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

Anna Offenwanger
anna.offenwanger@gmail.com
University of British Columbia
Vancouver, Canada

Alan Milligan
alanmil@student.ubc.ca
University of British Columbia
Vancouver, Canada

Minsuk Chang*
minsuk.chang@navercorp.com
School of Computing, KAIST
Daejeon, South Korea
Naver AI LAB
Seongnam, South Korea

Julia Bullard
julia.bullard@ubc.ca
University of British Columbia
Vancouver, Canada

Dongwook Yoon
yoon@cs.ubc.ca
University of British Columbia
Vancouver, Canada

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 1,147 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.

CCS CONCEPTS

• **Social and professional topics** → **Gender**; • **Human-centered computing** → Empirical studies in HCI; *User studies*.

KEYWORDS

gender, gender bias, user studies, participants, human subjects, HCI, CHI, human-computer interaction, research, data schema, dataset, meta-analysis

ACM Reference Format:
Anna Offenwanger, Alan Milligan, Minsuk Chang, Julia Bullard, and Dongwook Yoon. 2021. Diagnosing Bias in the Gender Representation of HCI Research Participants: How it Happens and Where We Are. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445383

*This work was done when this author was at KAIST

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445383

1 INTRODUCTION

Bias in the gender distribution of research participants has long been of concern within the human-computer interaction (HCI) community [32] because it can jeopardize the generalizability of research findings, perpetuate bias in data driven practices [38, 87], and make new technologies dangerous for underrepresented demographics [58, 109]. While gender bias has many dimensions [8, 33, 110], 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 [9, 29], 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. We supplement previous close reading approaches to investigating gender and related areas in HCI research [28, 69, 99] by sampling papers across all the 40 years of CHI [45] to provide a dataset that allows quantitative analysis of how gender is treated across different variables. This *data-driven* approach adds to the close reading approaches 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. Our analysis focuses on women and non-binary individuals who have historically been problematically underrepresented [2, 9, 29, 58, 109], though it is important to remember that gender underrepresentation affects dominant groups (typically men) as well [27].

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 measurable values that correlate with gender bias do not necessarily have a straightforward causal relationship with it. 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" [87, p. 17]. This can be seen in people reading "gender neutral" words as referencing men [22], which could be both an effect of women's exclusion from research [37, 61] and a cause for women to self-exclude [112, 113, 118]. 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 [58, 104], and there is no available data schema for participant gender representation 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. Barkhuus and Rode [9] extracted data from 358 papers published in 1983 to 2006, and Caine [29] from 465 papers published in 2014 through manual analysis. The rate of gender reporting is somewhere between 50% [9] and 70% [29], 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 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 examining 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 in HCI research participants: 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[1].
- Dataset Contribution: We provide the resulting gender dataset of study participants in the CHI proceedings, extracted from 1,147 papers published between 1981 and 2020, which includes gender reporting and reporting language, participant counts, and data on where participants were recruited from.

2 RELATED WORK

2.1 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 [32, 43, 66, 98], as well as supporting social justice [24, 68], 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 [8, 10, 11, 83]. Different kinds of gender bias crop up in different areas of HCI, including hiring [81, 114], retention [33, 110], publishing [17, 119], and willingness and ability to engage in computing fields [49, 80, 113, 116]. Stereotypes of technology users being predominantly men pervade the field [22], and it is possible for HCI researchers to intervene in these patterns with intentionally designed technology [8, 50]. Lack of attention to gender can cause harm through gender reductionism, especially to people with gender non-normative identities [65, 70], and through unintentional exclusion [87, 115], which can lead to role stereotyping [19, 101] and new technologies being difficult to use for underrepresented groups [2, 26, 79, 82, 96].

2.2 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 [53]. Gender Bias has been defined as "any set of attitudes and/or behaviors which favors one sex over the other" [16, 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" [95, 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 [7, 9, 55, 60]. The need to consider gender bias in subject sampling is acknowledged in other fields [36, 61, 89], especially medicine [12, 95], where

[1]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.

it has been linked to poor health outcomes and lower quality of life [21, 93]. Biased gender data can easily affect many kinds of HCI research because of gender differences in, for example, posture [120], finger size [77, 100], cognitive performance [56], learning [63], body image [42], and social behaviors [4, 41, 85] including technology acceptance [112, 123]. "More men than women participated in user studies" [29], but technologies are considered to be equally usable by all genders, which has led to problems [2, 109]. In addition to the problems that arise, researchers may miss opportunities, as gender analysis can lead to more accurate/statistically significant findings [109] and increases the audience for potential devices [108].

2.3 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 [112, 113, 118]. Researchers can purposefully exclude a gender for reasons connected to the research [34, 37]. There is also a link between author gender and attention to gender and sex analysis [84]. Since women are underrepresented as authors for computer science (CS) research publications [33, 75], 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" [76, p. 1678].

2.4 Data collection

Extracting gender data from publications is difficult because it is complex and inconsistent [99], and often absent altogether. While there are good reasons for data to be absent, such as a concern for participant privacy [72, p. 4], absent data contributes to the data silences [87], and these 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] [99], which makes it difficult to do any kind of automated extraction. In addition, the nuance of how researchers collect gender demographics is important [65]: for example, do "20 participants, (5 female)" and "5 of our 20 participants reported themselves as female" mean the same thing? Gender reporting guidelines are available [97], but older papers did not have these guidelines, and so they cannot be relied on. To handle this problem, we propose a data schema and gender collection guidelines, which we provide for the use of future researchers in this area.

2.5 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 [99]. There are many ways in which different dimensions of participant diversity can affect research. Such dimensions include ability [14, 20, 48, 92], gender modality [57], displacement [67], job stigma [107], homelessness [106], race and class, [121] and race and gender [25]. 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" [99, 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 [99]. 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.

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 conducted semi-structured interviews with HCI researchers about their research practices around participants, focusing on gender, research design, and decisions involved in recruiting and reporting. Each interview lasted about one hour. Detailed protocols are included in supplementary materials.

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 [45] and UbiComp'18 [46] as they are generally regarded as top-venues in HCI and for their scale and diversity. We 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).

To analyse the data, we used theoretical thematic analysis [23, p. 12]. The lead researcher conducted interviews and coded the transcripts, with two other researchers reading through six of them and discussing possible themes identified from the codes in multiple iterations. 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 [74] and psychology [47], 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 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 [97], and from the style guides [5, 86]. Caine [29] 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, which can be found in the supplementary materials, though it is too early to expect much adoption in the HCI community. We found no correlation 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.

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 section 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 [58] and stereotyping [22]. 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 is, 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 [97]. 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. This 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. 1, 1.2 - 1.4), and captured the words used to describe the genders (Fig. 1, 1.2.1), which allows us to 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 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. 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

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 1: Sample of the data schema developed for participant gender data in research publications.

was to add a data field to specify whether the author reported gender with a binary assumption (Fig. 1, 1.6). The full data dictionary is included in the supplementary materials.

4.3 Dataset

We choose to gather data from ACM SIGCHI [45]. 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 lead author for initial data exploration and 1,000 annotated half-and-half by the first two authors) 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 that we observed did indeed apply to the earlier years (1981 2000), 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 data fields outlined in the schema (Fig. 1). 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, Appendix Table 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 [29, p. 984]). We also classified papers by the gender words used in the paper (Gender language categorization, Appendix Table 2). We collected recruitment data verbosely, and used affinity diagrams [88] to find trends in reported recruiting practices. From this we created a classification code book, and categorized the papers into the recruitment classifications (Appendix Table 3, full code book in supplementary materials).

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 (584/1,147 papers, Appendix Table 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 was used under the assumption 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. It should also be noted that DER is confined to comparing the representation of men and women and cannot be used to analyse more complex gender representation.

The manual collection of data from PDFs was slow, taking 10 min per paper for our annotators, because finding data in PDFs has high

cognitive load. Also, the process was error prone as entering data into a form or spreadsheet often lead to mistakes. For faster and 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. As an annotation system MAGDA forces all data to come from text explicitly reported in the paper (G1). For data entries with unexpected data types and categories (e.g., "Gendered relationships") MAGDA offers free-form data entry that links back to the body text (e.g., "one housewife" [71], G2). To direct the annotator's attention to where gender data is likely to be found, we developed a machine learning system to highlight likely portions of the paper, and instructed annotators to focus on those sections (G4). Annotators found extraction time was reduced to 2.5 mins per paper with MAGDA. Claiming MAGDA as a systems contribution is beyond the scope of this paper, but using MAGDA helped us embrace the two core contributions during our data collection, the data schema and guidelines, and enabled the dataset contribution. For further information about the data collection tool, see the supplementary materials.

4.4 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, 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 [18] 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 [54]. We choose 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 [117]. Table 4 provides a list of the topics and the number of our 1,147 papers classified as each topic. Details of our analysis can be found in the supplementary materials. 10% of the data was annotated by both of the coders for calculating interrater reliability. The Cohen's Kappa statistic was over .6 for all data categories.

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. 2 (c), 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. 3), 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 [97, 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 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 [62] (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 for non-binary participants [30, 44, 64, 78].

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 [9, 29]. Unlike previous studies [29, p. 989], our investigation of the trend of the representation of women compared with men over time shows that while the *Distance from Even Representation of Men and Women* (DER) fluctuates around the average of -.15, the proportion of women participating in research does not appear to be increasing (Fig. 4). We applied linear regression to the data, and did not find a significant correlation between year and the DER of studies ($\beta = 0.002, t(582) = 0.41, p = 0.41$), indicating that there is no trend of 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

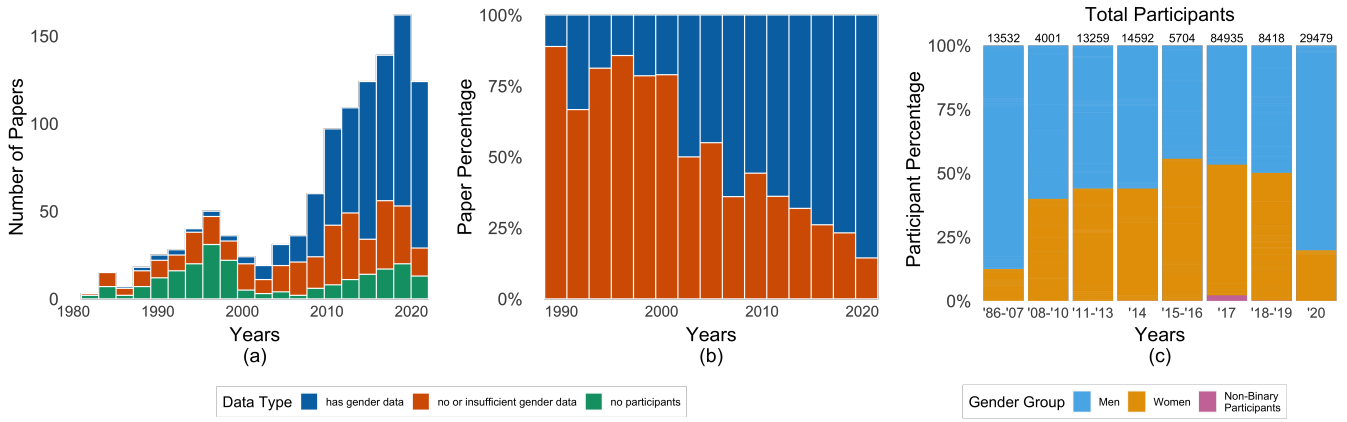

Figure 2: Gender reporting by year. (a) Number of papers published over the years of CHI, the number is higher after 2000. (b) The percent of papers with participants and gender reporting. Gender reporting is increasing. To smooth jagged trend, 1980-1990 grouped in first bar. (c) The total count of participants for the different year groups presented in proportion to each other. Non-binary participants are hardly visible.

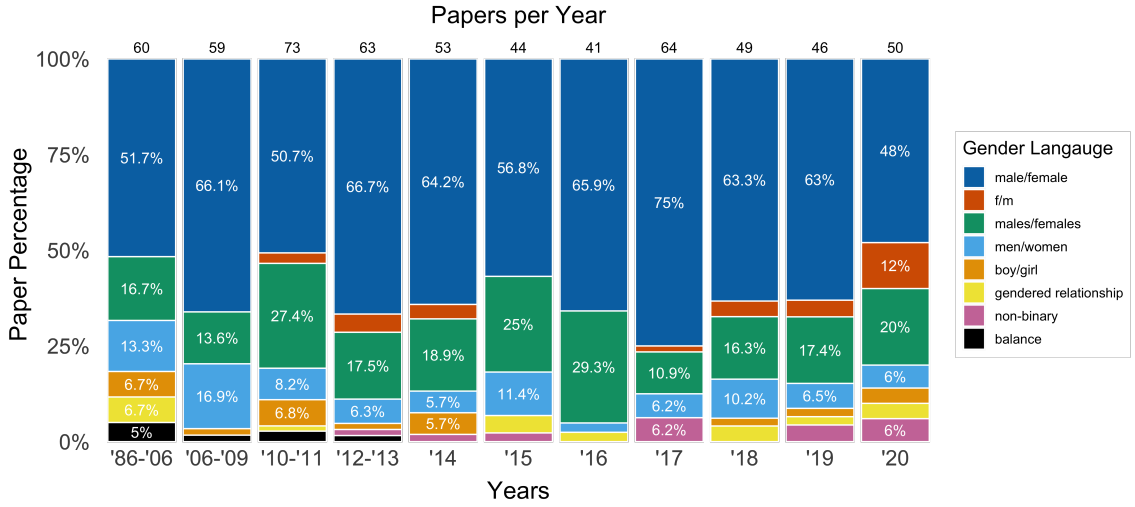

Figure 3: Percent of studies that fall into particular gender language categorization over the years of CHI. The use of "balance" language drops off as "non-binary" language appears, though the majority of reporting is still male(s)/female(s). A full breakdown of each categorization can been found in Table 2.

[13], the at other trans experiences of medical crowdfunding [51]. 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, $p = 0.24$).

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. 2 (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

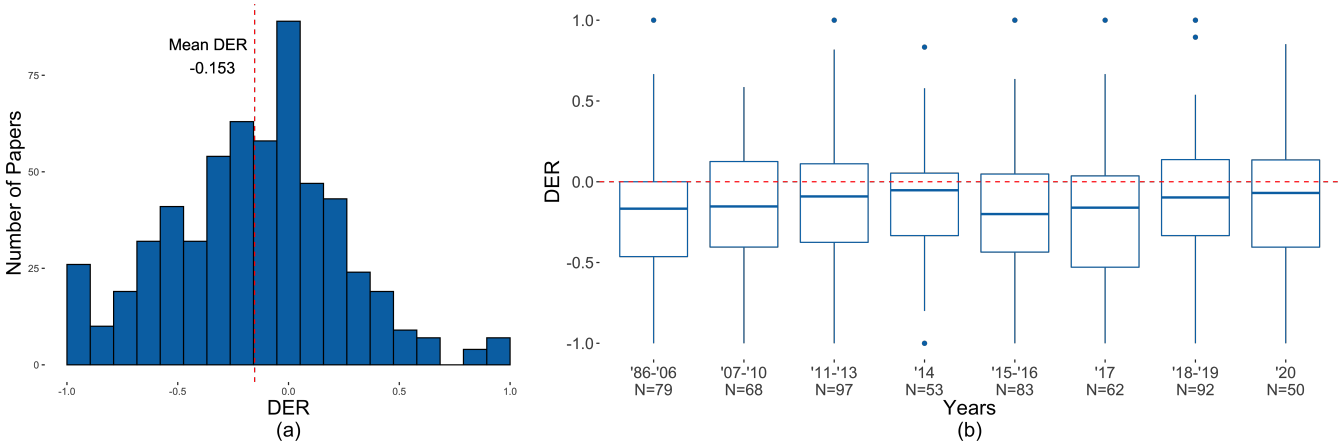

Figure 4: (a) The histogram of paper's Distance from Even Representation of Men and Women (DER), 584 papers have sufficient data to calculate DER. DER mean is -.153, meaning more men participate on average. (b) The chronological trend of DER. More men participate, and this trend is not changing.

language used in published papers (Fig. 3), 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) [105].

5.2 The gender bias in the participant source skews DER

Studies that recruit CS students are biased in favour of men and those that recruit psychology students in favour of women. Figure 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 [31]. 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 sufficiently ($adjustedR2 = .20, F(1, 32) = 9.03, p = .0051$). Study DER decreases as year increases (Fig. 5 (c)). We discuss the implications of the deterioration in section 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

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. 6 (a)), and topics that involve interactions situated in social or communicative contexts tended to have higher DER (the left side of Fig. 6 (a), Appendix Table 4). 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

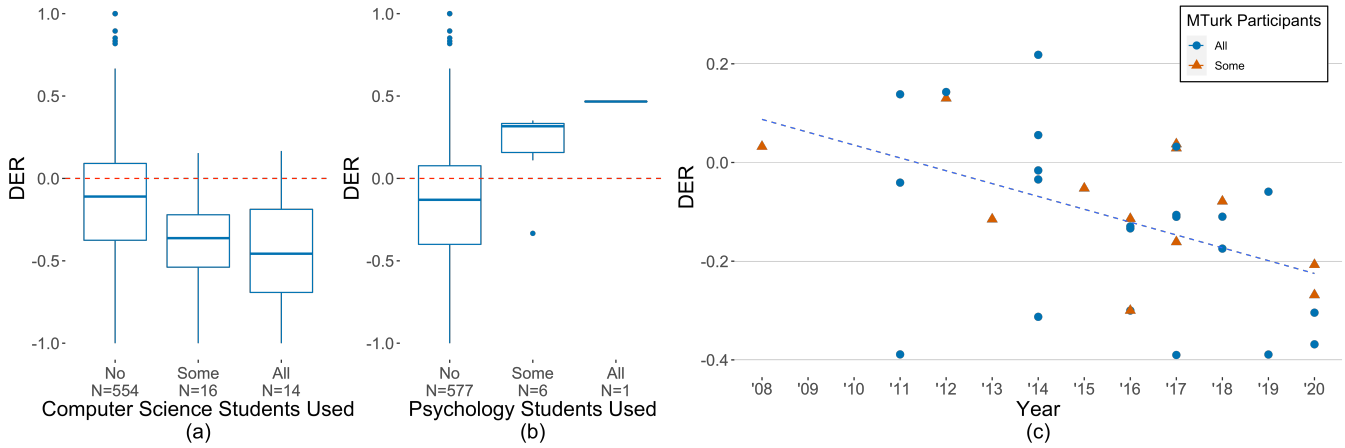

Figure 5: Participant recruitment. "All" means all participants belong to that classification, "Some" means only some of the participants belong to that classification. (a) Mean Distance from Even Representation of Men and Women (DER) for studies that use computer science (CS) students. The more CS students, the more men participate. (b) Mean DER for studies that use and psychology stduents. The more psychology students, the more women. (c) 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.

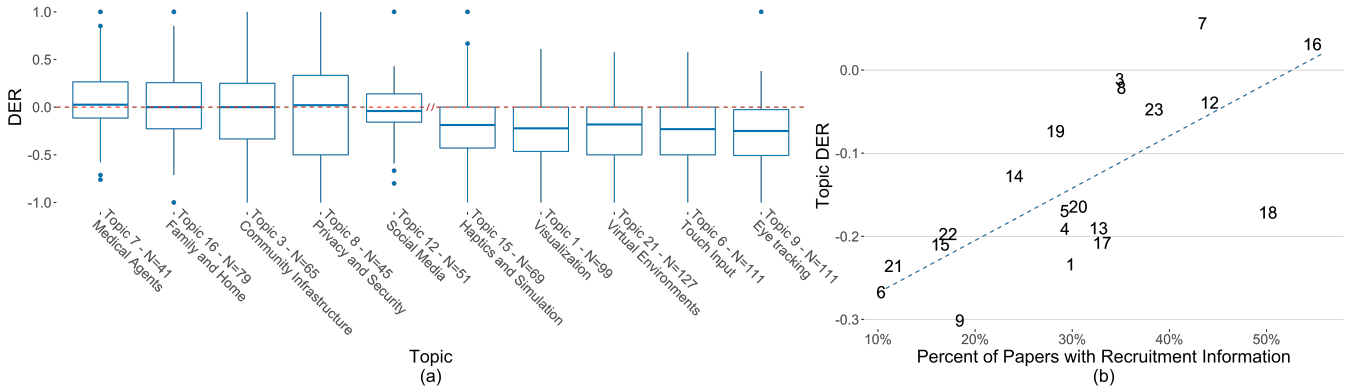

Figure 6: Comparison of topic to the *Distance from Even Representation of Men and Women* (DER). (a) The mean DER of papers classified in each topic, N = the number of papers classified. Showing the five topics with the highest mean DER, and the five with the lowest mean DER. Topics with high DER appear to be related to interaction in social and communicative contexts, 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.

of papers in a topic that reported participant recruitment information correlates with the proportion of women who participated in studies from that topic (Fig. 6 (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 ($adjusted R^2 = .50, F(1, 2355) = 2355, p < .001$).

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 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 (Section 4.1) and the data schema (Section 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 approach, 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 [6, 15]. 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 [26], 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." [108, p. 3]. Fully 30% of papers reported gender with an explicit binary assumption (179/584, Table 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 [30] 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 [1, 3]. Of the 1,147 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 [39]. 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" [87], 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 more on participants' lived experience [40], which could be the reason for the majority of studies shifting to including participants and reporting gender after 2000 (Fig. 2(a, b)). Methods also changed between the waves of HCI, shifting from quantitative to qualitative methods [40]. 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 method 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, the lack of gender reporting pre-2000 makes it difficult to do a data-driven analysis, so 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 [59, 94], 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 [94]. Previous research has extensively looked at how bias in machine learning algorithms can be traced back to biased datasets [25, 87], and as MTurk is often used to build datasets [90, 111], 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 such systems are used to gate gender restricted areas, like washrooms, this can have a hugely negative impact on a vulnerable population [70].

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" [34], and in one study, "female users tended to report feeling uncomfortable with putting the device on the chest" [34]. 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 [102]. 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 [97, 103], can serve as a foundation for this effort. Subcommittees that handle topics most prone to bias can raise awareness about this issues by questioning participant sources in publications, and inquiring whether the participants actually represent the population the results are meant to generalize to. 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 [97, 103].

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 [109] can remove barriers to equitable recruiting by ensuring funds are allocated for accommodations (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 [35, 52, 124], which is a positive sign that this attitude shift is taking place. We encourage researchers to make this inclusion standard practice.

7 CONCLUSION AND FUTURE WORK

We have provided empirical evidence for the under representation of women and non-binary participants in HCI research (RQ1).

Further, we have shown that recruiting is a key factor in gender representation, along with research area (RQ2). We have also presented gender data extraction guidelines and a participant gender data schema, instantiated in a system that has produced a structured and reliable set of gender data within HCI (RQ3). 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.

For future work, the relation of topic to DER suggests that method might be a factor, which previous studies have examined [29]. Despite the several ways in which author gender could impact participant recruiting [91, 122], we did not analyse author gender because of issues with collecting it [73]. Because we only collect things reported, if paper authors conflated sex and gender, then they will also be conflated in our data. Identifying this kind of conflation would require investigating gender analysis within publications, 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 participate in research. We therefore recommend a further qualitative investigation to prioritize non-binary perspectives.

ACKNOWLEDGMENTS

We would like to thank Austin Kobayashi for the work he did on building the MAGDA tool, and B. Watson for the early draft review. We also acknowledge the NSERC CGS-M, Discovery Grant, and Designing for People CREATE programs, as well as the generous gift from Adobe that funded this research.

REFERENCES

[1] Christopher Ahlberg and Ben Shneiderman. 1994. The Alphaslider: A Compact and Rapid Selector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 365–371. https://doi.org/10.1145/191666.191790

[2] Majed Al Zayer, Isayas B. Adhanom, Paul MacNeilage, and Eelke Folmer. 2019. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300584

[3] Florian Alt, Alireza Sahami Shirazi, Thomas Kubitza, and Albrecht Schmidt. 2013. Interaction Techniques for Creating and Exchanging Content with Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1709–1718. https://doi.org/10.1145/2470654.2466226

[4] Annalisa Anzani and Antonio Prunas. 2020. Sexual Fantasy of Cisgender and Nonbinary Individuals: A Quantitative Study. *Journal of Sex & Marital Therapy* 46, 8 (2020), 763–772. https://doi.org/10.1080/0092623X.2020.1814917 arXiv:https://doi.org/10.1080/0092623X.2020.1814917 PMID: 32880516.

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

[6] Bettina Bair and J. McGrath Cohoon. 2004. Special issue on gender-balancing computing education. *Journal on Educational Resources in Computing (JERIC)* 4, 1 (2004), 1–es.

[7] Sebastian Baltes and Stephan Diehl. 2016. Worse than spam: Issues in sampling software developers. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. Association for Computing Machinery, New York, NY, USA, 1–6.

[8] Shaowen Bardzell, Shad Gross, Jeffrey Wain, Austin Toombs, and Jeffrey Bardzell. 2011. The Significant Screwdriver: Care, Domestic Masculinity, and Interaction Design. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction* (Newcastle-upon-Tyne, United Kingdom) *(BCS-HCI '11)*. BCS Learning & Development Ltd., Swindon, GBR, 371–377.

[9] Louise Barkhuus and Jennifer A. Rode. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/1240624.2180963

[10] Laura Beckwith, Margaret Burnett, Valentina Grigoreanu, and Susan Wiedenbeck. 2006. Gender HCI: What about the software? *Computer* 39, 11 (2006), 97–101.

[11] Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. 2005. Effectiveness of End-User Debugging Software Features: Are There Gender Issues?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 869–878. https://doi.org/10.1145/1054972.1055094

[12] Annaliese K. Beery and Irving Zucker. 2011. Sex bias in neuroscience and biomedical research. *Neuroscience and biobehavioral reviews* 35, 3 (Jan 2011), 565–572. https://doi.org/10.1016/j.neubiorev.2010.07.002

[13] Rosanna Bellini, Simon Forrest, Nicole Westmarland, and Jan David Smeddinck. 2020. Mechanisms of Moral Responsibility: Rethinking Technologies for Domestic Violence Prevention Work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376693

[14] Cynthia L. Bennett and Os Keyes. 2020. What is the Point of Fairness? Disability, AI and the Complexity of Justice. *SIGACCESS Access. Comput.* 125, 5 (March 2020), 1. https://doi.org/10.1145/3386296.3386301

[15] Frank R. Bentley, Nediyana Daskalova, and Brooke White. 2017. 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* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1092–1099. https://doi.org/10.1145/3027063.3053335

[16] Nancy E. Betz and Louise F. Fitzgerald. 1987. *The career psychology of women*. Academic Press, San Diego, CA, US. xiii, 305–xiii, 305 pages.

[17] Vijay S Bhagat. 2018. Women authorship of scholarly publications in STEMM: Authorship puzzle. *Feminist Research* 2, 2 (2018), 66–76.

[18] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[19] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[20] Disha Bora, Hanlin Li, Sagar Salvi, and Erin Brady. 2017. ActVirtual: Making Public Activism Accessible. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 307–308. https://doi.org/10.1145/3132525.3134815

[21] Sophie H Bots, Floor Groepenhoff, Anouk LM Eikendal, Cara Tannenbaum, Paula A Rochon, Vera Regitz-Zagrosek, Virginia M Miller, Danielle Day, Folkert W Asselbergs, and Hester M den Ruijter. 2019. Adverse drug reactions to guideline-recommended heart failure drugs in women: a systematic review of the literature. *JACC: Heart Failure* 7, 3 (2019), 258–266.

[22] Adam Bradley, Cayley MacArthur, Mark Hancock, and Sheelagh Carpendale. 2015. Gendered or Neutral? Considering the Language of HCI. In *Proceedings of the 41st Graphics Interface Conference* (Halifax, Nova Scotia, Canada) *(GI '15)*. Canadian Information Processing Society, CAN, 163–170.

[23] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[24] Samantha Breslin and Bimlesh Wadhwa. 2014. Exploring Nuanced Gender Perspectives within the HCI Community. In *Proceedings of the India HCI 2014*

Conference on Human Computer Interaction (New Delhi, India) *(IndiaHCI '14)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/2676702.2676709

[25] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[26] M. Burnett, R. Counts, R. Lawrence, and H. Hanson. 2017. Gender HCI and microsoft: Highlights from a longitudinal study. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Raleigh, NC, USA, 139–143. https://doi.org/10.1109/VLHCC.2017.8103461

[27] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.

[28] Sabrina Burtscher and Katta Spiel. 2020. "But Where Would I Even Start?": Developing (Gender) Sensitivity in HCI Research and Practice. In *Proceedings of the Conference on Mensch Und Computer* (Magdeburg, Germany) *(MuC '20)*. Association for Computing Machinery, New York, NY, USA, 431–441. https://doi.org/10.1145/3404983.3405510

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

[30] Statistics Canada. 2020. Sex at birth and gender: Technical report on changes for the 2021 Census. https://www12.statcan.gc.ca/census-recensement/2021/ref/98-20-0002/982000022020002-eng.cfm Accessed: Sept 03, 2020.

[31] Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. 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 (2013), 2156–2160.

[32] Justine Cassell. 2002. *Genderizing Human-Computer Interaction*. L. Erlbaum Associates Inc., USA, 401–412.

[33] Stephen J Ceci and Wendy M Williams. 2011. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences* 108, 8 (2011), 3157–3162.

[34] Liwei Chan, Chi-Hao Hsieh, Yi-Ling Chen, Shuo Yang, Da-Yuan Huang, Rong-Hao Liang, and Bing-Yu Chen. 2015. Cyclops: Wearable and Single-Piece Full-Body Gesture Input Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3001–3009. https://doi.org/10.1145/2702123.2702464

[35] Irish Research Council. 2013. Irish Research Council (2013) Gender strategy and action plan 2013–20. http://research.ie/assets/uploads/2013/01/irish_research_council_gender_action_plan_2013_-2020.pdf Accessed: Nov 26, 2020.

[36] Rachel Croson and Uri Gneezy. 2009. Gender Differences in Preferences. *Journal of Economic Literature* 47, 2 (June 2009), 448–74. https://doi.org/10.1257/jel.47.2.448

[37] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[38] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT Press, USA.

[39] Emily Drabinski. 2018. Representing Normal: The Problem of the Unmarked in Library Organization Systems.

[40] Emanuel Felipe Duarte and M. Cecília C. Baranauskas. 2016. 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* (São Paulo, Brazil) *(IHC '16)*. Association for Computing Machinery, New York, NY, USA, Article 38, 4 pages. https://doi.org/10.1145/3033701.3033740

[41] Alice H Eagly. 1987. *Sex differences in social behavior: A social-role interpretation*. Psychology Press, New York.

[42] Alan Feingold and Ronald Mazzella. 1998. Gender differences in body image are increasing. *Psychological Science* 9, 3 (1998), 190–195.

[43] Casey Fiesler, Shannon Morrison, and Amy S. Bruckman. 2016. 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2574–2585. https://doi.org/10.1145/2858036.2858409

[44] AR Flores, JL Herman, GJ Gates, and TNT Brown. 2016. How Many Adults Identify as Transgender in the United States? Los Angeles, CA: The Williams Institute, UCLA School of Law.

[45] Association for Computing Machinery. 2020. Special Interest Group on Computer-Human Interaction. https://sigchi.org/ Accessed: Sept 09, 2020.

[46] Association for Computing Machinery. 2020. Ubiquitous Computing. https://ubicomp.org/ubicomp2020/ Accessed: Sept 09, 2020.

[47] Garth Fowler, C Cope, D Michalski, P Christidis, L Lin, and J Conroy. 2018. Women outnumber men in psychology graduate programs. *Monitor on Psychology* 49, 11 (2018), 21.

[48] Christopher Frauenberger, Julia Makhaeva, and Katta Spiel. 2016. Designing Smart Objects with Autistic Children: Four Design Exposès. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 130–139. https://doi.org/10.1145/2858036.2858050

[49] Vashti Galpin. 2002. Women in computing around the world. *ACM SIGCSE Bulletin* 34, 2 (2002), 94–100.

[50] Tripat Gill and Jing Lei. 2018. Counter-stereotypical products: Barriers to their adoption and strategies to overcome them. *Psychology & Marketing* 35, 7 (2018), 493–510.

[51] Amy Gonzales and Nicole Fritz. 2017. Prioritizing Flexibility and Intangibles: Medical Crowdfunding for Stigmatized Individuals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2371–2375. https://doi.org/10.1145/3025453.3025647

[52] Canadian Institutes of Health Research Government of Canada. 2019. Sex, Gender and Health Research. https://cihr-irsc.gc.ca/e/50833.html Accessed: Nov 26, 2020.

[53] Canadian Institutes of Health Research Government of Canada. 2020. What is gender? What is sex? https://cihr-irsc.gc.ca/e/48642.html Accessed: Aug 28, 2020.

[54] Shawn Graham, Scott Weingart, and Ian Milligan. 2012. *Getting started with topic modeling and MALLET*. Technical Report. The Editorial Board of the Programming Historian.

[55] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. https://doi.org/10.1145/1357054.1357074

[56] Georg Grön, Arthur P Wunderlich, Manfred Spitzer, Reinhard Tomczak, and Matthias W Riepe. 2000. Brain activation during human navigation: gender-different neural networks as substrate of performance. *Nature neuroscience* 3, 4 (2000), 404.

[57] Oliver L Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing" authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* 21, 6 (2016), 6791.

[58] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173582

[59] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V. Hanrahan, Jeffrey P. Bigham, and Chris Callison-Burch. 2019. 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* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312970

[60] Charles G. Hill, Maren Haag, Alannah Oleson, Chris Mendez, Nicola Marsden, Anita Sarma, and Margaret Burnett. 2017. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have It Both Ways?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6658–6671. https://doi.org/10.1145/3025453.3025609

[61] Anita Holdcroft. 2007. Gender bias in research: how does it affect evidence based medicine?

[62] Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. 2017. 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1882–1886. https://doi.org/10.1145/3025453.3025553

[63] Janet S. Hyde, Elizabeth Fennema, and Susan J. Lamon. 1990. Gender differences in mathematics performance: A meta-analysis. *Psychological bulletin* 107, 2 (1990), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

[64] Sandy James, Jody Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anafi. 2016. The report of the 2015 US transgender survey.

[65] Samantha Jaroszewski, Danielle Lottridge, Oliver L. Haimson, and Katie Quehl. 2018. "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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3173574.3173881

[66] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly* 44, 4 (1999), 741–763.

[67] Rikke Bjerg Jensen, Lizzie Coles-Kemp, and Reem Talhouk. 2020. 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376245

[68] Deborah G Johnson and Keith W Miller. 2002. Is diversity in computing a moral matter? *ACM SIGCSE Bulletin* 34, 2 (2002), 9–10.

[69] Gopinaath Kannabiran, Jeffrey Bardzell, and Shaowen Bardzell. 2011. 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* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 695–704. https://doi.org/10.1145/1978942.1979043

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

[71] Da-jung Kim and Youn-kyung Lim. 2019. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300714

[72] Sandjar Kozubaev, Fernando Rochaix, Carl DiSalvo, and Christopher A. Le Dantec. 2019. Spaces and Traces: Implications of Smart Technology in Public Housing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300669

[73] Stefan Krüger and Ben Hermann. 2019. Can an Online Service Predict Gender? On the State-of-the-Art in Gender Identification from Texts. In *Proceedings of the 2nd International Workshop on Gender Equality in Software Engineering (GE '19)*. IEEE Press, Montreal, Quebec, Canada, 13–16. https://doi.org/10.1109/GE.2019.00012

[74] Kathleen J Lehman, Linda J Sax, and Hilary B Zimmerman. 2016. Women planning to major in computer science: Who are they and what makes them unique? *Computer Science Education* 26, 4 (2016), 277–298.

[75] Marc J Lerchenmueller, Olav Sorenson, and Anupam B Jena. 2019. Gender differences in how scientists present the importance of their research: observational study. *BMJ* 367 (2019), l6573. https://doi.org/10.1136/bmj.l6573 arXiv:https://www.bmj.com/content/367/bmj.l6573.full.pdf

[76] Gilly Leshed, Theresa Velden, Oya Rieger, Blazej Kot, and Phoebe Sengers. 2008. In-Car Gps Navigation: Engagement with and Disengagement from the Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. https://doi.org/10.1145/1357054.1357316

[77] Yu-Cheng Lin. 2013. The Relationship between Touchscreen Sizes of Smartphones and Hand Dimensions. In *Universal Access in Human-Computer Interaction. Applications and Services for Quality of Life*, Constantine Stephanidis and Margherita Antona (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 643–650.

[78] Esther L Meerwijk and Jae M Sevelius. 2017. Transgender population size in the United States: a meta-regression of population-based probability samples. *American journal of public health* 107, 2 (2017), e1–e8.

[79] C. Mendez, A. Sarma, and M. Burnett. 2018. Gender in Open Source Software: What the Tools Tell. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*. IEEE, Gothenburg, Sweden, 21–24.

[80] Danaë Metaxa-Kakavouli, Kelly Wang, James A. Landay, and Jeff Hancock. 2018. Gender-Inclusive Design: Sense of Belonging and Bias in Web Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3174188

[81] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479.

[82] Justin Munafo, Meg Diedrick, and Thomas A Stoffregen. 2017. The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects. *Experimental brain research* 235, 3 (2017), 889–901.

[83] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.

[84] Mathias Wullum Nielsen, Jens Peter Andersen, Londa Schiebinger, and Jesper W Schneider. 2017. 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 (2017), 791–796.

[85] Tatsuya Nomura. 2017. Robots and gender. *Gender and the Genome* 1, 1 (2017), 18–25.

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

[87] Caroline Criado Perez. 2019. *Invisible women: Exposing data bias in a world designed for men.* Random House, New York City, United States.

[88] Craig Plain. 2007. Build an affinity for KJ method. *Quality Progress* 40, 3 (2007), 88.

[89] Kaitlin C. Rasmussen, Erin Maier, Beck E. Strauss, Meredith Durbin, Luc Riesbeck, Aislynn Wallach, Vic Zamloot, and Allison Erena. 2019. The Nonbinary Fraction: Looking Towards the Future of Gender Equity in Astronomy. arXiv:1907.04893 [astro-ph.IM]

[90] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (2017), 1–12.

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

[92] Kathryn E. Ringland, Christine T. Wolf, Heather Faucett, Lynn Dombrowski, and Gillian R. Hayes. 2016. "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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1256–1269. https://doi.org/10.1145/2858036.2858038

[93] Paula A Rochon, Jocalyn P Clark, Malcolm A Binns, Vishal Patel, and Jerry H Gurwitz. 1998. Reporting of gender-related information in clinical trials of drug therapy for myocardial infarction. *Cmaj* 159, 4 (1998), 321–327.

[94] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

[95] María Teresa Ruiz-Cantero, Carmen Vives-Cases, Lucía Artazcoz, Ana Delgado, Maria del Mar García Calvente, Consuelo Miqueo, Isabel Montero, Rocío Ortiz, Elena Ronda, Isabel Ruiz, et al. 2007. A framework to analyse gender bias in epidemiological research. *Journal of Epidemiology & Community Health* 61, Suppl 2 (2007), ii46–ii53.

[96] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 144 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359246

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

[98] Londa Schiebinger and Martina Schraudner. 2011. Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering1. *Interdisciplinary Science Reviews* 36, 2 (2011), 154–167.

[99] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5412–5427. https://doi.org/10.1145/3025453.3025766

[100] Brenda Scott and Vincent Conzola. 1997. Designing Touch Screen Numeric Keypads: Effects of Finger Size, Key Size, and Key Spacing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 41, 1 (1997), 360–364. https://doi.org/10.1177/107118139704100180

[101] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology* 71, 11 (2020), 1281–1294. https://doi.org/10.1002/asi.24335 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24335

[102] Luciano P. Soares, Leonardo Nomura, Marcio C. Cabral, Mario Nagamura, Roseli D. Lopes, and Marcelo K. Zuffo. 2005. Virtual Hang-Gliding over Rio de Janeiro. In *ACM SIGGRAPH 2005 Emerging Technologies* (Los Angeles, California) *(SIGGRAPH '05)*. Association for Computing Machinery, New York, NY, USA, 29–es. https://doi.org/10.1145/1187297.1187327

[103] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for HCI researchers. *interactions* 26, 4 (2019), 62–65.

[104] Katta Spiel, Os Keyes, and Pınar Barlas. 2019. Patching Gender: Non-Binary Utopias in HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310425

[105] Katta Spiel, Os Keyes, Ashley Marie Walker, Michael A. DeVito, Jeremy Birnholtz, Emeline Brulé, Ann Light, Pınar Barlas, Jean Hardy, Alex Ahmed, Jennifer A. Rode, Jed R. Brubaker, and Gopinaath Kannabiran. 2019. Queer(Ing) HCI: Moving Forward in Theory and Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.

org/10.1145/3290607.3311750
[106] Angelika Strohmayer, Rob Comber, and Madeline Balaam. 2015. Exploring Learning Ecologies among People Experiencing Homelessness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2275–2284. https://doi.org/10.1145/2702123.2702157
[107] Angelika Strohmayer, Mary Laing, and Rob Comber. 2017. 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3352–3364. https://doi.org/10.1145/3025453.3025615
[108] Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2019. Gender-Inclusive HCI Research and Design: A Conceptual Review. *Now Foundations and Trends® in Human–Computer Interaction* 13, 1 (2019), 1–69.
[109] Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. 2019. Sex and gender analysis improves science and engineering. *Nature* 575, 7781 (2019), 137–146.
[110] Divy Thakkar, Nithya Sambasivan, Purva Kulkarni, Pratap Kalenahalli Sudarshan, and Kentaro Toyama. 2018. 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173926
[111] Anh Truong, Sara Chen, Ersin Yumer, David Salesin, and Wilmot Li. 2018. Extracting Regular FOV Shots from 360 Event Footage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173890
[112] Sherry Turkle. 2004. Computational reticence: Why women fear the intimate machine. In *Technology and women's voices*. Routledge, Milton Park, Abingdon, Oxfordshire, 44–60.
[113] Sarah Theres Völkel, Wiktoria Wilkowska, and Martina Ziefle. 2018. Gender-Specific Motivation and Expectations toward Computer Science. In *Proceedings of the 4th Conference on Gender & IT* (Heilbronn, Germany) *(GenderIT '18)*. Association for Computing Machinery, New York, NY, USA, 123–134. https://doi.org/10.1145/3196839.3196858
[114] Carl L Von Baeyer, Debbie L Sherk, and Mark P Zanna. 1981. Impression management in the job interview: When the female applicant meets the male (chauvinist) interviewer. *Personality and Social Psychology Bulletin* 7, 1 (1981), 45–51.
[115] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *International AAAI Conference on Weblogs and Social Media* 9 (01 2015), 454–463.
[116] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education.
[117] Mika Westerlund, Seppo Leminen, and Mervi Rajahonka. 2018. A Topic Modelling Analysis of Living Labs Research. *Technology Innovation Management Review* 8 (07/2018 2018), 40–51. https://doi.org/10.22215/timreview/1170
[118] Michael W Wiederman. 1999. Volunteer bias in sexuality research using college student participants. *Journal of Sex Research* 36, 1 (1999), 59–66.
[119] Holly O Witteman, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 2019. Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet* 393, 10171 (2019), 531–540.
[120] Erik Won, Pete Johnson, Laura Punnett, Theodore Becker, and Jack Dennerlein. 2003. Gender Differences in Exposure to Physical Risk Factors during Standardized Computer Tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47, 10 (2003), 1155–1158. https://doi.org/10.1177/154193120304701011
[121] Sarita Yardi and Amy Bruckman. 2012. Income, Race, and Class: Exploring Socioeconomic Differences in Family Technology Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 3041–3050. https://doi.org/10.1145/2207676.2208716
[122] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye Text, Hello Emoji: Mobile Communication on WeChat in China. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 748–759. https://doi.org/10.1145/3025453.3025800
[123] Martina Ziefle and Anne Kathrin Schaar. 2011. Gender differences in acceptance and attitudes towards an invasive medical stent. *Electronic Journal of Health Informatics* 6, 2 (2011), 13.
[124] Fonds zur Förderung der wissenschaftlichen Forschung. 2019. FIX the Knowledge. https://www.fwf.ac.at/en/about-the-fwf/gender-issues/fix-the-knowledge Accessed: Nov 26, 2020.

A DATASET TABLES

A.1 Gender language coverage table

Table 1 provides the full classification schema for gender reporting coverage and how many papers of those we analyzed fall into each coverage classification.

A.2 Gender language categorization table

Table 2 provides the complete list of gender language categorizations and the number of papers which were assigned each categorization. In addition, all the gender words which appeared in the papers we analysed are listed. For the small amount of papers (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 2.

A.3 Recruitment classification table

Table 3 details all the recruitment categorizations we used, and briefly describes each categorization. A study would be labeled as "All" if it only reported participants that fell into the particular classification, and "No" if it did not report participants in that classification. "Some" means that the study reported participants in the classification, but they were not the only participants. "Yes" means that at least some of the participants belonged to that recruitment classification, but there were not enough studies to separate them into All and Some. The full code book is available in the supplementary material.

A.4 Topic classification table

Table 4 lists all 25 topics that were used to classify papers, and how many of the 1,147 papers fell into each topic. Papers can belong to multiple topics.

Table 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 2: Gender language categorization

Category	Papers classified	Words included
Non-binary	12	gender queer, nonbinary, non-gender-identifying individual, gender-fluid, transmen, 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

Table 3: Participant Recruitment Classifications

Recruitment Classification	Levels	Classification Description
Computer Science Students	All, Some, No	University students studying computer science
Psychology Students	All, Some, No	University students studying Psychology
Children	All, Some, No	Participants under 16 or highschool students.
Patients and Participants with Illnesses	Yes, No	Participants that are described as patients or as having an illness or disease.
Blind Participants	Yes, No	Participants described as blind or with visual impairments.
Amazon Mechanical Turk	All, Some, No	Participants recruited via Amazon's Mechanical Turk
Less-rigorous sampling	Yes, No	Less-rigorous sampling includes specifically "snowball sampling", "word of mouth", "convenience sampling", or "purposeful sampling".
Participant Pool	Yes, No	Studies which drew from a set of people designated as potential research participants (e.g. psychology student recruitment pool, participant recruitment mailing list)

Table 4: Topic table sorted by mean DER

Topic Number	Topic 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