Counts, R. Lawrence, and H. Hanson. Gender hcl and microsoft: Highlights from a longitudinal study. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 139–143. IEEE, 2017. → pages 5, 35

[35] S. Burtscher and K. Spiel. "but where would i even start?": Developing (gender) sensitivity in HCI research and practice. In *Proceedings of the Conference on Mensch Und Computer*, MuC '20, page 431–441, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375405. doi:10.1145/3404983.3405510. URL https://doi.org/10.1145/3404983.3405510. → page 1

[36] K. Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992, 2016. → pages 1, 2, 6, 14, 21, 28, 42

[37] K. Casler, L. Bickel, and E. Hackett. Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. *Computers in human behavior*, 29(6): 2156–2160, 2013. → page 30

[38] J. Cassell et al. Genderizing hci. *The Handbook of Human–Computer Interaction. Mahwah, NJ: Erlbaum*, pages 402–411, 2002. → pages 1, 5

[39] S. J. Ceci and W. M. Williams. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8):3157–3162, 2011. → pages 1, 5, 7

[40] J. Chan, J. C. Chang, T. Hope, D. Shahaf, and A. Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, 2018. → page 7

[41] L. Chan, C.-H. Hsieh, Y.-L. Chen, S. Yang, D.-Y. Huang, R.-H. Liang, and B.-Y. Chen. Cyclops: Wearable and single-piece full-body gesture input devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3001–3009, 2015. → pages 6, 38, 39

[42] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. → page 7

[43] R. Croson and U. Gneezy. Gender differences in preferences. 47(2): 448–474, 2009. → page 6

[44] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies. "yours is better!": Participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1321–1330, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi:10.1145/2207676.2208589. URL https://doi.org/10.1145/2207676.2208589. → pages 2, 6

[45] C. D'Ignazio and L. F. Klein. *Data feminism*. MIT Press, 2020. → page 1

[46] E. Drabinski. Representing normal: The problem of the unmarked in library organization systems. 2018. → page 37

[47] E. F. Duarte and M. C. C. Baranauskas. Revisiting the three hci waves: A preliminary discussion on philosophy of science and research paradigms. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, IHC '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450352352. doi:10.1145/3033701.3033740. URL https://doi.org/10.1145/3033701.3033740. → page 37

[48] A. H. Eagly. *Sex differences in social behavior: A social-role interpretation*. Psychology Press, 2013. → page 6

[49] A. Feingold and R. Mazzella. Gender differences in body image are increasing. *Psychological Science*, 9(3):190–195, 1998. → page 6

[50] C. Fiesler, S. Morrison, and A. S. Bruckman. An archive of their own: a case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2574–2585, 2016. → page 5

[51] A. Flores, J. Herman, G. Gates, and T. Brown. How many adults identify as transgender in the united states? los angeles, ca: The williams institute, ucla school of law, 2016. → page 27

[52] G. Fowler, C. Cope, D. Michalski, P. Christidis, L. Lin, and J. Conroy. Women outnumber men in psychology graduate programs. *Monitor on Psychology*, 49(11), 2018. → page 13

[53] C. Frauenberger, J. Makhaeva, and K. Spiel. Designing smart objects with autistic children: Four design exposès. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 130–139, 2016. → page 8

[54] V. Galpin. Women in computing around the world. *ACM SIGCSE Bulletin*, 34(2):94–100, 2002. → page 5

[55] T. Gill and J. Lei. Counter-stereotypical products: Barriers to their adoption and strategies to overcome them. *Psychology & Marketing*, 35(7): 493–510, 2018. → page 5

[56] A. Gonzales and N. Fritz. Prioritizing flexibility and intangibles: Medical crowdfunding for stigmatized individuals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2371–2375, 2017. → page 28

[57] S. Graham, S. Weingart, and I. Milligan. Getting started with topic modeling and mallet. Technical report, The Editorial Board of the Programming Historian, 2012. → page 24

[58] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 111–120, 2008. → page 6

[59] G. Grön, A. P. Wunderlich, M. Spitzer, R. Tomczak, and M. W. Riepe. Brain activation during human navigation: gender-different neural networks as substrate of performance. *Nature neuroscience*, 3(4):404, 2000. → page 6

[60] O. L. Haimson and A. L. Hoffmann. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday*, 2016. → page 8

[61] F. Hamidi, M. K. Scheuerman, and S. M. Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018. → pages 1, 2, 17

[62] K. Hara, A. Adams, K. Milland, S. Savage, B. V. Hanrahan, J. P. Bigham, and C. Callison-Burch. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019. → page 38

[63] C. G. Hill, M. Haag, A. Oleson, C. Mendez, N. Marsden, A. Sarma, and M. Burnett. Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 6658–6671, 2017. → page 6

[64] A. Holdcroft. Gender bias in research: how does it affect evidence based medicine?, 2007. → pages 2, 6

[65] A. J. Hornof and A. Cavender. Eyedraw: enabling children with severe motor impairments to draw with their eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–170, 2005. → page 2

[66] B. Huber, K. Reinecke, and K. Z. Gajos. The effect of performance feedback on social media sharing at volunteer-based online experiment platforms. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1882–1886, 2017. → page 26

[67] J. S. Hyde, E. Fennema, and S. J. Lamon. Gender differences in mathematics performance: A meta-analysis. 107(2):139–155, 1990. ISSN 1939-1455(ELECTRONIC),00332909(PRINT). doi:10.1037/0033-2909.107.2.139. → page 6

[68] S. James, J. Herman, S. Rankin, M. Keisling, L. Mottet, and M. Anafi. The report of the 2015 us transgender survey. 2016. → page 27

[69] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARs)*, pages 83–103, 2015. → page 1

[70] S. Jaroszewski, D. Lottridge, O. L. Haimson, and K. Quehl. "genderfluid" or "attack helicopter" responsible HCI research practice with non-binary gender variation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2018. → pages 5, 7

[71] K. A. Jehn, G. B. Northcraft, and M. A. Neale. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly*, 44(4):741–763, 1999. → page 5

[72] R. B. Jensen, L. Coles-Kemp, and R. Talhouk. When the civic turn turns digital: Designing safe and secure refugee resettlement. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. → page 8

[73] D. G. Johnson and K. W. Miller. Is diversity in computing a moral matter? *ACM SIGCSE Bulletin*, 34(2):9–10, 2002. → page 5

[74] G. Kannabiran, J. Bardzell, and S. Bardzell. How HCI talks about sexuality: discursive strategies, blind spots, and opportunities for future research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 695–704, 2011. → page 1

[75] O. Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018. → pages 5, 38

[76] D.-j. Kim and Y.-k. Lim. Co-performing agent: Design for building user-agent partnership in learning and adaptive services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. → page 20

[77] S. Kozubaev, F. Rochaix, C. DiSalvo, and C. A. Le Dantec. Spaces and traces: Implications of smart technology in public housing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. → page 7

[78] T. Kulesza, D. Charles, R. Caruana, S. A. Amershi, and D. A. Fisher. Structured labeling to facilitate concept evolution in machine learning, June 11 2019. US Patent 10,318,572. → page 7

[79] K. J. Lehman, L. J. Sax, and H. B. Zimmerman. Women planning to major in computer science: Who are they and what makes them unique? *Computer Science Education*, 26(4):277–298, 2016. → page 13

[80] M. J. Lerchenmueller, O. Sorenson, and A. B. Jena. Gender differences in how scientists present the importance of their research: observational study. *bmj*, 367, 2019. → page 7

[81] G. Leshed, T. Velden, O. Rieger, B. Kot, and P. Sengers. In-car gps navigation: engagement with and disengagement from the environment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1675–1684, 2008. → page 7

[82] Y.-C. Lin. The relationship between touchscreen sizes of smartphones and hand dimensions. In *International Conference on Universal Access in Human-Computer Interaction*, pages 643–650. Springer, 2013. → page 6

[83] A. Lundgard, Y. Yang, M. L. Foster, and W. S. Lasecki. Bolt: Instantaneous crowdsourcing via just-in-time training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2018. → page 7

[84] E. L. Meerwijk and J. M. Sevelius. Transgender population size in the united states: a meta-regression of population-based probability samples. *American journal of public health*, 107(2):e1–e8, 2017. → page 27

[85] C. Mendez, A. Sarma, and M. Burnett. Gender in open source software: what the tools tell. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, pages 21–24, 2018. → page 5

[86] D. Metaxa-Kakavouli, K. Wang, J. A. Landay, and J. Hancock. Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018. → page 5

[87] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012. → page 5

[88] J. Munafo, M. Diedrick, and T. A. Stoffregen. The virtual reality head-mounted display oculus rift induces motion sickness and is sexist in its effects. *Experimental brain research*, 235(3):889–901, 2017. → page 5

[89] C. Nass, Y. Moon, and N. Green. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10):864–876, 1997. → page 5

[90] M. W. Nielsen, J. P. Andersen, L. Schiebinger, and J. W. Schneider. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nature human behaviour*, 1(11): 791–796, 2017. → page 7

[91] T. Nomura. Robots and gender. *Gender and the Genome*, 1(1):18–25, 2017. → page 6

[92] C. M. of Style Online. *The Chicago manual of style*. The University of Chicago Press, Chicago, seventeenth edition, 2017. → page 14

[93] C. C. Perez. *Invisible women: Exposing data bias in a world designed for men*. Random House, 2019. → pages 1, 2, 5, 7, 37, 38

[94] C. Plain. Build an affinity for kj method. *Quality Progress*, 40(3):88, 2007. → page 21

[95] K. C. Rasmussen, E. Maier, B. E. Strauss, M. Durbin, L. Riesbeck, A. Wallach, V. Zamloot, and A. Erena. The nonbinary fraction: Looking towards the future of gender equity in astronomy. *arXiv preprint arXiv:1907.04893*, 2019. → page 6

[96] A. G. Reece and C. M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):1–12, 2017. → page 38

[97] H. M. Reeder. The effect of gender role orientation on same-and cross-sex friendship formation. *Sex Roles*, 49(3-4):143–152, 2003. → page 42

[98] K. E. Ringland, C. T. Wolf, H. Faucett, L. Dombrowski, and G. R. Hayes. "will i always be not social?" re-conceptualizing sociality in the context of a minecraft community for autism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1256–1269, 2016. → page 8

[99] P. A. Rochon, J. P. Clark, M. A. Binns, V. Patel, and J. H. Gurwitz. Reporting of gender-related information in clinical trials of drug therapy for myocardial infarction. *Cmaj*, 159(4):321–327, 1998. → page 6

[100] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. 2010. → page 38

[101] M. T. Ruiz-Cantero, C. Vives-Cases, L. Artazcoz, A. Delgado, M. d. M. G. Calvente, C. Miqueo, I. Montero, R. Ortiz, E. Ronda, I. Ruiz, et al. A framework to analyse gender bias in epidemiological research. *Journal of Epidemiology & Community Health*, 61(Suppl 2):ii46–ii53, 2007. → page 6

[102] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi:10.1145/3359246. URL https://doi.org/10.1145/3359246. → page 5

[103] M. K. Scheuerman, K. Spiel, O. L. Haimson, F. Hamidi, and S. M. Branham. HCI guidelines for gender equity and inclusivity, 2020. URL https://www.morgan-klaus.com/gender-guidelines.html. Accessed: Aug 26, 2020. → pages 7, 14, 17, 25, 39

[104] L. Schiebinger and M. Schraudner. Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering1. *Interdisciplinary Science Reviews*, 36(2):154–167, 2011. → page 5

[105] A. Schlesinger, W. K. Edwards, and R. E. Grinter. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5412–5427, 2017. → pages 1, 7, 8

[106] B. Scott and V. Conzola. Designing touch screen numeric keypads: Effects of finger size, key size, and key spacing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 41, pages 360–364. SAGE Publications Sage CA: Los Angeles, CA, 1997. → page 6

[107] V. K. Singh, M. Chayko, R. Inamdar, and D. Floegel. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, 2020. → page 5

[108] L. P. Soares, L. Nomura, M. C. Cabral, M. Nagamura, R. de Deus Lopes, and M. K. Zuffo. Virtual hang-gliding over rio de janeiro. In *SIGGRAPH Emerging Technologies*, page 29, 2005. → page 39

[109] K. Spiel, O. L. Haimson, and D. Lottridge. How to do better with gender on surveys: a guide for hci researchers. *interactions*, 26(4):62–65, 2019. → page 39

[110] K. Spiel, O. Keyes, and P. Barlas. Patching gender: Non-binary utopias in HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–11, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi:10.1145/3290607.3310425. URL https://doi.org/10.1145/3290607.3310425. → page 2

[111] K. Spiel, O. Keyes, A. M. Walker, M. A. DeVito, J. Birnholtz, E. Brulé, A. Light, P. Barlas, J. Hardy, A. Ahmed, J. A. Rode, J. R. Brubaker, and G. Kannabiran. Queer(ing) HCI: Moving forward in theory and practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–4, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi:10.1145/3290607.3311750. URL https://doi.org/10.1145/3290607.3311750. → page 29

[112] A. Strohmayer, R. Comber, and M. Balaam. Exploring learning ecologies among people experiencing homelessness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2275–2284, 2015. → page 8

[113] A. Strohmayer, M. Laing, and R. Comber. Technologies and social justice outcomes in sex work charities: fighting stigma, saving lives. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3352–3364, 2017. → page 8

[114] S. Stumpf, A. Peters, S. Bardzell, M. Burnett, D. Busse, J. Cauchard, and E. Churchill. Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends® in Human–Computer Interaction*, 13(1), 2019. → pages 6, 36

[115] L. Takayama. Toward a science of robotics: Goals and standards for experimental research. In *RSS workshop on good experimental methodology in robotics*, 2009. → page 35

[116] C. Tannenbaum, R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger. Sex and gender analysis improves science and engineering. *Nature*, 575(7781): 137–146, 2019. → pages 1, 2, 6, 39

[117] D. Thakkar, N. Sambasivan, P. Kulkarni, P. Kalenahalli Sudarshan, and K. Toyama. The unexpected entry and exodus of women in computing and HCI in india. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. → pages 1, 5

[118] A. Truong, S. Chen, E. Yumer, D. Salesin, and W. Li. Extracting regular fov shots from 360 event footage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018. → page 38

[119] S. Turkle. Computational reticence: Why women fear the intimate machine. In *Technology and women's voices*, pages 44–60. Routledge, 2004. → pages 2, 6

[120] S. T. Völkel, W. Wilkowska, and M. Ziefle. Gender-specific motivation and expectations toward computer science. In *Proceedings of the 4th Conference on Gender & IT*, pages 123–134, 2018. → pages 2, 5, 6

[121] C. L. Von Baeyer, D. L. Sherk, and M. P. Zanna. Impression management in the job interview: When the female applicant meets the male (chauvinist) interviewer. *Personality and Social Psychology Bulletin*, 7(1): 45–51, 1981. → page 5

[122] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *arXiv preprint arXiv:1501.06307*, 2015. → page 5

[123] M. West, R. Kraut, and H. Ei Chew. I'd blush if i could: closing gender divides in digital skills through education. 2019. → page 5

[124] M. Westerlund, S. Leminen, and M. Rajahonka. A topic modelling analysis of living labs research. *Technology Innovation Management Review*, 8(7), 2018. → page 24

[125] M. W. Wiederman. Volunteer bias in sexuality research using college student participants. *Journal of Sex Research*, 36(1):59–66, 1999. → pages 2, 6

[126] H. O. Witteman, M. Hendricks, S. Straus, and C. Tannenbaum. Are gender gaps due to evaluations of the applicant or the science? a natural experiment at a national funding agency. *The Lancet*, 393(10171):531–540, 2019. → page 5

[127] E. Won, P. Johnson, L. Punnett, T. Becker, and J. Dennerlein. Gender differences in exposure to physical risk factors during standardized computer tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 47, pages 1155–1158. SAGE Publications Sage CA: Los Angeles, CA, 2003. → page 6

[128] S. Yardi and A. Bruckman. Income, race, and class: exploring socioeconomic differences in family technology use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3041–3050, 2012. → page 8

[129] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018. → page 20

[130] R. Zhou, J. Hentschel, and N. Kumar. Goodbye text, hello emoji: Mobile communication on wechat in china. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 748–759, 2017. → page 42

[131] M. Ziefle and A. K. Schaar. Gender differences in acceptance and attitudes towards an invasive medical stent. *Electronic Journal of Health Informatics*, 6(2):13, 2011. → page 6

# Appendix A

# Interview Protocols

These protocols were used to elicit information about research practices regarding gender and participant recruitment in research.

## Preparation

Paper to have

- Print copy of this script
- Printed consent form
- Print copy of any provided materials
- Printed backup of demographics survey
- Compensation $ and receipt
- Pen

Electronics to have

- Audio Recorder (make sure it's charged)
- Computer as a backup for recording
- Online version of the survey

Test recording

## Introduction

Hello and Welcome.

So what do you know so far about the study?

Historically, HCI publications have had a gender imbalance in their research participants. With shifts in modern attitudes towards gender, we want to know how gender is currently being treated in HCI research.

We're interested in the intersection between gender and research practices. As you are well aware, a lot of the factors that go into the conception and design of research never make it into the research publications. Some common things that we're interested in learning about are things like who did what in the research, the theoretical and educational background of the people who were involved, and challenges that were encountered that changed the design of the research.

Through these interviews, we intend to gain a nuanced understanding of the explicit information published with regards to gender in research participants, as well as the implicit factors that both affect and are affected by the gender of research participants, across a variety of different research methods and departments.

Do you have any questions?

## Consent form, demographics, and recording

So the first thing I have is a consent form which you should have received, feel free to take as much time as you like to read this over, and if you have any questions, don't hesitate to ask me.

*provide consent form*

Are you ok with this interview being recorded?

*possibly begin recording*

Alright so the first thing I'm going to get you to do is fill out the short demographics survey.

*provide survey, good opportunity to get the recording going*

Thank you very much, now we can start with the interview.

## Demographics Questionnaire

1. Participant number (ask facilitator):
2. What age bracket do you fall into?
    - 20
    - 21-40
    - 41-55
    - 56-65
    - 66+
3. What is your gender? (multiple selection possible)
    - Man
    - Woman
    - Non-Binary
    - Prefer not to disclose
    - Prefer to self-describe:
4. How long have you been working with Human Subject Research?
    - 0-3 years
    - 3-5 years
    - 5-10 years
    - 10+ years
5. What is your current primary role in human subject research? (if not ongoing, what was your last primary role?)
    - Principle investigator
    - Research assistant
    - Supervisor not directly involved
    - Participant
    - Other

## Interview (Primary Researcher)

1. RESEARCH SUMMARY Ask researcher to describe the study
    (a) 30 second summary of study is sufficient
    (b) What was your role in the study?
        i. We're trying to get at the process of doing research, so if you can clarify as we go who was really in charge of doing what, that would be great.
    (c) Ask the user to describe the participants in the study
    (d) Is gender important in any part of your research? Why?
        i. If yes, What efforts did you make because of that?
        ii. If no, What would have had to been different about your study for gender to be important?
2. RECRUITING Examine recruitment material
    (a) IF NONE PROVIDED What recruitment methods did you use?
    (b) Did you have to modify your research method because of your participants or change your participant recruiting because of your research method?
    (c) What were the study participants told they would be doing?
    (d) How was this decided on, and who decided it?
        i. Was based on your past work, or past work from your research group, or past training?
    (e) What were you looking for in participants?
    (f) Is this a usual recruitment method?
    (g) Was it effective?
    (h) What else did you use?
3. RESULTS OF RECRUITING Examine demographics results (in paper if nowhere else)

(a) What demographics results did you collect? Why?

i. IF GENDER COLLECTED How did you collect the gender results and why? What gender categories did you use and why?

(b) Was the demographics collection straight forward or were there obstacles?

(c) How about things you didn't expect?

4. REPORTING Back to the paper

(a) Was reporting gender something that was considered? Why/why not?

(b) What did you change because of peer review feedback?

i. Did they require more from the research around the participant portion?

ii. Did they pick on any of the gender language?

iii. Did they question what you reported for participants?

5. GENDER THEORY

These questions are conceptual, you might feel like you need some time to think about it, and it's perfectly ok to not have a solid answer.

(a) What does Gender mean to you?

(b) Is how you think about gender represented in your research?

(c) Where is gender important in research in general?

(d) Did theories of gender come up in your formal education?

6. RESEARCH PARTICIPATION

(a) Why do you think the people in your study participated?

(b) If you have participated in research, why did you do it?

7. IF IT HASN'T BEEN COVERED

(a) How did you get into human subject research?

i. Was there any formal training

ii. Did gender come into the training at all?

(b) What guidance did you give you other researchers participating in the research?

## Interview (Research Manager)

For this interview, we're trying to get at the process of doing research. As we go, if you can make a point of mentioning who was in charge of doing what, that would be great. If there are questions about things you were not involved in, we can skip those questions, but it would be great to know who was in charge of doing them, and what role other research team members had in them.

1. ABOUT RESEARCH MANAGERS RESEARCH

   These questions are about research participants in your research, and how gender factors into research in general. We are trying to understand what intersections there might be between gender and research in your research area and group.

   We are interested in the end to end process of research, from design and funding to publication.

   (a) Is gender important in any part of your research, or in your research group? Why?
      i. If yes, What efforts do you make because of that?
      ii. If no, What would need to be different about your research for gender to be important?

   (b) At what stage is gender most likely to be important in your research group?

   (c) How did you get into human subject research?
      i. Was there any formal training?
      ii. Did gender come into the training at all?

   (d) What guidance about doing research with human subjects do you give other researchers?

(e) Can you recall any experience (in this study or any other) where you had to modify your research method because of your participants or change your participant recruiting because of your research methods?

2. SUMMARY OF THIS RESEARCH

This section is to get a general idea about this specific study from your perspective. We chose this study to ground the interview, but if you feel it's relevant, please feel free to talk about any of your other research.

(a) Can you give me a 30 second summary of this research from your perspective? What was it about?

(b) What was your role in the study?

(c) Why did you recruit participants for this study?

(d) Did it matter who was recruited?

(e) How was this decided on, and who decided it?

i. Was based on your past work, or past work from your research group, or past training?

3. REPORTING

For this section, we want to understand end to end what was eventually reported about the participants.

(a) Who made decisions about what demographics to collect from participants?

(b) What demographics results were collected from the participants?

i. IF GENDER COLLECTED How did you collect the gender results and why? What gender categories did you use and why?

(c) How involved were you in writing the paper and revisions to the paper?

(d) Was reporting gender something that was considered? Why/why not?

(e) What did you change because of peer review feedback?

i. Did they require more from the research around the participant portion?

ii. Did they pick on any of the gender language?

iii. Did they question what you reported for participants?

4. GENDER THEORY

   These questions are conceptual, you might feel like you need some time to think about it, and it's perfectly ok to not have an answer.

   (a) What does Gender mean to you?

   (b) Is how you think about gender represented in your research?

   (c) Where is gender important in research in general?

   (d) Did theories of gender come up in your formal education?

5. RESEARCH PARTICIPATION

   Extra questions if we have time.

   (a) Why do you think the people in your study participated?

   (b) If you have participated in research, why did you do it?

# Appendix B

# Guidelines and Best Practices for Gender Reporting

Our interviews with researchers revealed that researchers are aware of non-binary gender, but do not know how to handle it. We therefore investigated current guidelines for handling gender in research reporting, and present the results of this investigation here.

We investigated whether there were any correlations between convergence on these guidelines and gender representation, but found no interesting trends. However, these guidelines are extremely new, so it might be worth tracking them in future publications.

These guidelines are provided by the APA[1] and Chicago[2] styleguides, and the Gender in HCI page by Scheuerman[3]. We list each guideline here, along with the relevant recommendation from the various sources.

1. Use 'men/women' in preference to 'male/female'

[1]American Psychological Association. 2020.Publication manual of the American Psychological Association: the official guide to APA style(7th ed.).American Psychological Association, Washington, DC.

[2]Chicago Manual of Style Online. 2017.The Chicago manual of style(seventeenth ed.). The University of Chicago Press, Chicago.

[3]Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI Guidelines for Gender Equity andInclusivity. Retrieved Aug 26, 2020 from https://www.morgan-klaus.com/gender-guidelines.html

(a) Scheuerman et. al., A-3: “instead of writing ‘Female participants were more likely to disagree’ you could write ‘Women were more likely to disagree.’ ”

(b) APA, Gender and Noun Usage: “to reduce the possibility of stereotypic bias and avoid ambiguity, use specific nouns to identify people or groups of people (e.g., women, men, transgender men, trans men, transgender women, trans women, cisgender women, cisgender men, gender-fluid people)”

2. Avoid use of female, male as nouns

(a) Scheuerman et. al., C-2 : “Transgenders (or ‘a transgender’) is similarly incorrect and offensive. Choosing to use transgender as a noun, rather than an adjective, removes ‘people’ or ‘individuals,’ thus dehumanizing transgender people.”

(b) APA, Gender and Noun Usage: “Use ‘male’ and ‘female’ as adjectives (e.g., a male participant, a female experimenter) when appropriate and relevant. Use ‘male’ and ‘female’ as nouns only when the age range is broad or ambiguous or to identify a transgender person’s sex assignment at birth (e.g., ‘person assigned female at birth’ is correct, not ‘person assigned girl at birth’). Otherwise, avoid using ‘male’ and ‘female’ as nouns and instead use the specific nouns for people of different ages (e.g., women).”

(c) APA, Examples of Bias-Free Language: Males/females is described as problematic men/women as prefered.

3. Avoid binary language

(a) Scheuerman et. al., A7: ”Avoid binary language.”

(b) Scheuerman et. al. G-5: “Avoid using the term ‘Other’ on surveys, as it implies gender norms that are othering to non-binary participants.”

(c) APA, Terms That Imply Binaries: “Avoid referring to one sex or gender as the ‘opposite sex’ or ‘opposite gender’; appropriate wording may be ‘another sex’ or ‘another gender.’ ”

4. Report all genders

    (a) Scheuerman et. al. A-8: “Do not define demographics of participants by the ‘outlier.’ For example, only stating that ‘47% of our participants were women’ without also defining the remaining 53%, which could be made up of men, non-binary and/or agender people”

5. Do not use “he” as a gender neutral pronoun. Use “they”

    (a) Scheuerman et. al., A-2: “Use ‘they’ as a gender neutral pronoun, rather than ‘he or she,’ ‘he/she,’ or ‘s/he’ ”

    (b) APA, Gender and Pronoun Useage: “When referring to individuals whose identified pronouns are not known or when the gender of a generic or hypothetical person is irrelevant within the context, use the singular ‘they’ to avoid making assumptions about an individual’s gender. Use the forms ‘they,’ ‘them,’ ‘theirs,’ and so forth. Sexist bias can occur when pronouns are used carelessly, as when the pronoun ‘he’ is used to refer to all people, when a gendered pronoun is used exclusively to define roles by sex (e.g., ‘the nurse . . . she’)”

    (c) Chicago, 5.48: “When referring specifically to a person who does not identify with a gender-specific pronoun, however, they and its forms are often preferred.

# Appendix C

# Data Dictionary

This data dictionary structures participant gender data from published research. It is not meant to be representative of gender itself, only of what is published about participant gender in research.

## Attribute Explanation

| | |
|---|---|
| Semantic unit | Names are descriptive and unique within the Data Dictionary. |
| Semantic components | The Semantic components each have their own entries later in the Data Dictionary. A Semantic unit that has Semantic components does not have any value of its own. Only Semantic units at the lowest level have values. |
| Definition | The meaning of the Semantic unit. |
| Rationale | Why the Semantic unit is needed, if this is not self-evident from the definition |

| | |
|---|---|
| Data constraint | How the value of the Semantic unit should be encoded. Some common data constraints are:<br>• Container - The Semantic unit is an umbrella for two or more Semantic components and has no value of its own.<br>• None - The Semantic unit can take any form of value.<br>• Number - The value must be a numeric integer<br>• Text - The value must be text<br>• Controlled Vocabulary - The value must be one of a set of values.<br>• Reference - A reference to another item in the dataset |
| Examples | One or more examples of values the Semantic unit may take. Examples are intended to be illustrative. |
| Repeatability | A Semantic unit designated as "Repeatable" can take multiple values under its parent unit. A unit designated as "Not Repeatable" will have a single value under its parent, though there may be multiple values of this unit under multiple parents. |
| Obligation | Whether a value for the Semantic unit is mandatory or optional. |
| Creation / Maintenance notes | Notes about how the values for the Semantic unit may be obtained and/or updated. |
| Usage notes | Information about the intended use of the Semantic unit, or clarification of the definition. |

## Overview

- 1 Paper
  - 1.1 Participant set
    - 1.1.1 Participant total count
    - 1.1.2 Participants Reported as ⚧
      - 1.1.2.1 Text Indicator ⚧

        - 1.1.2.2 Number Classified As ⚧
        - 1.1.2.3 ⚧Numeric Text
    - 1.1.3 Participants Reported as ♂
        - 1.1.3.1 Text Indicator ♂
        - 1.1.3.2 Number Classified As ♂
        - 1.1.3.3 ♂Numeric Text
    - 1.1.4 Participants Reported as ♀
        - 1.1.4.1 Text Indicator ♀
        - 1.1.4.2 Number Classified As ♀
        - 1.1.4.3 ♀Numeric Text
    - 1.1.5 Participants Reported as Did Not Disclose
        - 1.1.5.1 Text Indicator for Did Not Disclose
        - 1.1.5.2 Number Classified as Did Not Disclose
        - 1.1.5.3 Did Not Disclose Numeric Text
    - 1.1.6 Gender Reported with Binary Assumption
        - 1.1.6.1 One Gender Reporting
        - 1.1.6.2 Balanced Gender Reporting
    - 1.1.7 Participant Source
    - 1.1.8 Participant Total Count Numeric Text
    - 1.1.9 Estimated Participant Count
    - 1.1.10 Auxiliary Participant Count
  - 1.2 Number of Participant Sets
  - 1.3 Paper DOI
- 2 ML Sentence
  - 2.1 Sentence Text
  - 2.2 Sentence Schema Target

## Data Dictionary

| Semantic unit | 1 Paper |
|---|---|
| Semantic components | 1.1 Participant Set<br>1.2 Number of Participant Sets<br>1.3 Paper DOI |
| Definition | A paper refers to a complete publication, put out in PDF format at a conference or in a journal. |
| Data constraint | Container |
| Repeatability | Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | The database of papers will be created prior to data extraction. |
| Usage notes | |

| Semantic unit | 1.1 Participant set |
|---|---|
| Semantic components | 1.1.1 Participant total count<br>1.1.2 Participants Reported as ⚧<br>1.1.3 Participants Reported as ♂<br>1.1.4 Participants Reported as ♀<br>1.1.5 Participants Reported as Did Not Disclose Gender<br>1.1.6 Nothing About Gender Reported<br>1.1.7 Participant Source<br>1.1.8 Participant Total Count Numeric Text<br>1.1.9 Estimated Participant Count |
| Definition | A participant set is a batch of people sampled in a recruitment. This is a container for the data about a set of participants. |
| Rationale | Papers can contain multiple studies, and those studies will often recruit participants and report information about their participants differently. We therefore need to keep these sets of participants separate. |

| | |
|---|---|
| Data constraint | Container |
| Repeatability | Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | If a paper reports multiple studies with the exact same group of participants, these studies should be counted as one participant set. If there is a difference in the participant population, then count it as two. (for example, if the researcher recruits a large sample for one study, and then a uses a deliberate subset of those participants, report as two sets) |
| Usage notes | |

| Semantic unit | 1.1.1 Participant total count |
|---|---|
| Semantic components | |
| Definition | The total number of participants in this set. |
| Rationale | Often only a portion of participants will be classified with the various attributes, so having the total number is important for cross validation and validating participation rates. |
| Data constraint | Number |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | Occasionally the exact number of participants is withheld, this value is therefore optional. |
| Usage notes | |

| Semantic unit | 1.1.2 Participants Reported as ⚧ |
|---|---|
| Semantic components | 1.1.2.1 Text Indicator ⚧<br>1.1.2.2 Number Classified As ⚧<br>1.1.2.3 ⚧Numeric Text |
| Definition | Container for reporting complex participant gender. |

| | |
|---|---|
| Rationale | This container will collect everything in the participant descriptions that fall into the category of complex gender. If there are multiple descriptors, they should be summed under this category. It is possible that a descriptor will fall under this category as well as others such as 1.1.4 Participants Reported as ♀and 1.1.3 Participants Reported as ♂. The study may not report any ⚧gender, therefore, this category is optional. |
| Data constraint | Container |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| | |
|---|---|
| Semantic unit | 1.1.2.1 Text Indicator ⚧ |
| Semantic components | |
| Definition | The word or phrase that indicates ⚧ |
| Rationale | Collected for verification and to analyse exact word usage. There may be multiple words used. |
| Data constraint | Text |
| Examples | non-binary, trans, genderfluid, two-spirit, questioning |
| Repeatability | Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | There must be a word or phrase that indicates ⚧for 1.1.2 Participants Reported as ⚧to be included. Identical words should be collapsed |
| Usage notes | |

| | |
|---|---|
| Semantic unit | 1.1.2.2 Number Classified As ⚧ |
| Semantic components | |

| | |
|---|---|
| Definition | The number of participants classified as ⚧ |
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.2.3 ⚧numeric text' should be filled, not both. |
| Data constraint | Number |
| Examples | 6 |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.2.3 ⚧Numeric Text |
|---|---|
| Semantic components | |
| Definition | Text indicating the number of participants classified as ⚧, but that cannot be converted to an integer. |
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.2.2 Number Classified As ⚧' should be filled, not both. |
| Data constraint | Text |
| Examples | Predominantly, more than half, around ten, some |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | If a paper gives a clear fraction "half of the participants", and provides the total number of participants, this can be converted to an integer, and should be recorded under Number Classified As. However, if they say "Close to half" or do not provide the total number of participants, this cannot be converted to an integer. |
| Usage notes | |

| Semantic unit | 1.1.3 Participants Reported as ♂ |
|---|---|

| Semantic components | 1.1.3.1 Text Indicator<br>1.1.3.2 Number Classified As ♂<br>1.1.3.3 ♂Numeric Text |
|---|---|
| Definition | Container for participant gender reported variously as man, male, masculine, etc. |
| Rationale | This container will collect everything in the participant descriptions that fall into the above categories. If there are multiple descriptors, they should be summed under this category. The study may not report any ♂gender, therefore, this category is optional. |
| Data constraint | Container |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.3.1 Text Indicator ♂ |
|---|---|
| Semantic components | |
| Definition | The word or phrase that indicates ♂ |
| Rationale | Collected for verification and to analyse exact word usage. There may be multiple words used. |
| Data constraint | Text |
| Examples | Man, male, boy |
| Repeatability | Repeatable |
| Obligation | Mandatory |

| | |
|---|---|
| Creation / Maintenance notes | There must be a word or phrase that indicates ♂for 1.1.3 Participants Reported as ♂to be included. Identical words should be collapsed. Note that the only words that are used to describe participants should be included. Words used to talk about the participants, (i.e. "When we talked to participant 1, he said") should not be considered. |
| Usage notes | |

| Semantic unit | 1.1.3.2 Number Classified As ♂ |
|---|---|
| Semantic components | |
| Definition | The number of participants classified as ♂ |
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.3.2 Number Classified As ♂' should be filled, not both. |
| Data constraint | Number |
| Examples | 6 |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.3.3 ♂Numeric Text |
|---|---|
| Semantic components | |
| Definition | Text indicating the number of participants classified as ♂, but that cannot be converted to an integer. |
| Rationale | This is an if provided value, hence optional. |
| Data constraint | Text |
| Examples | Predominantly, more than half, around ten |
| Repeatability | Repeatable |
| Obligation | Optional |

| | |
|---|---|
| Creation / Maintenance notes | If a paper gives a clear fraction “half of the participants”, and provides the total number of participants, this can be converted to an integer, and should be recorded under Number Classified As. However, if they say “Close to half” or do not provide the total number of participants, this cannot be converted to an integer. |
| Usage notes | |

| Semantic unit | 1.1.4 Participants Reported as ♀ |
|---|---|
| Semantic components | 1.1.4.1 Text Indicator<br>1.1.4.2 Number Classified As ♀<br>1.1.4.3 ♀Numeric Text |
| Definition | Container for participant gender reported variously as woman, female, feminine, etc. |
| Rationale | This container will collect everything in the participant descriptions that fall into the above categories. If there are multiple descriptiors, they should be summed under this category. The study may not report any ♀gender, therefore, this category is optional. |
| Data constraint | Container |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.4.1 Text Indicator ♀ |
|---|---|
| Semantic components | |
| Definition | The word or phrase that indicates ♀ |
| Rationale | Collected for verification and to analyse exact word usage. There may be multiple words used. |

| | |
|---|---|
| Data constraint | Text |
| Examples | Woman, female, girl |
| Repeatability | Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | There must be a word or phrase that indicates ♀for 1.1.4 Participants Reported as ♀to be included. Identical words should be collapsed. Note that the only words that are used to describe participants should be included. Words used to talk about the participants, (i.e. "When we talked to participant 1, she said") should not be considered. |
| Usage notes | |

| Semantic unit | 1.1.4.2 Number Classified As ♀ |
|---|---|
| Semantic components | |
| Definition | The number of participants classified as ♀ |
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.4.3 ♀Numeric Text' should be filled, not both. |
| Data constraint | Number |
| Examples | 6 |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.4.3 ♀Numeric Text |
|---|---|
| Semantic components | |
| Definition | Text indicating the number of participants classified as ♀, but that cannot be converted to an integer. |

| | |
|---|---|
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.4.2 Number Classified As ♀' should be filled, not both. |
| Data constraint | Text |
| Examples | Predominantly, more than half, around ten |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | If a paper gives a clear fraction "half of the participants", and provides the total number of participants, this can be converted to an integer, and should be recorded under Number Classified As. However, if they say "Close to half" or do not provide the total number of participants, this cannot be converted to an integer. |
| Usage notes | |

| | |
|---|---|
| Semantic unit | 1.1.5 Participants Reported as Did Not Disclose |
| Semantic components | 1.1.5.1 Text Indicator for Did Not Disclose<br>1.1.5.2 Number Classified as Did Not Disclose<br>1.1.5.3 Did Not Disclose Numeric Text |
| Definition | This captures the data that the author reported that some participant(s) did not disclose their gender. |
| Rationale | Giving participants the option to withhold gender is a standard survey option, and needs to be captured when reported. |
| Data constraint | Container |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.5.1 Text Indicator for Did Not Disclose |
|---|---|
| Semantic components | |
| Definition | The phrase that indicates that gender was not disclosed by the participant. |
| Rationale | Collected for verification and to analyse exact word usage. There may be multiple words used. |
| Data constraint | Text |
| Examples | Preferred not to disclose, did not disclose |
| Repeatability | Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | There must be a phrase that indicates that gender was not disclosed, otherwise there was nothing reported. |
| Usage notes | |

| Semantic unit | 1.1.5.2 Number Classified as Did Not Disclose |
|---|---|
| Semantic components | |
| Definition | The number of participants classified as Did Not Disclose |
| Rationale | This is an if provided value, hence optional. |
| Data constraint | Number |
| Examples | 6 |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.5.3 Did Not Disclose Numeric Text |
|---|---|
| Semantic components | |
| Definition | Text indicating the number of participants classified as Did Not Disclose, but that cannot be converted into an integer. |

| | |
|---|---|
| Rationale | This is an if provided value, hence optional. Either this OR '1.1.5.2 Number Classified As Did Not Disclose' should be filled, not both. |
| Data constraint | Text |
| Examples | Predominantly, more than half, around ten, some, a few |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | If a paper gives a clear fraction "half of the participants", and provides the total number of participants, this can be converted to an integer, and should be recorded under Number Classified As. However, if they say "Close to half" or do not provide the total number of participants, this cannot be converted to an integer. |
| Usage notes | |

| | |
|---|---|
| Semantic unit | 1.1.6 Gender Reported with Binary Assumption |
| Semantic components | 1.1.6.1 One Gender Reporting<br>1.1.6.2 Balanced Gender Reporting |
| Definition | A container to capture all information about the author reporting gender which must be interpreted under a binary assumption of gender (i.e. two genders, men and women) |
| Rationale | The gender numbers reported are often insufficient to capture this nuance, therefore it should be captured separately. |
| Data constraint | Container |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.6.1 One Gender Reporting |
|---|---|
| Semantic components | |
| Definition | The text the author used to report a single gender when it is obvious that the author means to indicate that the rest of the participants are the binary gender not reported. |
| Rationale | |
| Data constraint | Text |
| Examples | “12 participants (6 female)”, “57% of the participants were men” |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.6.2 Balanced Gender Reporting |
|---|---|
| Semantic components | |
| Definition | The text which reports gender as ‘balanced’, in other words that the author wishes to indicate that the set contains an equal number of men and women participated in the study. |
| Rationale | |
| Data constraint | Text |
| Examples | “Gender was balanced”, “we balanced gender in all conditions” |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.7 Participant Source |
|---|---|
| Semantic components | |
| Definition | Text containing information about where the participants were obtained from |
| Rationale | Very general bucket to collect this information as the reporting varies widely. |
| Data constraint | Text |
| Examples | Facebook post, posters, department email lists, interest groups, approaching people in the hallway, convenience sampling, snowball sampling, “Our participants were all students in our department”, “We recruited dancers from a local club” |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |
| Usage notes | |

| Semantic unit | 1.1.8 Participant Total Count Numeric Text |
|---|---|
| Semantic components | |
| Definition | The numeric information about the total number of participants that took part in this study. |
| Rationale | This is for when total count for participants is not reported, but some numeric information is provided that cannot be converted to an integer. Either this OR ‘1.1.1 Participant Total Count’ should be filled, not both. |
| Data constraint | Text |
| Examples | Several, around half of the survey participants, close to a dozen, some |
| Repeatability | Not Repeatable |
| Obligation | Optional |

| | |
|---|---|
| Creation / Maintenance notes | Occasionally the exact number of participants is withheld, this value is therefore optional. |
| Usage notes | |

| Semantic unit | 1.1.9 Estimated Participant Count |
|---|---|
| Semantic components | |
| Definition | Estimated total number of participants |
| Rationale | The total number of participants can sometimes be estimated even if not directly stated. |
| Data constraint | Number of participants can sometimes be estimated or guessed at. If there is no definite total count of participants. |
| Examples | ”We completed 27 interviews.” |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |

| Semantic unit | 1.1.10 Auxiliary Participant Count |
|---|---|
| Semantic components | |
| Definition | Number of participants that were part of an auxiliary set |
| Rationale | Participants that take part in pilot studies are generally not subjected to the same rigour with regards to reporting and recruiting. Therefore they should be considered separately for main study participants. |

| | |
|---|---|
| Data constraint | Participant counts should only be marked as Auxiliary when the participants set is associated with a pilot study and there is a main study in the paper. Some papers will contain only one study which they call a pilot study, but in this case, as the 'pilot' study is the main study for the purposes of the publication, it should be marked as a participant count. |
| Examples | We conducted a pilot study with 4 participants. |
| Repeatability | Not Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | |

| Semantic unit | 1.2 Number of Participant Sets |
|---|---|
| Semantic components | |
| Definition | The number of participant sets the paper contains. |
| Rationale | The number of participant sets in a paper can easily be 0. If this is the case the paper will contain no participant data, and that fact should be captured. |
| Data constraint | Number |
| Repeatability | Not Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | This number will vary from 0 upwards. |
| Usage notes | |

| Semantic unit | 1.3 Paper DOI |
|---|---|
| Semantic components | |
| Definition | An ID that uniquely identifies the paper and allows it to be connected with other metadata. |
| Data constraint | Text |

| | |
|---|---|
| Repeatability | Not Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | |
| Usage notes | |

## Auxiliary Data

This is not part of the official study data, but it is additional data which will be included with the data set as it is of use to future researchers. As such, it is worth having a formal definition of it included with the above data schema.

| | |
|---|---|
| Semantic unit | 2 ML Sentence |
| Semantic components | 2.1 Text<br>2.2 Schema Target |
| Definition | A sentence from a paper which directs a user to where a piece of data might be found. |
| Rationale | These sentences are needed to train a machine learning module to detect relevant sentences |
| Data constraint | Container |
| Repeatability | Repeatable |
| Obligation | Optional |
| Creation / Maintenance notes | Should be created with every piece of data that is pulled from the text.<br>In most instances, the piece of data should be directly copyable from the sentence text.<br>In some cases direct copy is not possible as the piece of data may require interpretation (e.x. "40 undergraduate students at a large university in the northeastern United States (22.5% male)" = [1.1.3.2 Number Classified As ♂] 9). In that case the sentence which contains the data required for interpretation should be taken. |

| | |
|---|---|
| Usage notes | |

| Semantic unit | 2.1 Sentence Text |
|---|---|
| Semantic components | |
| Definition | The text of the sentence |
| Rationale | |
| Data constraint | Text |
| Examples | “That leaves 187 participants.”, “Participants consisted of 40 undergraduate students at a large university in the northeastern United States (22.5% male).” |
| Repeatability | Not Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | In cases where multiple sentences may be said to contain the data, the best or most informative sentence should be selected. |
| Usage notes | |

| Semantic unit | 2.2 Sentence Schema Target |
|---|---|
| Semantic components | |
| Definition | A link to the piece of data which this ML Sentence provides. |
| Rationale | Required in order to sort the sentences for training different models. |
| Data constraint | Reference |
| Examples | “[1.1.8 Participant Total count]”, “[1.1.3.2 Number Classified As ♂]” |
| Repeatability | Not Repeatable |
| Obligation | Mandatory |
| Creation / Maintenance notes | |
| Usage notes | |

# Appendix D

# Gender Keywords

The list in Fig D.1 was compiled from several sources, including the *HCI Guidelines for Gender Equity and Inclusivity* page[1], a list of available gender identities from Facebook[2], and some words taken from a list used to create gender neutral word embeddings[3]. We used this list for a custom gender search feature in MAGDA. We do not claim that this list is comprehensive, but we found it effective as an additional check for gender data in papers, and provide it here for the use of other researchers.

[1] https://www.morgan-klaus.com/gender-guidelines.html, accessed 2020-09-13

[2] https://abcnews.go.com/blogs/headlines/2014/02/heres-a-list-of-58-gender-options-for-facebook-users, accessed 2020-09-13

[3] https://github.com/uclanlp/gn_glove/tree/master/wordlist, accessed 2020-09-13

afab, agender, amab , androgyne, androgynous, aporagender, assigned female at birth, assigned male at birth, aunt, aunts, bachelor, bachelorette, bachelors, bigender, boy, boyfriend, boyfriends, boyhood, boys, brother, brothers, cis, cis female, cis male, cis man, cis woman, cisgender, cisgender female, cisgender male, cisgender man, cisgender woman, daughter, daughters, demi-agender, demi-boy, demi-fluid, demi-gender, demi-girl, demi-non-binary, effeminate, f2m, father, fathers, female, female to male, females, feminism, ftm, gay, gays, gender, gender confusion, gender f*ck, gender fluid, gender indifferent, gender neutral, gender nonconforming, gender questioning, gender variant, genderfluid, genderflux, genderless, genderqueer, gentleman, gentlemen, girl, girlfriend, girlfriends, girlhood, girls, granddaughter, granddaughters, grandfather, grandfathers, grandma, grandmother, grandmothers, grandson, grandsons, graygender, he, her, hers, herself, him, himself, his, househusband, househusbands, housewife, housewives, husband, husbands, intergender, intersex, ladies, lady, lesbian, lesbians, m2f, male, male to female, males, man, masculism, maverique, maxigender, men, mother, mothers, mr., mrs., mtf, multigender, nephew, nephews, neutrois, niece, nieces, non-binary, omnigender, pangender, paternity, polygender, she, sir, sister, sisters, son, sons, stepdaughter, stepdaughters, stepfather, stepfathers, stepmother, stepmothers, stepson, stepsons, trans, trans fem, trans female, trans feminine, trans femme, trans male, trans man, trans masc, trans masculine, trans person, trans woman, trans*, trans* female, trans* male, trans* man, trans* person, trans* woman, transfeminine, transgender, transgender female, transgender male, transgender man, transgender person, transgender woman, transmasculine, transsexual, transsexual female, transsexual male, transsexual man, transsexual person, transsexual woman, trigender, two-spirit, uncle, uncles, widow, widower, widowers, widows, wife, wives, woman, women

**Figure D.1:** The list of gender words used in the MAGDA gender word search system

# Appendix E

# Recruitment Classification Codebook

This table lists out the recruitment classifications, and the criteria for a paper to be labeled with the given classification.

**Table E.1:** The list of all recruitment classifications for research publications

| Recruitment Classification | Value | Criteria |
|---|---|---|
| Psychology Students | All | The participants were called psychology students, were in a psychology class, or were called students and recruited from a psychology university or department, or from a psychology student pool |
| | Some | Some of the participants fall under the above classification, but were not the only participants reported. |
| | No | There is not enough information to be able to say for certain that the study belongs under the above headings. |

| | | |
|---|---|---|
| Computer Science Students | All | The participants were called computer science majors or students, or were stated as coming from a Computer science class, or were specified as being graduate or undergraduate students from a Computer Science department.<br>Additionally, HCI, Computer Engineering, Electrical Engineering, Information Systems, CyberArts, Programming, Machine Learning, UX design, and any interdisciplinary Computer science programs, are all close enough to be considered CS |
| | Some | Participants fall into the above classification, but were not the only participants described.<br>For the sake of consistency, a study that states that some of the participants were students, and some of the participants came from a computer science department or background (or any of the others), will be included under this heading. |
| | No | There is not enough information to be able to say for certain that the study belongs under the above headings.<br>Industrial and mechanical engineering are not CS |
| Children | All | Participants must be all under 16 or highschool students |

| | | |
|---|---|---|
| | Some | Must include participants under 16 or highschool students, as well as participants who are not. |
| | No | No participants specified as under 16 or a highschool student |
| Patients and Participants with Illnesses | Yes | Described as patients, in a hospital, clinic, or care facility, or as having a disease or illness, e.g. cancer, lower back pain, diabetes, Parkinson's Disease, etc. |
| | No | No disease or illness described. Blindness, deafness, autism, etc. are disabilities, not diseases. Motor impairments, while they can be caused by diseases, are not diseases themselves. A disorder is not a disease. |
| Blind Participants | Yes | Some participants were described as blind or visually impaired. |
| | No | No participants described as blind. Colorblind was not counted. |
| Amazon Mechanical Turk (MTurk) | All | The participants were all recruited from MTurk. |
| | Some | There were MTurk participants, but they were not the only participants. |
| | Other | No participants were recruited from MTurk, but some participants were recruited from another crowdsourcing platform (Prolific, Crowdflower, etc.). |
| | No | The paper did not report sufficient information to classify as any of the above. |

| | | |
|---|---|---|
| Convenience sampling | Yes | "Snowball sampling", "word of mouth", "convenience sampling", or "purposeful sampling" were reported as having been used. 'Snowball recruiting', 'snowball sampling', and 'snowball sample' are equivalent.<br>Participants being requested to share the study also belongs in this category, but as the words used to describe this vary widely, we do not include it in the classification |
| | No | None of the above specified. Personal networks, unless specified as a convenience sample, were not counted |
| Participant Pool | Yes | Participants come from a designated research participant pool, mailing list of people interested in research participation, or participant database. The mailing list must be for people interested in research, department or other mailing lists are not counted. |
| | No | No pool is specified. |

# Appendix F

# Topic Modelling Additional Material

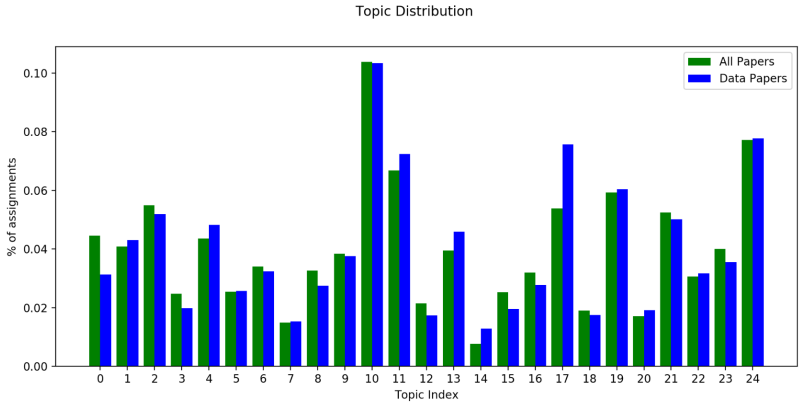


**Figure F.1:** The distribution of topics in the sample vs the distribution of CHI overall.

To classify papers by research area, we assigned topics using the MALLET library. Topic modeling can be used to categorize large text corpuses automatically. Mallet uses Latent Dirichlet Allocation (LDA) to infer a probability distribution of topics on each paper of our dataset. We chose twenty-five topics through both numerical analysis of perplexity and coherence scores and expert opinions from experienced HCI researchers, displayed in table F.1 and figure F.2. Papers were

cleaned of stopwords, malformed data, and venue words prior to modelling. As cross validation of our paper sampling technique, we compared the distribution of topics in our sampled papers to that of CHI over, and found our papers were representative (see Fig F.1).

**Table F.1:** The 25 topics selected to classify the CHI corpus from 1981 to 2020, sorted by $\alpha$ value.

| Topic | Assigned Label | Dirichlet $\alpha$[1] | Top 20 Topic Words |
|---|---|---|---|
| 10 | Usability Study (Generic Topic) | 0.31787 | participants study task results condition effect significant conditions experiment time tasks participant table found studies number effects test differences questions |
| 24 | Design (Generic Topic) | 0.26953 | design digital people hci work research experience technology experiences interaction ways paper social space practices sense make participants personal e.g |
| 11 | Usability Study (Generic Topic) | 0.21943 | design research process designers users user tools methods usability project ideas development work participants evaluation paper researchers tool software problems |
| 19 | Mobile Computing | 0.18919 | participants users mobile phone time people devices user study device app information messages data work application usage location email day |
| 2 | Data Analysis (General Topic) | 0.17998 | data model models user accuracy algorithm set features system work figure number results approach based analysis dataset time training human |

[1]Topics with large $\alpha$ values are generally built on topics that are very frequent in documents, but are seldom the "main research focus." This can be seen with topic 10 Usability Study, for example, which contains words regarding experiments involving participants. Many CHI papers will have these as a part of their research, but the experiment itself is not the research focus. As a general rule, topics with a more moderate $\alpha$ values are the most interesting topics.

| | | | |
|---|---|---|---|
| 17 | Programming Tools | 0.15998 | user interface system code users task application programming computer figure model program systems interfaces tasks software time command state actions |
| 21 | Virtual Environments | 0.1479 | display user virtual interaction objects physical object system users figure camera space displays view hand environment motion image screen participants |
| 4 | Collaboration | 0.14726 | work group information workers communication shared team time support collaborative groups tasks task members collaboration activities data activity working system |
| 1 | Visualization | 0.1406 | data visual visualization figure color visualizations design map view elements time layout space drawing interface tool graph users tools image |
| 13 | Information Search | 0.13283 | search information users web text user page content pages reading query results document items images tags participants number queries image |
| 23 | Social Media | 0.10873 | social online people users media content facebook community information friends posts news support twitter comments communities post sharing participants research |
| 9 | Eye tracking | 0.10859 | target time task performance figure movement experiment gaze targets cursor pointing tasks selection error techniques mouse technique eye distance trials |

| | | | |
|---|---|---|---|
| 0 | Analysis (Generic Topic) | 0.10361 | paper participants page work study first data significant user's specific prior experience found research find analysis future based findings don't |
| 8 | Community Infrastructure | 0.10204 | community local data work public technology social hci research people service communities women services practices technologies access city issues infrastructure |
| 16 | Family and Home | 0.09273 | children participants family home parents technology people child social adults older activities support technologies families care study work research time |
| 5 | Audio-Visual Media | 0.09246 | video audio speech voice videos music language sound content system words live speaker recording speakers participants sounds time english viewers |
| 22 | Devices and Fabrication | 0.08983 | figure sensor device sensing force devices paper design sensors shape surface material materials objects user physical work fabrication touch printed |
| 6 | Touch Input | 0.08576 | gestures gesture touch input users participants hand user interaction keyboard finger figure study text screen pen device key typing devices |
| 18 | Teaching and Learning | 0.07758 | students learning feedback student teachers training skills learners questions educational classroom education class school learn teacher group knowledge system teaching |

| | | | |
|---|---|---|---|
| 3 | Health Metrics | 0.07753 | data health participants food emotional behavior activity support change mental people stress emotions positive study sleep tracking time emotion negative |
| 15 | Haptics and Simulation | 0.07035 | feedback participants visual tactile haptic driving stimuli system sound information study figure drivers blind audio driver cues car auditory thermal |
| 12 | Privacy and Security | 0.06845 | privacy participants users security data information password user access trust energy passwords study authentication control system account concerns online risk |
| 20 | Video Games | 0.05868 | game games players player play experience playing gaming social played avatar gameplay level challenge control character experiences motivation world characters |
| 7 | Medical Agents | 0.04666 | patients patient robot information agent medical care agents robots health human clinical system clinicians hospital conversational healthcare treatment study paper |
| 14 | Wikipedia | 0.00891 | wikipedia editors edits article osm work articles figure localness edit quantum vgi titles credibility dandelion editor wiki editing title map |

Roughly 7500 documents were used in modelling, consisting of CHI papers published between 1981 and 2020. Documents were obtained partially as text files provided by the ACM, and partially downloaded by researchers and converted to text files. The Python library pdftotext was used for conversion. Extensive data cleaning was necessary in order to prevent "junk word" topics from forming. These


**Figure F.2:** The word cloud that reflects the individual weight of each word in the 25 topics.

words included malformed data, words about venues (which needed to be removed from converted PDFs), and DOI links. A comprehensive list is provided in Fig F.3.

A few individual characters such as the UTF-8 “soft hyphen” were removed as the topic modelling library we used (MALLET) was splitting words based on these hidden characters. We also encountered an issue where “ligatures” (Fig F.4) were replaced with “.”. To combat this, words matching the regular expression [a-zA-Z]˙[a-zA-Z] had the “.” substituted for the character combinations below, and

“glasgow”, “scotland”, “canada”, “montreal”, “montréal”, “honolulu”, “hawaii”, “hi”, “denver”, “U.S.”, “usa”, “U.K.”, “ing”, “tion”, “con”, “rst”, “tions”, “pro”, “chi”, “inter”, “par”, “ment”, “partici”, “tive”, “pants”, “ndings”, “ticipants”, “ments”, “ers”, “thhe”, “tthe”, “thee”, “annd”, “oof”, “wwe”, “andd”, “ffor”, “ases”, “inn”, “thhat”, “tto”, “wwas”, “e.g.”, “i.e.”, “april”, “https://doi.org”, “pparticipants”, “wwith”, “aas”, “aand”, “abby”, “gendermag”, “ofthe”, “oon”, “iin”, “bby”, “e-nable”, “particippants”, “bbe”, “ffigure”, “wwere”, “aare”, “thhis”, “weere”, “foor”, “facet”, “nnot”, “participannts”, “onn”, “tim”, “withh”, “wiith”, “oour”, “subtle”, “aan”, “loci”, “forr”, “figuure”, “dis”, “nger”, “it’s”, “ass”, “twwo”, “canbe”, “fromm”, “froom”, “wwhen”, “wee”, “figuree”, “tthis”, “thaat”, “e-nable”, “bignav”, “aadhaar”, “http[s]://doiȯrg[0-9/j̇]*”, “https://dxdoiȯrg[0-9/j̇]*”, “\\xad”, “\\xa0”

**Figure F.3:** Junk words removed from text corpus.

“ff”, “fi”, “fl”, “ft”, “ffi”, “ffl”

**Figure F.4:** Ligatures which had to be filled in via a dictionary search.

checked against a dictionary of roughly 466000 english words. If a match was found, the word was substituted.

All topic modelling was done using the Latent Dirichlet Allocation implementation from the MALLET Library[2], we ran each model for 1000 iterations, with an optomization interval of 10. Python’s matplotlib was used for data visualizations.

The numerical measures perplexity and coherence are often used in deciding the number of topics suitable for a topic model. We calculated these for models ranging from 5 to 35 topics, and found a general trend of more topics optimizing their respective measure. Model visualizations (similar to Fig F.2) were judged by multiple experienced HCI researchers, and 25 topics was selected to be the optimal number.

The 25 topics are listed with the number of papers with enough data to calculate and the average DER of those papers in table F.2.

[2] http://mallet.cs.umass.edu/

**Table F.2:** Topic table sorted by mean DER

| Topic | Assigned Label | Paper Count | Mean DER | sd |
|---|---|---|---|---|
| 7 | Medical Agents | 39 | .06 | .39 |
| 16 | Family and Home | 70 | .04 | .44 |
| 3 | Health Metrics | 55 | .00 | .45 |
| 8 | Community Infrastructure | 39 | -.01 | .58 |
| 12 | Privacy and Security | 48 | -.03 | .30 |
| 23 | Social Media | 86 | -.04 | .40 |
| 24 | Design (Generic Topic) | 120 | -.04 | .45 |
| 19 | Mobile Computing | 160 | -.07 | .36 |
| 11 | Usability Study (Generic Topic) | 107 | -.11 | .40 |
| 0 | Analysis (Generic Topic) | 79 | -.11 | .38 |
| 14 | Wikipedia | 16 | -.12 | .30 |
| 20 | Video Games | 50 | -.16 | .38 |
| 5 | Audio-Visual Media | 61 | -.16 | .35 |
| 10 | Usability Study (Generic Topic) | 289 | -.16 | .34 |
| 18 | Teaching and Learning | 41 | -.16 | .41 |
| 13 | Information Search | 87 | -.18 | .34 |
| 4 | Collaboration | 69 | -.18 | .42 |
| 22 | Devices and Fabrication | 67 | -.19 | .41 |
| 17 | Programming Tools | 88 | -.20 | .39 |
| 15 | Haptics and Simulation | 64 | -.20 | .37 |
| 1 | Visualization | 90 | -.23 | .35 |
| 21 | Virtual Environments | 111 | -.23 | .34 |
| 2 | Data Analysis (Generic Topic) | 88 | -.24 | .37 |
| 6 | Touch Input | 98 | -.26 | .33 |
| 9 | Eye tracking | 100 | -.29 | .36 |

# Appendix G

# Inter-rater Reliability Statistics

We calculated Cohen's kappa for each item in our data classification. For each round of data collection, 10% of the data was annotated by both coders. We first did a small subset of the papers, and ran the inter-rater reliability on 16 papers which both coders annotated. After discussion, we proceeded to complete the full set. We calculated inter-rater reliability for participant count and gender statistics on 95 papers which both coders annotated from the second round of annotation.

After data was collected, we classified the recruitment information, and applied recruitment labels to each paper. Each coder labeled the duplicated papers they annotated, and we calculated Cohen's Kappa. After a discussion, we clarified the recruitment classifications, and relabeled the papers. Cohen's Kappa reported for the final classification. 136 papers were used in this classification.

[1] Inter-rater reliability for non-binary participants not calculated due to the extremely low number of studies containing this data.

**Table G.1:** Inter-rater reliability scores for numeric participant data

| Data Type | Kappa |
|---|---|
| Total count of participants | 0.729 |
| Number of women | 0.858[1] |
| Number of men | 0.869[1] |
| Gender words used | 0.816 |
| Reporting coverage | 0.889 |

**Table G.2:** Inter-rater reliability scores for participant recruitment data

| Recruitment Classification | Kappa |
|---|---|
| Computer Science students | 0.682 |
| MTurk | 0.794 |
| Psyc students | 1.0[2] |
| Children | 0.661 |
| Less-rigorous Sampling | 0.562[3] |
| Participant Pool | 0.664 |
| Illness | 1.0[2] |

[2] In both of these cases, only two or three instances appeared in the 136 papers, in which the annotators agreed in their tagging. Inter-rater reliability could not be calculated for studies using Blind participants as they did not appear in the duplicated set.

[3]This low score is primarily due to the fact that one annotator did not tag lines including "Word of Mouth" and "Sampling". The coverage of data is therefore somewhat questionable, but as each annotator did half the data, randomly selected, the missed instances are unlikely to skew analysis results.