AN ABSTRACT OF THE DISSERTATION OF

Abrar Fallatah for the degree of Doctor of Philosophy in Computer Science presented on June 26, 2023.

Title: Inclusive Software Design and Its Methods for Beyond-WIMP User Interfaces

Abstract approved: ______________________________

Margaret Burnett

While digital inclusivity researchers and software practitioners have been trying to address exclusion biases in Windows, Icons, Menus, and Pointers (WIMP) user interfaces (UIs) for a long time, little has been done to investigate if and how inclusive software design and its methods that have been devised for WIMP UIs can be used effectively to design and evaluate beyond-WIMP UIs. To that end, this dissertation investigated the use of inclusive software design on a selection of different beyond-WIMP UIs at 3 stages. In Chapter 3, we explored the applicability of inclusive software design as an evaluation approach with a social robot that interacts with diverse people in diverse social places. In Chapter 4, we examined using a particular inclusive software design method to evaluate and redesign a multiple robots controller. Finally, in Chapter 5, we investigated whether and how a family of inclusive software design methods can be used analytically to evaluate the multidimensionality of a Hands-Free Integrated Development Environment (IDE). This work contributes design implications, new technology development, and empirical contributions to designing beyond-WIMP UIs for diverse humans.

Inclusive Software Design and Its Methods for Beyond-WIMP User Interfaces

by

Abrar Fallatah

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 26, 2023
Commencement June 2024

Doctor of Philosophy dissertation of Abrar Fallatah presented on June 26, 2023.

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Abrar Fallatah, Author

ACKNOWLEDGEMENTS

I wish to extend my most profound appreciation to the remarkable intellects behind this outstanding work and those who have accompanied and supported me throughout this transformative journey.

First and foremost, I must express my heartfelt gratitude to *Margaret Burnett* (affectionately known as Peggy), my esteemed Ph.D. advisor, whose unwavering belief in my abilities surpassed my own during moments of self-doubt. Furthermore, I extend my profound thanks to the esteemed members of my committee—*Naomi Fitter, Christopher Sanchez, Parsad Tadepall,* and *Geoff Hollinger*—who not only bore witness to my academic progress but also provided invaluable guidance and mentorship throughout the arduous expedition.

I am also indebted to my esteemed collaborators, who diligently aided in translating my conceptualizations into scholarly journals. Professors *Heather Knight, Bohkyung Chun, Martin Erwig, Christopher Bogart, and Anita Sarma*, and a distinguished ensemble including *Samarendra Hedaoo, Sogol Balali, Brett Shawn Stoddard, Yao-Lin Tsai, Jiaji Sun, Jillian Renee Emard, Denisse A Alvarado, Brett Altena, Jeremy Urann, Janani Swaminathan, Alexandra Bacula, Abhijeet Agnihotri, Brian Layng*, together with the exceptional contributions of *Fatima A Moussaoui, Md Montaser Hamid, and Chimdi Chikezie*, have significantly enriched and elevated my work.

Moreover, I wish to convey my immeasurable gratitude to my devoted family, cherished friends, and esteemed mentors who have played pivotal roles in shaping my path. Foremost among them is my mother, *Khadijah Haji*, whose unwavering emotional and financial support has been a cornerstone of my journey, always nurturing my aspirations and instilling the belief that no dream is beyond reach. I must also acknowledge the transformative influence of my brother, *Eyad Fallatah*, who exemplifies that life extends far beyond our humble hometown's confines, inspiring an insatiable curiosity and an unwavering drive to explore uncharted territories. Equally deserving of recognition is my brother, *Anwar Fallatah*, whose sage counsel and mentorship have endowed me with the invaluable qualities of independence, enabling me to navigate the world with confidence and fearlessness. I am also grateful for my brother *Omar Fallath*, whose positive impact on my life cannot be understated, and my sister, *Amjad Fallatah*, who has been

an unwavering source of strength, fortifying my resilience and teaching me the virtues of determination and assertiveness. I am also profoundly grateful for the unwavering support of my closest confidante and best friend, *Samia Aldhahi*, whose steadfast presence and unwavering support have seen me through countless emotional tribulations.

Lastly, I must express my gratitude to *Prof. Emad Aboelela*, my esteemed advisor during my Bachelor's degree studies, whose professionally crafted and genuinely heartfelt recommendation letter opened doors and paved the way for my future endeavors, and *Prof. Iman Alanasri*, a luminary at Taibah University, whose unwavering belief in my potential and constant encouragement throughout my academic odyssey encouraged me to pursue a Ph.D. with unwavering determination.

To every individual mentioned above, I owe a debt of gratitude beyond words for their unwavering support, guidance, and belief in my capabilities, without which this remarkable journey would not have been possible.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

Chapter 1: General Introduction

Exclusion bias is a substantial problem because humans gain and lose physical and cognitive abilities throughout their lifetime through illnesses, injuries, social identities, and even cultural backgrounds. Thus, eventually, all humans are excluded by designs that do not fit their ever-changing physical and cognitive needs. This exclusion by design is not limited to keyboard-plus-mouse User Interfaces (UIs) and extends to all ubiquitous computing systems that inherit the unconscious biases of the creators behind them [38].

While human-computer interaction researchers and practitioners have been adopting inclusive design to weed out exclusion by design from keyboard-plus-mouse UIs and its dominating WIMP (Windows, Icons, Menus, and Pointers) interaction modalities, little has been done in the field of ubiquitous computing systems that are beyond-WIMP. Beyond-WIMP UIs, such as robots that uses motion and audio, are becoming a reality and replacing WIMP UIs. This dissertation investigates the following thesis: *Inclusive software design and its methods that have been devised for WIMP UIs can be used effectively to design and evaluate beyond-WIMP UIs.* To that end, this thesis contains a three-stage investigation of inclusive software design for a selection of three beyond-WIMP UIs: a social robot, a multiple robots controller, and a Hands-Free Integrated Development Environment (IDE). The first stage (Chapter 3) explored the applicability of inclusive software design as an evaluation approach with a social robot that interacts with diverse people in diverse social places. The second stage (Chapter 4) examined the use of a particular inclusive software design method to evaluate and redesign a multiple robots controller. In this stage, the method was used twice: empirically in a lab study and then analytically with design experts. Finally, stage 3 (Chapter 5) investigated whether and how a family of inclusive software design methods can be used analytically to evaluate the multidimensionality of a Hands-Free IDE.,

This dissertation makes several design implications (Chapter 3), new technology development (Chapter 4), and empirical contributions (Chapter 3, Chapter 4, and Chapter 5) to designing beyond-WIMP UIs for diverse humans. Specifically, the relevancy of inclusivity to beyond-WIMP UIs and whether and how inclusivity methods can be used empirically and analytically to design and evaluate inclusive beyond-WIMP UIs.

Figure 1.1 summarizes the thesis statement and its three stages, each of which we discussed in the previous chapters.

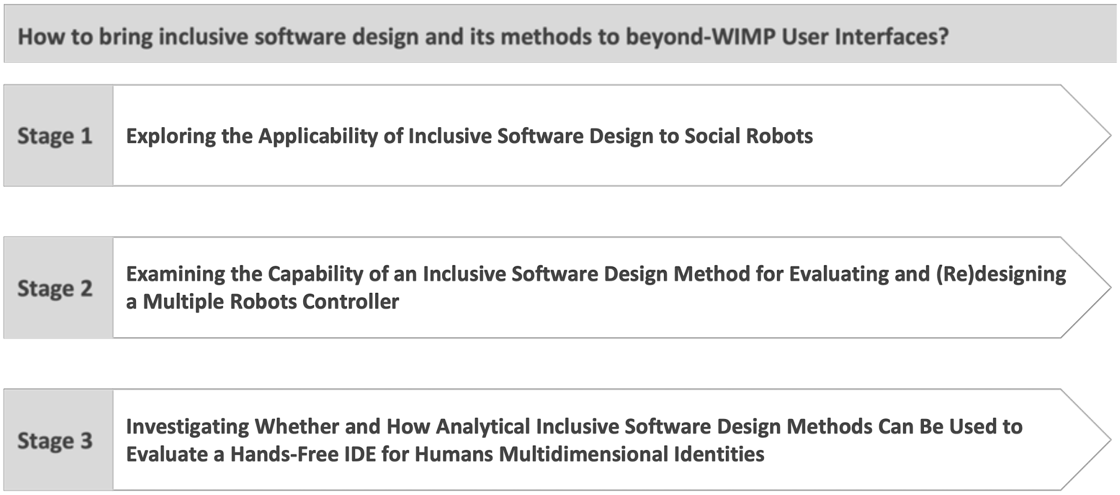

How to bring inclusive software design and its methods to beyond-WIMP User Interfaces?
Stage 1
Exploring the Applicability of Inclusive Software Design to Social Robots
Stage 2
Examining the Capability of an Inclusive Software Design Method for Evaluating and (Re)designing a Multiple Robots Controller
Stage 3
Investigating Whether and How Analytical Inclusive Software Design Methods Can Be Used to Evaluate a Hands-Free IDE for Humans Multidimensional Identities

Figure 1.1: Summary of the dissertation goal and stages.

Chapter 2: Overall Background and Related Work

The feasibility of this dissertation rests on weeding out *Digital Exclusion Biases* (aka *Exclusion by Design*), which happens when designers instinctively build solutions for users like them, either unaware of the needs of users with different cognitive and physical attributes or do not know how to accommodate others' needs into the design cycle [41, 38, 59]. These *Disabled Designs* not only stop users from accomplishing their tasks but also decide where humans belong and where they are outsiders, shape humans' sense of value, and what they can contribute [43, 31, 19, 82, 81]. Exclusion and the social rejection that often accompanies interacting with disabled designs are human experiences that inclusivity researchers and practitioners aim to eliminate by using *Inclusive Design* [38, 19].

This chapter provides background on *Inclusive Design* and its evaluation methods that have been devised for WIMP and beyond-WIMP UIs. This chapter also covers the related work that applies to thesis statement. Additional related work specific to a research stage is covered throughout this document in the context of that stage.

2.1 What is Inclusive Design?

Inclusive Design is a set of approaches designers embrace to ensure that buildings, public spaces, services, products, and technology address the needs of the broadest possible audience [15, 55, 14, 38]. The essence of inclusive design is understanding and enabling often excluded users, such as minorities and older adults, to be in the mainstream of everyday life rather than users who succeed only with particular attention or special design solutions [15, 41]. This shift from special solutions and assistive gadgets toward inclusivity was first introduced to Human-Computer Interaction (HCI) in 1994 by Roger Coleman to enable independence for older adults and people with physical disabilities [14]. Ever since, inclusive design has been growing to include additional *Diversity Dimensions*, such as gender, socioeconomic status (SES), race, ethnicity, income, language, education, and more [15, 55, 14].

Although researchers have been adopting several human-centered design approaches to address inclusivity, the prominent approach to inclusive beyond-WIMP design focuses on developing guidelines for a specific diversity dimension(s) based on an intensive literature review. For example in the automotive domain, Roundtree et al. reviewed the needs of various road users (i.e. children, adults, older adults, and individuals with visual, auditory, and cognitive impairments) to apprise inclusive design guidelines for an external human-vehicle interface [69]. Roundtree et al. reported that while the average reaction time for road users in crosswalks is approximately one second, external human-vehicle interfaces should take into consideration that children pedestrians' reflexes are not trained for emergency scenarios and that their field of vision, which is one-third narrower than adults, hinders their ability to judge speed and distance.

Here Roundtree et al. approached inclusivity from human factors and ergonomics perspective, but this thesis focuses on Inclusive Software Design that supports the human's cognitive problem-solving diversity. For example of inclusive software design in a WIMP interface was Vorvoreanu et al. summative evaluation of an inclusive software design method that finds and fixes the gender biases of a search engine [84]. Vorvoreanu et al. used a set of problem-solving factors, such as self-efficacy and information processing styles, to identify biased UI elements and patterns and fix them. This thesis use similar inclusive software design methods to find and fix cognitive problem-solving biases of two beyond-WIMP user interfaces.

2.2 Evaluation Methods for Interaction Design

Evaluation methods for interaction design fall into two categories: *Analytical or Empirical.*

Analytical Evaluation Methods (aka discount evaluation) rely on using a set of design principles to systematically evaluate a design without users and with the help of a small set of User eXperience (UX) expert evaluators [56, 14]. Examples of analytical evaluations methods include cognitive walkthroughs [86, 46], heuristic evaluations [57], expert reviews [51], and applying HCI guidelines to a design [5, 49]. The *Cognitive Walkthrough* for example is an action-based systematic approach to evaluate whether the order of cues of a use-case reflect the way people cognitively process tasks and anticipate next-steps of a platform [61, 86, 56]. For example, Wharton's cognitive walkthrough uses a set of

4 questions to systematically evaluate the actions needed to complete a use-case(s) [86]. The cognitive walkthrough evaluators answer these 4 questions for each action of a use-case in sequence and justify each answer based on their HCI expertise and understanding of the users. In a similar way, *Heuristic Evaluation* uses a set of agreed upon best practices or usability “rules of thumb” to evaluate a design [83]. For example, Nielsen’s 10 heuristics are used as principles to uncover usability problems in a design and their severity [57]. The heuristic evaluation evaluators check design compliance with each heuristic independently before aggregating their findings [47]. The evaluation methods we are using in Chapter 4 and Chapter 5 rely on these analytical evaluation methods.

Unlike analytical evaluation methods, *Empirical Evaluation Methods* rely on data that users generate, such as their feedback on a UI, behaviors and interaction patterns with an interface, artifact they create, etc. Empirical data is obtained by several means such as ethnography [21, 54], usability studies [78], controlled lab studies, interviews, contextual inquiries, and observations. *Ethnography*, the method we are using in Chapter 3, investigates the contextual construction of social behaviors and perceptions that are linked to local cultures [33]. Ethnography is known as being relevant in HCI studies as it helps increase the social and contextual investigation of and around computational systems [21, 25]. An example of using ethnography as an evaluation design method in a hospital units uncovered how patient profiles, workflow, and social dynamics dramatically influenced the staff’s interaction with an autonomous delivery robot [54]. Although the post-partum units integrated the robot into their workflow and social context, the low tolerance for interruptions in the medical units caused the robot to have a negative impact on the workflow and staff resistance. *User studies* on the other hand are widely known for focusing on users and their tasks while seeking evidence about how to improve the usability of an interface [32]. User studies, which we are using in Chapter 4, are sometimes referred to as usability testing [40] and are usually performed on controlled and staged settings [78, 52]. Showkat et al., for example, ran a 12-participants users study to identify gender differences when Tele-operating the PR2 humanoid robot (a mobile robot with 2 arms) to manipulate four objects [77]. Showkat et al. revealed that male participants were more confident and tinkered more with the robot than females which resulted in greater task success and lower task completion time for males.

2.3 InclusiveMag and GenderMag

InclusiveMag [50] is a meta-method that enables HCI researchers to generate systematic analytical design methods for a given diversity dimension (Figure 2.1). The generated methods are framed along *facet types* and *facet values*. Designers and other software practitioners can then use the InclusiveMag-generated methods to analytically evaluate user experiences from the perspective of users across the diversity dimensions.

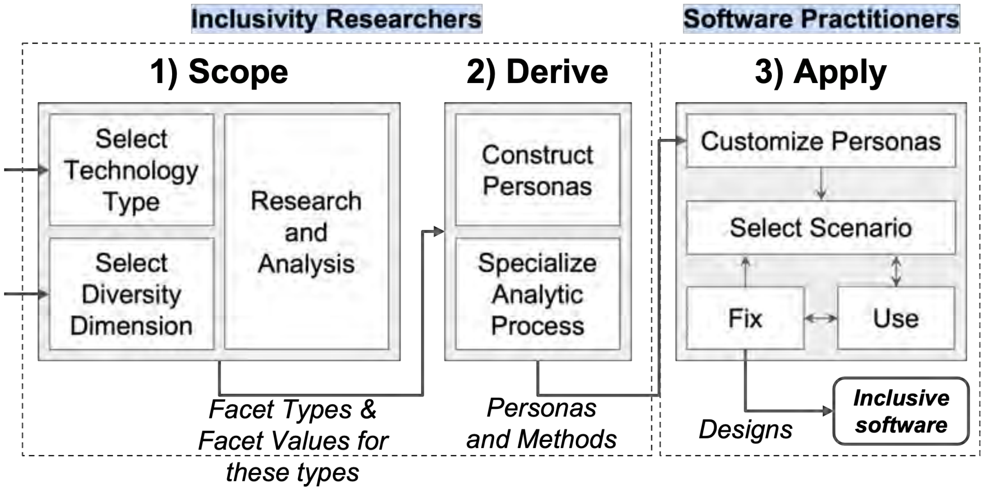
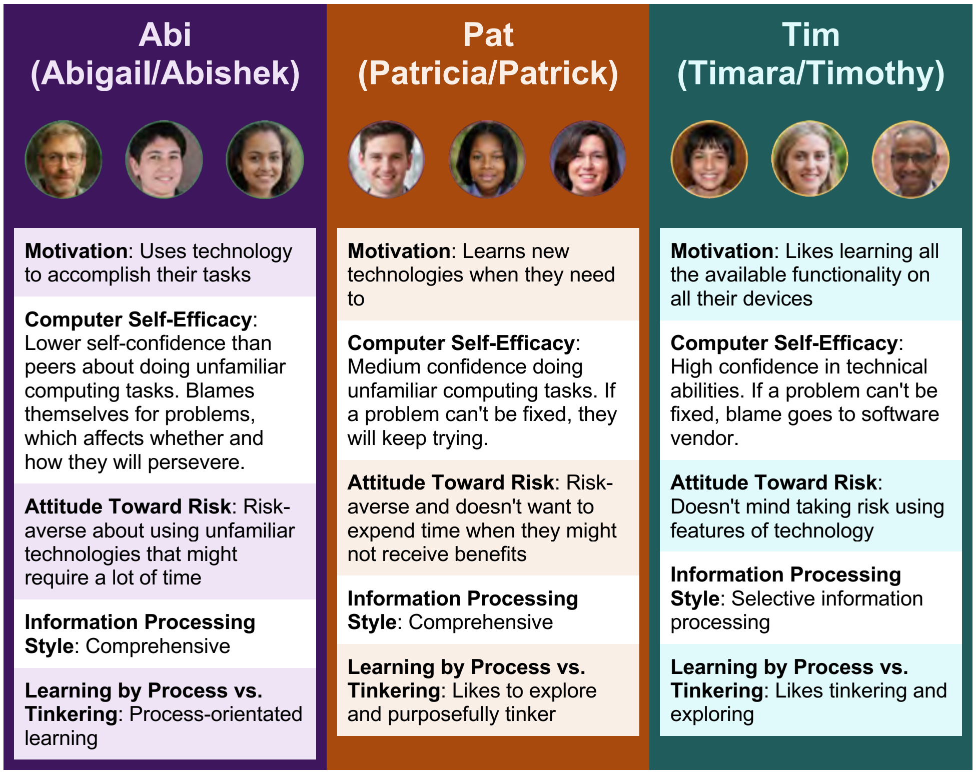


Figure 2.1: The InclusiveMag meta-method has three steps. The first and second steps output facet types and the range of possible values that each facet type can have framed along personas and analytical design methods. These facet types and values provide the starting point for the inclusivity design methods in Chapter 4 and Chapter 5.

For example, InclusiveMag was used to generate GenderMag, a systematic analytical method for the diversity dimension of gender [9, 50]. HCI practitioners have used the GenderMag method to find, avoid, and fix inclusivity issues in a variety of domains, such as education software [9, 35, 76, 17], machine learning aids [10], office productivity software [37], open source project sites [12, 24, 58], robotics [77, 3], software tools [29], and search interfaces [84]. Other offspring of InclusiveMag include SESMag to support users in diverse socioeconomic situations [39], AgeMag to evaluate age bias in e-commerce applications [48], and a collection of eight pilot InclusiveMag-generated methods to support eight diversity dimensions (e.g., eyesight, attention span, position along the autism spectrum) [50].

Figure 2.1 gives an overview of InclusiveMag with its 3 steps: Scope, Derive, and Apply. The first step, Scope, produces a set of facet types for the *diversity dimension* of interest (e.g., gender, in the GenderMag example). These *facet types* are traits for which individuals at opposite ends of the diversity dimension can differ significantly from each other. For example, the GenderMag personas in Figure 2.2 have GenderMag's five facet types and some of the possible facet values different individuals might have. This thesis use GenderMag's personas in Chapter 4 and Chapter 5.



Abi (Abigail/Abishek)	Pat (Patricia/Patrick)	Tim (Timara/Timothy)
Motivation: Uses technology to accomplish their tasks	**Motivation**: Learns new technologies when they need to	**Motivation**: Likes learning all the available functionality on all their devices
Computer Self-Efficacy: Lower self-confidence than peers about doing unfamiliar computing tasks. Blames themselves for problems, which affects whether and how they will persevere.	**Computer Self-Efficacy**: Medium confidence doing unfamiliar computing tasks. If a problem can't be fixed, they will keep trying.	**Computer Self-Efficacy**: High confidence in technical abilities. If a problem can't be fixed, blame goes to software vendor.
Attitude Toward Risk: Risk-averse about using unfamiliar technologies that might require a lot of time	**Attitude Toward Risk**: Risk-averse and doesn't want to expend time when they might not receive benefits	**Attitude Toward Risk**: Doesn't mind taking risk using features of technology
Information Processing Style: Comprehensive	**Information Processing Style**: Comprehensive	**Information Processing Style**: Selective information processing
Learning by Process vs. Tinkering: Process-orientated learning	**Learning by Process vs. Tinkering**: Likes to explore and purposefully tinker	**Learning by Process vs. Tinkering**: Likes tinkering and exploring

Figure 2.2: Portions of GenderMag's three personas—Abi (left), Pat (middle), and Tim (right)—as customized by a faculty member who was applying GenderMag to college-level students [44], with each persona's facet value for the GenderMag facet types. In Chapter 4 we used the Abi persona only, while we used both Abi and Tim personas in Chapter 5. More than two personas is useful in emphasizing to other humans the diversity of the target population.

In InclusiveMag's second step, Derive, inclusivity researchers use the facet types they created in the Scope step to derive mechanisms for HCI practitioners to use when designing/evaluating a system's inclusivity, such as Figure 2.2's personas. The researchers also specialize an existing analytic method, such as a cognitive walkthrough or set of design heuristics, using the facet types.

Finally, in the third step, Apply, software practitioners customize and apply the generated method(s) or other facet-based artifacts (e.g., personas) to evaluate their technology to increase its inclusivity across that diversity dimension. For example, if in the Derive step the HCI researchers chose to specialize a cognitive walkthrough for the analytic process, then in the Apply step the practitioners will be conducting this specialized cognitive walkthrough. The fifth chapter of this thesis, Chapter 5, shows how these practitioners can compose one or more of these methods to evaluate a free-hand IDE across multiple intersecting diversity dimensions.

Chapter 3: Exploring the Applicability of Inclusive Software Design to Social Robots

The first stage to defend or reject the thesis was exploring the applicability of inclusive software design to evaluate the inclusivity of a beyond-WIMP UI. The beyond-WIMP UI in this stage was a robotic chair (ChairBot) that asked bystanders to help it order food items from six various cafes around a University Campus. In this stage, we studied the inclusivity requirements in terms of culture and situational factors that predicted participants' likelihood to help and care for the robot. A chi-square test revealed that cultural and situational factors such as the cafe's overall mood (social atmosphere) and the robot's motion characteristics (approaching styles) predicted participants' likelihood to help and care for the robot. This stage showed that software inclusivity is relevant to evaluating beyond-WIMP UIs that use motion only to communicate with humans. This stage also showed that robotics engineer must consider the cultural and situational factors at which the interaction is taking place to develop inclusive help-seeking robots.

3.1 Introduction

Prior work established that as robots leave the lab and join our daily life in numerous forms providing countless services, these robots are yet still limited and can benefit from human help to overcome their actuation limitations and exceptional situations [1, 2]. For example, mobile delivery robots must stay idle until a human picks up the packages [11] or push elevator buttons to travel between building floors [67]. These examples of how robots might still need help performing subtasks encouraged other researchers to examine the tasks that robots can ask for help with, such as directions [85], image labeling [72], and reaching objects [23]. Rosenthal et al., Bajones et al., and Rose advanced the literature investigating the best human candidate to approach for help location-wise [67, 2, 66]. Rosenthal et al. found that robots should ask for help in their location before finding someone nearby because the tradeoff between the robot's task completion and human-helper interruption impacts the likelihood of receiving help [67].

Bajones et al. and Rose also found that closer human and those in the same social range as the robot (roughly 10 feet) are more likely to help it [2, 66]. However, humans come from diverse backgrounds, have different abilities, and follow various cultural norms. While prior studies have focused on optimizing help for a particular task and recruiting the best candidate to help, we focused on exploring the cultural and situational inclusivity requirements for robots to succeed in seeking help in diverse locations.

To find the inclusivity requirements we sought to answer two research questions. The first research question focuses on the inclusivity requirements and the second question provides qualitative insights into why would people help and care for a robot.

RQ_Inclusivity: *What cultural and situational factors predict diverse people's likelihood to help and care for a robot?*

RQ_Reasons: *For what diverse reasons do people help and care for a robot?*

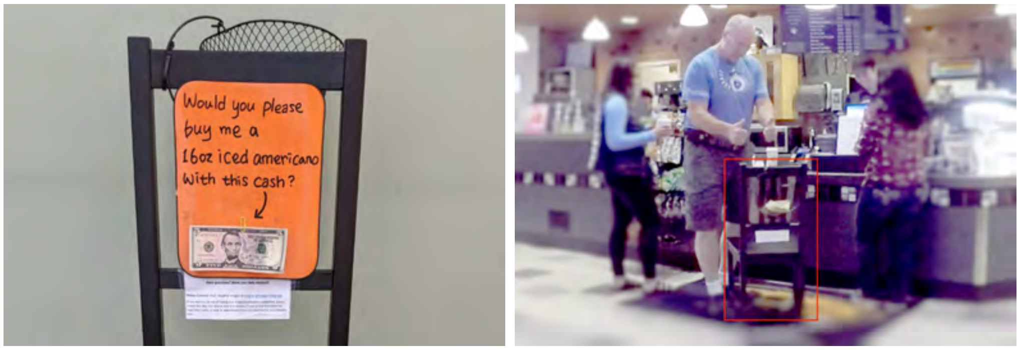
To answer the research questions, the ChairBot asked 268 participants to help it order food items from six cafes. We collected data about participants helping and caring behaviors using Ethnography and in-the-wild study. We analyzed the data to define six cultural and situational factors using ethnographic theory building and deductive data analysis. We then run a sequence of chi-square and simple logistic regression tests. Our statistical results revealed that four out of the six cultural and situational factors, significantly predicted participants' likelihood to help and care for the robot.

To provide qualitative insights into why would people help and care for the robot, we analyzed the data using inductive data analysis. We found that while the how varied participants had three main reasons for their behaviors: (1) Experiencing entertaining interaction, (2) Helping the individuals behind the robot deployment, and (3)Increasing the cafe's revenue.

The utilization of ethnography and in-the-wild study as evaluation methods, in this chapter, revealed how inclusive design might be applied to evaluate the inclusivity of social robots. Additional to the four design implications, this work showed that software inclusivity is relevant to evaluating beyond-WIMP UIs that use motion only to communicate with diverse humans.

3.2 Methodology

To answer the research questions, we controlled the ChairBot using the Wizard of Oz technique [1] in diverse locations, where we made it ask 268 participants to help it order food in a cafe. The robot asked for help via a small whiteboard to display various requests (e.g., “Would you please buy me a 16 oz iced americano with this cash?”, “Would you please buy me a blueberry muffin with this cash?”). Figure 3.1a shows how the robot asked for help and Figure 3.1b shows an interaction with a participant.



(a)

(b)

Figure 3.1: The ChairBot in Action. (a) A closup on the ChairBot with an adhered whiteboard, a clipped $5 bill, and a basket asking for help. (b) A cafe customer helping the robot buy an item. After placing the item in the basket, the customer gives the robot (in the red box) a thumbs up as it drives away.

We used two empirical evaluation methods to answer the research questions: ethnography and in-the-wild user studies (recall Chapter 2). Similar to [73, 54, 13], we chose ethnography to provide a detailed view and ensure a prosperous understanding of inhabited cultural factors by observing human behaviors in diverse social settings. We also picked *in-the-wild* user studies to discover realistic aspects of the interactions with the robot and its design [87, 65]. Thus, before presenting our results, here, we explain how we approached the study focusing on Research Sites, Data Collection, and Data Analysis in the following subsections.

[1] A user-testing technique allows a researcher (the “wizard”) to generate system responses as users interact with an interface without knowing that the wizard is behind the scenes pulling the levers and flipping the switches. The Wizard of Oz technique enables the interface to *appear to be real* as an up and running system [18, 42, 34, 68].

3.2.1 Research Sites

We conducted the study at six various cafes over eight weeks at a University campus. We visited each cafe twice for 12 total visits. We chose the six cafes from buildings that covered a range of pedagogical topics (engineering, English-as-a-second-language (ESL), business), and a variety of activities (library, dining hall, student union). All visits occurred around lunchtime and between 11:00am - 3:00pm, based on managerial approval, and lasted two hours. The robot interacted with diverse people, including faculty and staff, students, and parents on college site visit days. This study was approved by the university IRB. Figure 3.2 shows a map of the research sites and Figure 3.3 shows an abstract illustration of each one.

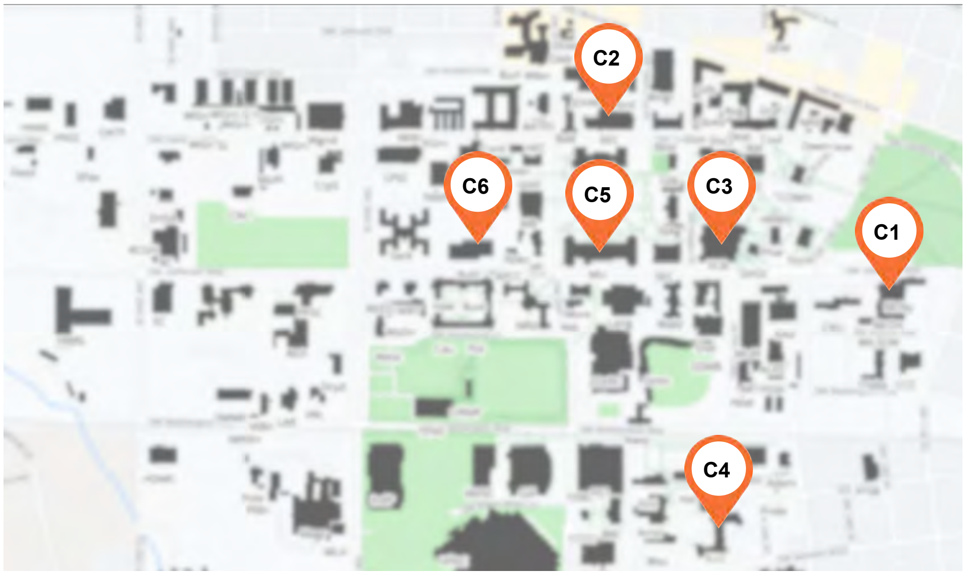
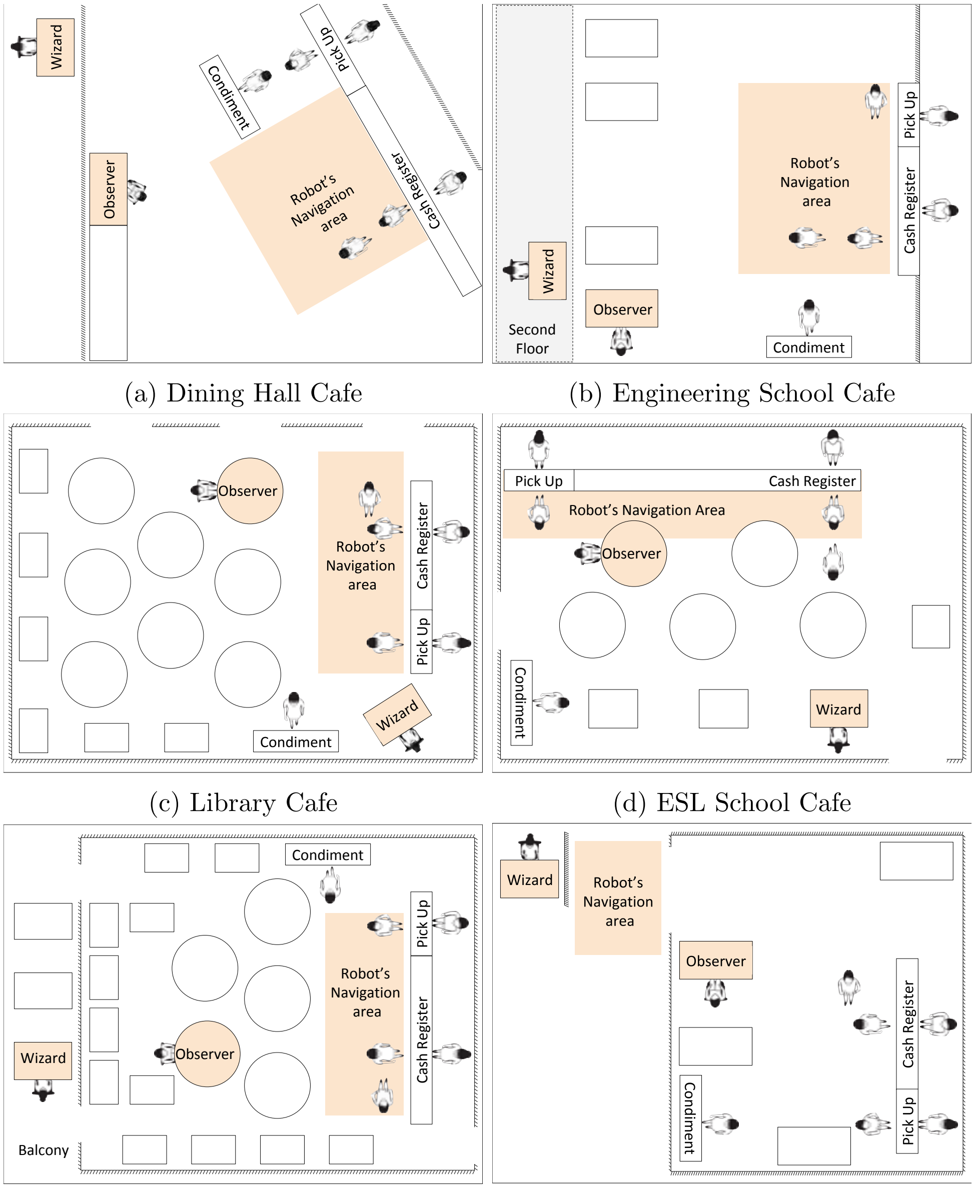


Figure 3.2: The research sites/cafes depicted on a map of the university campus. C1= Dining Hall Cafe, C2= Engineering School Cafe, C3= Library Cafe , C4= ESL School Cafe, C5= Student Union Cafe, and C6= Business School Cafe)



(a) Dining Hall Cafe

(b) Engineering School Cafe

(c) Library Cafe

(d) ESL School Cafe

(e) Student Union Cafe

(f) Business School Cafe

Figure 3.3: An abstract illustration of the research sites sorted by their architecture characteristics: Open-Space (a,b), Semi-Open (c,d), and Self-Contained (e,f). The locations of the ethnographer/observer, wizard and the robot are highlighted in orange.

The ***Dining Hall cafe*** featured in Figure 3.3a is located at the left corner of a first-floor food court in a student housing facility. It shares a wide-open hall area with six other food chains and a dine-in area without being barricaded. The cafe is not contained and connected to different stores in a dine-in area where a flow of people pass by. This cafe is located at one of the university's largest dining halls and serves a diverse range of customers coming in for a quick bite, a filling meal, or a relaxing break.

The ***Engineering School cafe*** featured in Figure 3.3b is located at the center corner of a wide-open first-floor atrium in a 153,000-square-foot building. The building's first floor extends up through several floors with a glass roof; thus, the cafe space can be viewed from upstairs and from stairways. There are classrooms, conference/meeting rooms, offices and a fishbowl computer lab near the cafe. There also is a spacious lounge area facing the cafe and a public piano for passersby to play. Similar to the Dining Hall cafe, this cafe serves a diverse range of customers, not only engineering faculty and staff.

The ***Library cafe*** featured in Figure 3.3c is located on the first floor of the university's main library. It features a semi-open architecture and occupies most of the floor. The cafe is bordered with a narrow barrier and pillars, allowing three access points. The remaining area outside of the cafe hosts three vending machines, restrooms, an elevator, and stairs to the library's main entrance. The cafe's customers are people who work and study in the library, in addition to customers who drop by to grab beverages and snacks.

The ***ESL School cafe*** featured in Figure 3.3d is located on the first floor of a living and learning center for international students. International students attending the university's ESL programs live and study on upstairs floors, where teachers and staff occupy the first floor in the daytime. Within the building, the cafe near a convenience market and residence's kitchen. The cafe area is semi-open with three transparent glass walls, and two open entrances between the glass walls.

The ***Student Union cafe*** featured in Figure 3.3e is located on the second floor of the university's student union building, in which large event rooms, dining facilities, offices, and lounges are clustered. Even though the building is usually crowded, the cafe is contained within opaque walls with a single narrower doorway. A variety of students, faculties, employees, and community members come to this cafe for numerous purposes ranging from formal interviews to casual chatting.

The ***Business School cafe*** featured in Figure 3.3f is located on the first floor of the business school building. It is barricaded with an opaque wall with a single entrance connected to the lobby where a lounge, a stairway, and elevators are located. The cafe occupies a smaller space compared with other semi-open and contained cafes, and has four tables. The customer group of this cafe is usually business students and faculty members. It is worth noting that the cafe manager at this cafe didn't allow the robot to be deployed inside the cafe to avoid crowdedness; thus, we deployed the robot in the lobby area right next to the cafe's entrance.

3.2.2 Data Collection

Similar to Yang et al. in-the-wild user study [87], we tracked participants helping and caring behaviors across all research sites by *Recording Videos*. We collected video footage using two cameras: primary and supplementary. The primary video footage was collected with a stationary wireless camera, and the supplementary video footage was collected with the USB camera attached to the wizard computer. We analyzed the video footage only from the primary camera; however, we occasionally referred to the supplementary one since both cameras offered different angles.

Additional to video recording and similar to Mutlu and Forlizzi [54], we collected data using two ethnographic methods: *Participant Observation* and *Fly-on-the-Wall Observations*. We used *Participant Observation* to gain intimate familiarity with and in-depth understanding of how the social context influences participant behavior and how the participant behavior influences the social context, while conducting open-ended interviews (aka unstructured conversations [53, 60]) with the participants. We used *Fly-on-the-Wall Observation* by following the robot from a distance as it asked participants for help, to observe the interactions as they took place in the natural context without influencing the social context.

In each site-visit, one of our team of two researchers played the ethnographer role (here forward observer) and conducted *Participant Observations*. To eliminate biases, halfway through each site-visit, the two researchers switched, allowing both to play the role of the observer in a counterbalanced manner. The observer wrote field notes in a shared notebook based on participants and cafes' workers interactions with the robot.

Interviews with the participants took place sporadically and were initiated by either the observer or the participant. The observer asked questions regarding the participants' interactions and their perceptions of the robot. When participants initiated conversations, some voluntarily reported their reactions, and others asked about details of the research and technical features of the robot. Meanwhile, interviews with workers took place when the customer flow was slow, or at the beginning and the end of the robot's deployment. Workers talked about their experiences and perceptions of the robot hanging around their workplace and gave the researchers clues on customers' reactions toward the robot. Such unstructured interviews allowed customers and workers to articulate their experience with the robot freely and in their own words.

Additional to participant observation, we used *fly-one-the-wall observation.* While participant observations were taking place, the second team member played the role of the wizard and conducted fly-on-the-wall observations through the supplementary camera from a distance. Because the wizard was operating the robot and simultaneously performing fly-on-the-wall observations, the wizard wrote the filed notes after finishing up each site-visit, or at times of no interactions. The two researchers also counterbalanced between the roles (observer and wizard) by switching after an hour of the site visit.

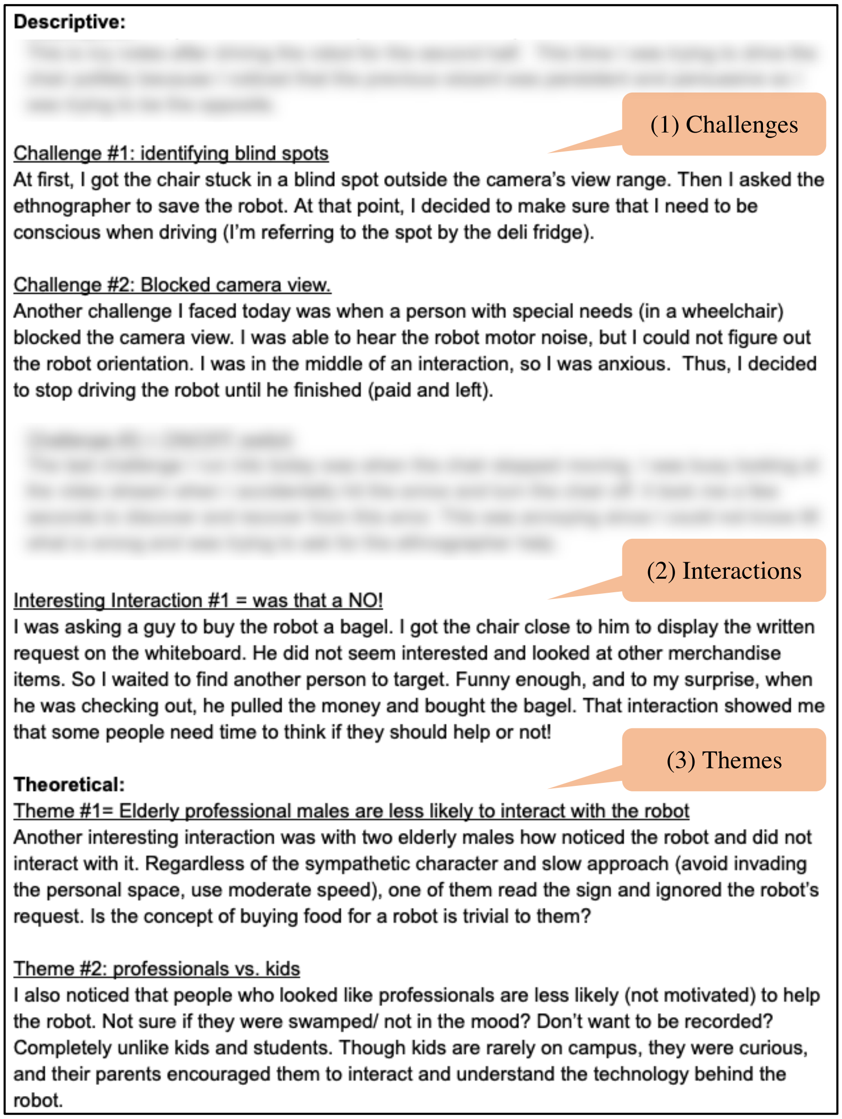
The two researchers shared notes were in a spontaneous (non-technical) language and reflected their reasoning and conclusion. The researchers recorded participants' interactions and initial theoretical themes. While playing the role of the wizards, the researchers also recorded the challenges encountered in operating the robot (i.e., getting the robot stuck in a blind spot). Figure 3.4 shows a sample of a wizard notes and its reflective nature including the operating challenges.

3.2.3 Data Analysis

To ensure an adequate understanding of the cultural and situational factors, data analysis consisted of ethnographic theory-building [60] in conjunction with inductive and deductive qualitative data analysis. The next three subsections cover each in details.

3.2.3.1 Ethnographic Theory Building

To develop the testable theories in this research, we rapidly analyzed the data using four major steps: identify themes, refine themes, link themes and develop theories (Figure 3.5). The four steps were repeated twice, once with two researchers and once with the whole research team. After each visit, two researchers examined the collected textual data, manually identified the initial themes, and noted them. In this process, the two re-



Descriptive:

[illegible]

Challenge #1: identifying blind spots
At first, I got the chair stuck in a blind spot outside the camera's view range. Then I asked the ethnographer to save the robot. At that point, I decided to make sure that I need to be conscious when driving (I'm referring to the spot by the deli fridge).

Challenge #2: Blocked camera view.
Another challenge I faced today was when a person with special needs (in a wheelchair) blocked the camera view. I was able to hear the robot motor noise, but I could not figure out the robot orientation. I was in the middle of an interaction, so I was anxious. Thus, I decided to stop driving the robot until he finished (paid and left).

[illegible]

Interesting Interaction #1 = was that a NO!
I was asking a guy to buy the robot a bagel. I got the chair close to him to display the written request on the whiteboard. He did not seem interested and looked at other merchandise items. So I waited to find another person to target. Funny enough, and to my surprise, when he was checking out, he pulled the money and bought the bagel. That interaction showed me that some people need time to think if they should help or not!

Theoretical:
Theme #1= Elderly professional males are less likely to interact with the robot
Another interesting interaction was with two elderly males how noticed the robot and did not interact with it. Regardless of the sympathetic character and slow approach (avoid invading the personal space, use moderate speed), one of them read the sign and ignored the robot's request. Is the concept of buying food for a robot is trivial to them?

Theme #2: professionals vs. kids
I also noticed that people who looked like professionals are less likely (not motivated) to help the robot. Not sure if they were swamped/ not in the mood? Don't want to be recorded? Completely unlike kids and students. Though kids are rarely on campus, they were curious, and their parents encouraged them to interact and understand the technology behind the robot.

Figure 3.4: A sample from a wizard notes with callouts to (1)Challenges while operating the robot, (2)Interactions description, and (3)Themes identification.

searchers asked each others' questions to justify the reasoning behind each theme, refine questionable themes, and link the themes to develop theories that answer the research questions. The two researchers were able to develop 8 theories to propose to the team.

The whole research team had two major meetings to finalize the developed theories.

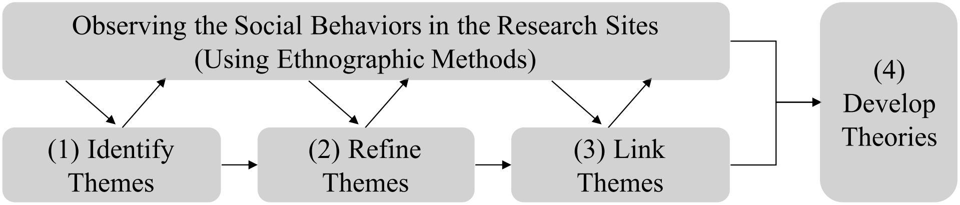

Observing the Social Behaviors in the Research Sites
(Using Ethnographic Methods)
(1) Identify Themes
(2) Refine Themes
(3) Link Themes
(4) Develop Theories

Figure 3.5: Our ethnographic theory-building process.

The initial meeting was two weeks into the study, and the research team gathered to identify and refine prominent themes (Figure 3.5, steps 1 and 2). At this meeting, the team noticed that some participants took extra measures to help the robot, thus, the researchers started referring to these behaviors as "caring behaviors". The second meeting was four weeks into the study, and aimed to refine the themes, link the themes and finalize a list of theories (Figure 3.5, steps 2, 3 and 4). At this meeting, the team interpreted the cultural factors that may predict participants likelihood to help the robot. The team also eliminated weak themes such as using the perceived age and gender of participants, since they might introduce biases.

3.2.4 Inductive Data Analysis

We also analyzed all the data using three levels of coding: *open, axial, and selective.* In *open coding*, one researcher (the reliability coder) segmented the video footage into interactions and assigned codes to the observed behaviors. For example, when the cafe at the ESL ran out of blueberry muffins, a participant went to a nearby store to buy a muffin for the ChairBot. This behavior was coded as *"going to a nearby store"* because it was further than the expected help.

For *axial coding*, we related the codes to each other using inductive reasoning. We used inductive reasoning to code the behaviors that stood out from a simple help, labeled them as "caring" behaviors, and initially categorized them to 5 categories; these were reduced to 3 categories afterwords.

Finally, we chose the core categories and its subcategories that emerged from the total 268 interactions with cafe customers for *selective coding.* For example, we decided during selective coding the 5 categories for the "caring" behaviors can be regrouped to form 3 categories. Additionally in this level we decided that care was a dependent variable we needed to analyze and report as a subset of help. Table 3.1 shows the finale set of categories and subcategories of the caring behaviors.

Table 3.1: The categories and subcategories for the caring behaviors.

Category	**Subcategory**
Volunteering Without Request	Helping without ordering Rejoining to help Going to a nearby store Interfering to help Staff help
Anticipating the Robot Needs	Opting for gourmet options Substituting to heather ingredients Opting for a similar item Adding Straw Including napkins Asking to have the order toasted Including utensils
Emotional Expression	Affirmation Patting Cautiousness

To ensure the *reliability of the analysis*, another researcher joined the reliability coder in coding the data. The two independently coded 20% of the video data and reached more than 80% agreement (counted using MAXQDA [2]). Given this agreement, the reliability coder coded the remaining data.

3.2.5 Deductive Data Analysis

A further dive into the deviations at each research site highlighted three situational factors. We used deductive reasoning to code these factors since no specific behavioral observations or patterns recognition needs to be established. For example, identifying *Food Items* and its levels { Grab and Go, Drink Order, Meal Order } as a situational fac-

[2] MAXQDA is a software program for computer-assisted qualitative and mixed methods data analysis in academic and scientific settings.

tor to assess the impact of preparation time for each food item on how likely participants help or care for the robot. We also used this top-down approach to why participants helped and cared for the robot.

3.3 Results

3.3.1 How Much Helping and Caring

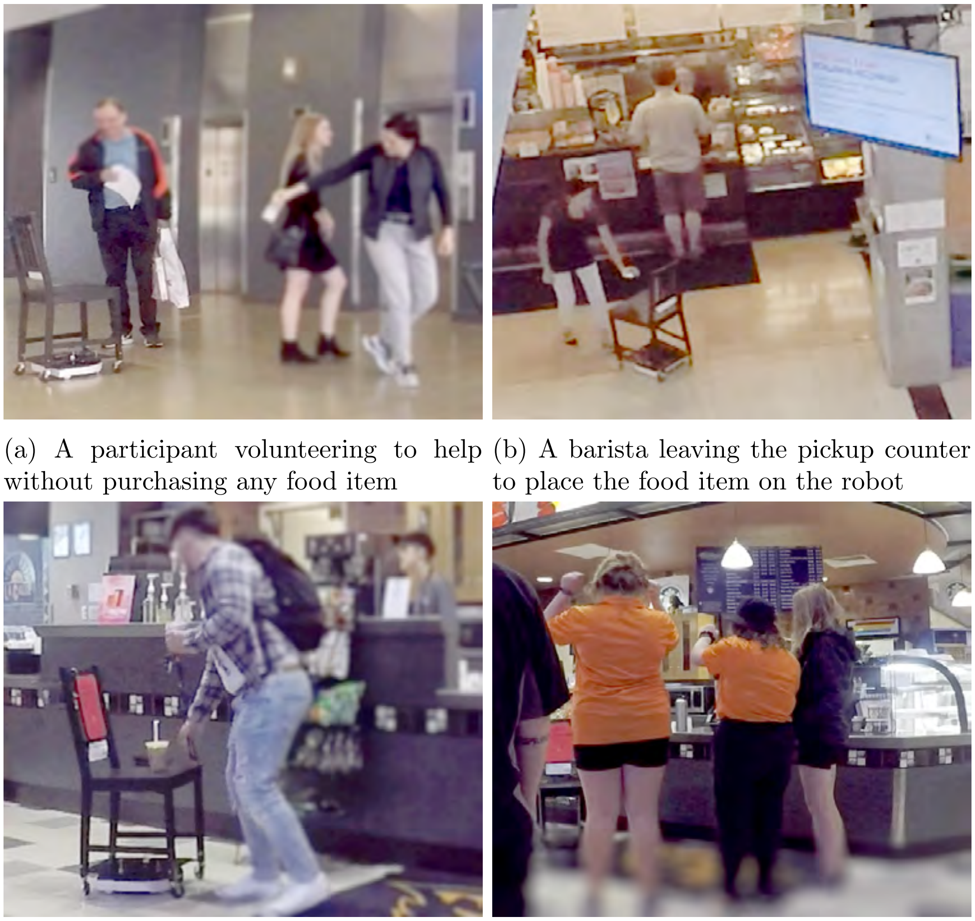
268 [3] different participants interacted with the robot. We defined an *interaction* as any encounter the robot had with a person while when it approached them with a help request or got their attention. If the targeted person did not look at the robot while it was in motion, we considered that as one-sided and not an interaction.

25% of participants who interacted with the robot also helped it. Participants helped the robot by buying the requested item it was asking for. Refusing to help the robot was either by ignoring the robot, stopping the interaction midway, or using any means of communication channels to indicate an unwillingness to help. Examples of the communication channels included shaking their head at the robot, saying "NO"or that they are busy. In a single extreme case, one participant picked up the robot and moved it out of their way.

A subset of *"helping"* was *"caring"*. 60% of participants who helped the robot went above and beyond just giving it what it wanted, demonstrating a level of care for the robot (recall Table 3.1). Caring behaviors had three main categories: (1)volunteering without request, (2)anticipating the robot's needs, and (3)encouraging the robot with positive statements or gestures. A peer-reviewed video with clips of theses behaviors is available [22], and Figure 3.6 shows some caring behaviors.

The most common category of caring behaviors was *volunteering without request* (18/41). The behaviors in this category included: placing and picking up the robot order without buying anything from the cafe (Figure 3.6a) and rejoining the order line after leaving the cafe. Some participants went to a nearby market to help, when a food item was sold out in the ESL school cafe. Five other participants took the initiative to help when noticing that the current participant is not taking action toward placing an

[3] To compare different levels of each cultural and situational factor, each participant was only exposed to a single level of each factor in a between-subjects (or between-groups) design fashion.



(a) A participant volunteering to help without purchasing any food item

(b) A barista leaving the pickup counter to place the food item on the robot

(c) A participant adding a straw and napkins to the requested drink item

(d) A group of participants cheering for the robot after helping

Figure 3.6: Four Examples of Caring Behaviors: (a) and (b) volunteering to help, (c) anticipating the robot needs, and (d) emotional expressions.

order. In some cases, cafe staff volunteered to help by leaving their position behind the register, placing the order, picking it up, and then putting it on the robot (Figure 3.6b).

Participant's *anticipation of the robot needs* was the second category caring behaviors, based on frequency (13/41). Considering that most participants were placing orders at a cafe shop themselves, this category included opting for gourmet option, substituting an ingredient with a healthier one, and picking up complimentary items(Figure 3.6c). In an interview with a participant who ordered a gourmet berries tea instead of a standard iced tea, the interviewee stated that they had never had this tea before but opted for it because it sounded like flavorful and refreshing summer beverage. Such behaviors showed that people were willing to go the extra mile to help the robot, instead of Just Do The Work.

The third category of caring behaviors was participants usage of *emotional expression* while interacting with the robot (10/41). This category included affirmative gestures and statements of encouragement such as patting on the robot, writing a positive letter on a napkin, and even cheering for the robot (Figure 3.6d). Participants used these expressions to confirm that they purchased the item and that the robot is ready to go. There were also a few participants who leaned on the cautious side and apologized to the robot itself when placing the wrong order or misplacing the reminding change.

3.3.2 The cultural and situational factors for predicting participants likelihood to help and care for the robot

To answer **RQ_Inclusivity** about the cultural and situational factors that predicted participants likelihood to *help* and *care* for the robot, we analyzed the diverse characteristics of the 268 interactions (recall 3.3.3 Data Analysis). In this study, we used R as a language and environment to statistically predict participants' likelihood to help and care for the robot. The *dependent variables* were help and care, and *independent variables* were the six cultural and situational factors. Chi-square tests predicted the statistical significance of each factor, and a simple logistic regression ,GLM[4], predicted the statistical significance of each level within a factor. We considered a *P<.05* significant*, *P<.01* significant** and *P<.001* significant***. This subsection defines, lists, and reports the cultural factors followed by the situational ones.

Cultural Factors

Cultural Factor is an ethnographical term refers to the shared patterns of behaviors and norms in a social group and are linked to local sets of social-atmosphere and conditions [4, 16]. The three cultural factors in this study were: *Social Atmosphere*, *Worker Attitude*, and *Architecture* (Table 3.2).

[4] GLM stands for Generalized Linear Model, a simple version of logistic regression for predicting categorical/binary outcomes. For example, predicting class's effect (the predictor) on the Titanic passengers' survival (the binary outcome).

Table 3.2: The three cultural factors and its levels.

Factor	**Levels**	**Research Site**
Social Atmosphere	Work Mood	Business School Cafe
	Mixed	Library Cafe
		Engineering Cafe
		Student Union Cafe
	Playful	Dining Hall Cafe
		ESL School Cafe
Worker Attitude	Friendly	Engineering Cafe (Visit#1)
		Dining Hall Cafe (Visit#1)
		ESL School Cafe (Visit#1)
	Neutral	Library Cafe (Visit#1)
		Dining Hall Cafe (Visit#2)
		ESL School Cafe (Visit#2)
		Student Union Cafe (Visit#1)
		Engineering Cafe (Visit#2)
		Business School Cafe (Visit#2)
	Unfriendly	Business School Cafe (Visit#1)
		Student Union Cafe (Visit#2)
		Library Cafe (Visit#2)
Architecture	Self-Contained	Business School Cafe
		Student Union Cafe
	Semi-Open	ESL School Cafe
		Library Cafe
	Open Space	Dining Hall Cafe
		Engineering Cafe

The first cultural factor, *Social Atmosphere*, referred to the overall mood of a cafe that influenced its norms and its levels were: { Work Mood, Mixed, Playful }. Social atmosphere was a significant factor in predicting participant's likelihood to help (X^2 (2,268)=10.802, $p = 0.004$**) and care for (X^2 (2,268)=7.4627, $p = 0.023$*) the robot. Figure 3.7 shows that out of the three social atmospheres the Mixed and Playful ones exhibited the most instances of help and care compared to the Work Mood.

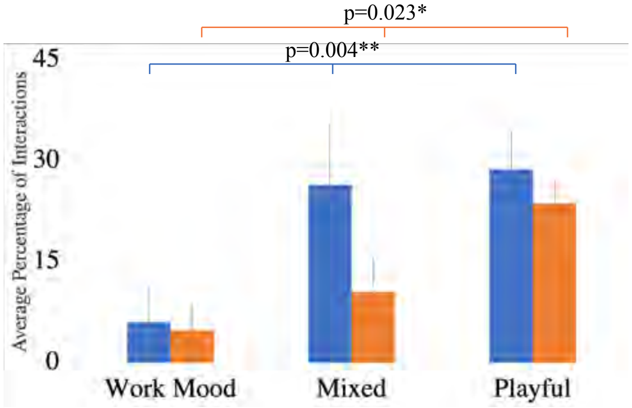


Figure 3.7: Average percentage of help and care interactions sectioned by the three levels of social atmosphere { Work Mood, Mixed, and Playful }.

The simple logistic regression test showed participants in the Mixed and Playful social atmospheres have higher odds (4.33 and 3.57 more times respectively) of helping the robot compared to participants in the Work Mood social atmosphere (Table 3.3). In the playful ESL school, where international students live and study, the robot's presence worked as a special event for them, especially during their first visit. The ESL students took videos and pictures of the robot and even interacted with the robot before the study began. After one student read the sign, they told the robot "Ok, come on. I will buy you a muffin! Can you follow me? Wow, you can follow me." Table 3.3 shows the odds of participants caring for the robot in the Playful social atmosphere were 3.64 times more than in the Work Mood social atmosphere. For instance, in the playful dining hall, three of the interactions were made by participants who had rejoined the physical area where the robot was wandering. One student said,"Oh, no one wants to buy a tea for you? I'll do it for you." On a different example, a woman who passed by the robot at first but doubled back to help, saying "Okay, I will buy you the chips, but you should go for a healthier option next time, okay?" All of these examples were distinct from the serious and work-like business school cafe where only 5 participants took the time to help the robot. This result suggests ***Design Implication #1:*** *Help-seeking robots should ask for help where the predominant behavioral norms are socially oriented rather than serious.*

Table 3.3: The p values and odds of participants helping (Top in Blue) and caring (Bottom in Orange) for the robot sectioned by the three levels of social atmosphere { Work Mood, Mixed, and Playful }.

Level	**P Value**	**Estimate Std.** (β)	**Odds** (e^{β})
Work Mood (*ref.*)	<.001 ***	-2.2618	0.10
Mixed	0.004 **	1.4660	4.33
Playful	0.015 *	1.2714	3.57
Work Mood (*ref.*)	<.001 ***	-2.5055	0.08
Mixed	0.333	0.5692	–
Playful	0.024 *	1.2925	3.64

The second cultural factor was *worker attitude*, and referred to cafes' staffs' friendliness towards the robot and its levels were: { Friendly, Neutral, Unfriendly }. Worker's attitude was a significant factor in predicting participant's likelihood to help (X^2 (2,268)=6.4448, p = 0.039*) and care for (X^2 (2,268)=11.239, p = 0.004**) the robot. Figure 3.8 shows participants helped and cared for the robot more when the worker's attitude toward the robot were friendly and natural compared to when the attitude was unfriendly.

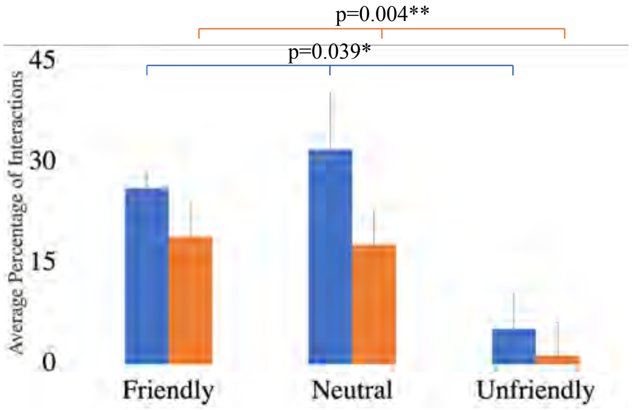


Figure 3.8: Average percentage of help and care interactions sectioned by the three levels of Worker Attitude { Friendly, Neutral, Unfriendly }.

The simple logistic regression test, however, did not support the previous numerical observation showing that only Friendly worker's attitude predicted participant's likelihood to help and care for the robot (Table 3.4). This result can be explained by Sauppe et al. observation that workers relate to robots as a social entities and operators with them as adjacent coworkers [73]. Mutlu et al. also found that when organizations adopt a beyond-WIMP system such as service robots, the robots impact the social dynamic and workload of the field workers, which in turn contribute to possible positive or negative reactions and attitudes toward the robot [54]. Based on the observations and numerical data during the two library visits were workers shifted from friendly at Visit#1 to friendly at Visit#2, we infer that this theory would be interesting to explore further in a more controlled setting. This result suggests ***Design Implication #2:*** *People would help robots if the people working around the robot have positive reactions and responses.*

Table 3.4: The p values and odds of participants helping (Top in Blue) and caring (Bottom in Orange) for the robot sectioned by the three levels of worker attitude { Friendly, Neutral, and Unfriendly }.

Level	**P Value**	**Estimate Std.** (β)	**Odds** (e^{β})
Friendly (*ref.*)	<.001 ***	-1.1474	0.32
Neutral	0.359	0.2872	–
Unfriendly	0.102	-0.9628	–
Friendly (*ref.*)	<.001 ***	-2.1068	0.12
Neutral	0.055	0.7787	–
Unfriendly	0.169	-1.4767	–

Architecture was the third cultural factor and referred to the spatial characteristics of a cafe, be it contained with opaque walls, semi-open, or open without being barricaded. Thus, the levels were { Self-Contained, Semi-Open, Open Space }. Figure 3.9 shows that the Cafe's architecture was a significant factor in predicting participant's likelihood to help the robot (X^2 (2,268) = 9.168, p = 0.010*) but not care (X^2 (2,268) = 3.6413, p = 0.162).

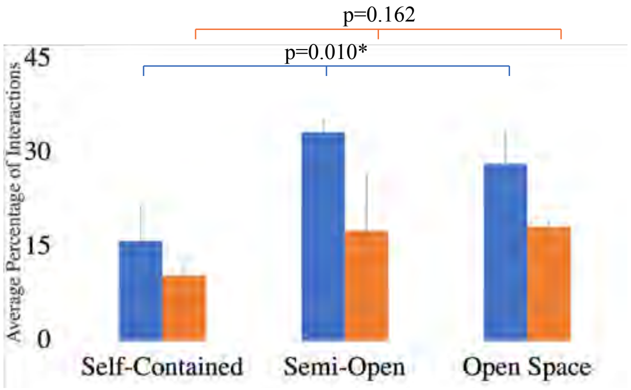


Figure 3.9: Average percentage of help and care interactions sectioned by the three levels Architecture { Self-Contained, Semi-Open, Open Space }.

The simple logistic regression test showed the odds of participants helping the robot in a Self-Contained cafe were 0.15 times lower than Semi-Open and Open Spaces (Table 3.5). We inferred that semi-open and open spaces were suited for informal social gatherings where giving a thumbs up to a robot is not outside the social norms. The dining hall's cafe, for example, is located at the corner of a first-floor food court in a student housing facility where it is common for students to study, unwind, and even watch TV. In contrast, the self-contained business school's cafe where helping the ChairBot garnered others' attention, and many participants did not appreciate the stares or scrutinizing glances. The simple logistic regression test also showed that odds of participants caring for the robot were only significant in the Self-Contained spaces (Table 3.5). This result suggests ***Design Implication #3:*** *help-seeking robots can ask for help in places where people are chilling and unwinding rather than engaged.*

Table 3.5: The p values and odds of participants helping (Top in Blue) and caring (Bottom in Orange) for the robot sectioned by the three levels of architecture { Self-Contained, Semi-Open, Open Space }.

Level	**P Value**	**Estimate Std.** (β)	**Odds** (e^{β})
Self-Contained (*ref.*)	<.001 ***	-1.8718	0.15
Semi-Open	0.005 **	1.1580	3.18
Open Space	0.017 *	0.9438	2.57
Self-Contained (*ref.*)	<.001 ***	-2.2736	0.10
Semi-Open	0.289	0.5444	–
Open Space	0.069	0.8344	–

Situational Factors *Situational factor* is another ethnographical term that refers to minutely and opportunistically constructed conditions of a setting where moment-by-moment social interactions are embedded. In our study, each interaction between the robot and a participant had 3 situational factors: The type of the requested food item, the robot approaching styles, and the number of people in an interaction (Table 3.6).

Table 3.6: The three situational factors and its levels.

Factor	**Level**	**Examples**
Food Items	Grab and Go	Bakery or pre-packaged goods
	Drink Orders	Coffee, tea, or smoothie
	Meal Orders	Grilled sandwiches or wraps
Robot Approaching Style	Pushy	Address and repeat at list twice
	Subtle	Address and repeat at most twice
Number of people	Individual	Single person
	Group	Two or more people

To investigate the influence of different cafes' different menu items, the robot ordered a variety of food items from each cafe. This helped us assess the impact of preparation time for each food item on how likely participants help or care for the robot. Food item, the first situational factor, was categorized as { Grab and Go, Drink Order, Meal Order } based on preparation time. Grab and go orders required no preparation time, drink orders required some preparation time, and meal orders referred to advanced food items that required longer preparation time.

While the numerical distribution of helping and caring instances in Figure 3.10 shows that most participants helped the robot when it asked for grab and go item, the chi-square test showed that food item was not a significant factor in predicting participant's likelihood to help the robot (X^2 (2,268) = 4.7662, p = 0.092) or care for it(X^2 (2,268) = 3.71, p = 0.157). This result contradict our initial theory that time allocation influenced participates likelihood to help and care for the robot. We suggests that the theory about time allocation would be interesting to explore further in a more controlled setting, similar to how [67] investigated the influence of request time and repetition on participates likelihood to help the robot.

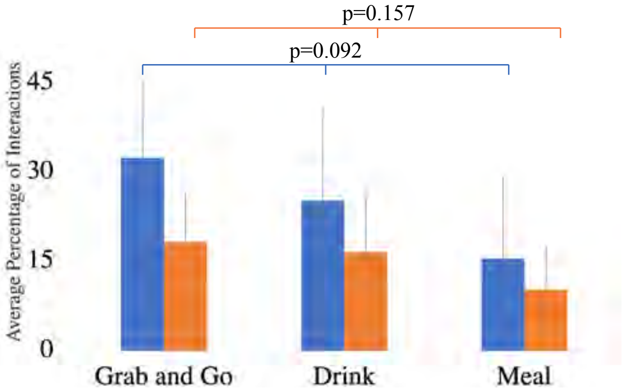


Figure 3.10: Average percentage of help and care interactions sectioned by the three levels of Food Items { Grab and Go, Drink Order, Meal Order }.

The second situational factor was the *robot's approaching style* and referred to the robot's Motion characteristics, where the robot could either be { Subtle, Pushy }. When the robot was pushy, it addressed a participant and repeated the help request at least twice, but did not stop requesting help until the participant showed signs of refusing to help. If the participant pulled out the clipped cash, the robot span in a happy-like dance. On the other hand, the subtle-approaching ChairBot addressed a participant and repeated the help request at most twice. If the participant pulled the clipped cash, the robot accompanied the participant to the register.

Figure 3.11 shows that the robot's approaching style was a significant factor in predicting participant likelihood to help the robot (X^2 (2,268) = 4.7867, p = 0.029 *) but not care for it (X^2 (2,268) = 3.5945, p = 0.058). The simple logistic regression test showed the odds of participants helping a the Subtle robot were 0.25 times lower than when it was Pushy(Table 3.5). While the robot's approaching style was not a significant factor in predicting participant likelihood to care for the robot, numerically participants cared for the pushy robot more than the subtle one by customizing the order to include healthier options such as ordering a crepe without cream. Even participants who did not help the pushy robot expressed a sense of care saying "Sorry, I'm in a hurry." Patting the robot was another common behavior that more frequently happened with the pushy robot. Participants engaged more with the pushy robot, such as one participant saying: "Turn around if you want a muffin!" This result suggests ***Design Implication #4: help-seeking robots should be direct and persistent when asking people for help.***

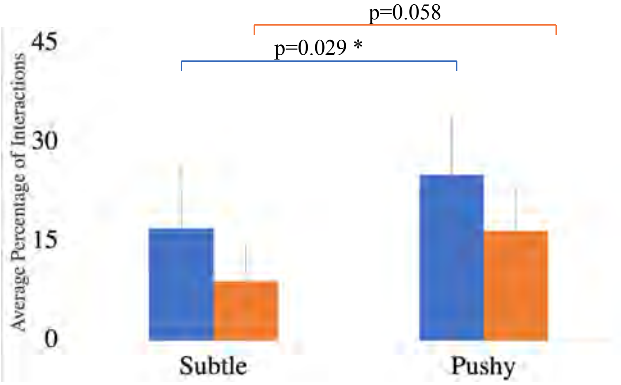


Figure 3.11: Average percentage of help and care interactions sectioned by the two levels of Robot's Approaching Styles { Subtle, Pushy }.

Table 3.7: The p values and odds of participants helping (Top in Blue) and caring (Bottom in Orange) for the robot sectioned by the two levels of the robot approaching styles { Subtle, Pushy }.

Level	**P Value**	**Estimate Std.** (β)	**Odds** (e^{β})
Subtle(*ref.*)	<.001 ***	-1.3952	0.25
Pushy	0.030 *	0.6182	1.86
Subtle(*ref.*)	<.001 ***	-2.0557	0.13
Pushy	0.061	0.6496	–

The third situational factor was *the number of people* in the interaction. The real-world nature of the study did not exclude anyone from interacting with the robot, and we identified two levels for this factor{ Individual, Group }. Similar to [27, 71], the term groups referred to "two or more individuals who are apparently connected by a social relationship before interacting with the robot." Participants exhibited caring behaviors toward the robot when alone than in groups of two or more participants (Figure 3.12). However, this was not a statistically significant factor for predicting participant's likelihood to help (X^2 (2,268) = 0.025043, p = 0.874) or care for the robot (X^2 (2,268) = 0.0021311, p = 0.963). We infer that this result would be interesting to explore further by collecting data about group dynamics. This inference is concurrent with Sebo et al.'s argument that when interacting with robots, groups exhibit behaviors that differ from individuals, especially in decision-making behaviors [75] such as helping a robotic chair.

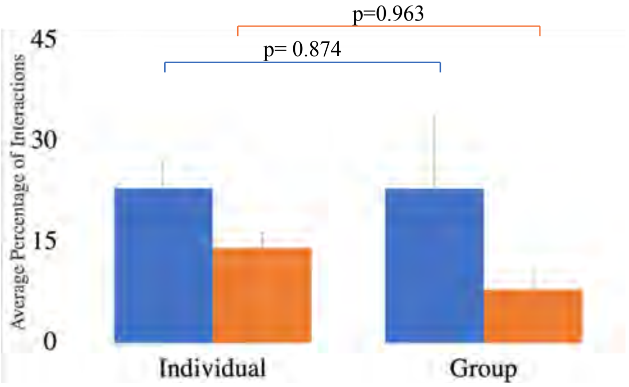


Figure 3.12: Average percentage of help and care interactions sectioned by the two levels of Number of People { Individual, Group }.

3.3.3 Why participants helped and cared for the robot

When answering **RQ_Reasons**, about why these diverse participants ended up helping and caring for the robot, participants had varying 3 main reasons for their behaviors: *(1) Amusement/Curiosity, (2) Helping other people,* and *(3) Increase revenue.*

(1) Amusement/Curiosity Some participants were amused by their interaction with the ChairBot, even if they had seen it before. At the engineering school, where researchers performed two previous ChairBot studies, the cafe manager introduced the robot to the customers as a "regular customer" and encouraged customers to interact with it. The employee told a customer, "This little guy gave me a candy last Halloween, and this time it is asking me to buy a coffee, haha, how fun!" Some participants were also curious about how the robot worked. In one instance, two students jointly tested the robot's perception and path planning by moving around the robot, then went to the researcher asking technical questions "Does the robot see people or have object detection functionalities? How does the robot realize if it has received food?"

(2) Helping other people Some participants helped the robot to help the research team. Other reasons for assisting the robot included a desire to help the person behind the robot. For example, a student who was going through his mid-term said: "I imagined that a student who is preparing a mid-term had sent the robot to the cafe to buy food. I wanted to help the busy student." Similarly, a participant came to the researcher after helping the robot and stated: "I didn't help the robot but you, a researcher who is running this study. I also do my own experiments as a graduate student, and I know how hard it is. If I didn't notice you, I wouldn't help the robot."

(3) Increase revenue Finally, some cafe staff helped the robot to promote their business and increase revenue. For example, the cafe manager at the library said she wanted to help the robot as long as it garnered customers' attention and increased revenue. The manager at the dining hall said that he wanted to have the robot in as long as it increased revenue and brought joy. A staff member followed his comment: "Oh you are talking about Charles, we call him Charlie the chair here. Look at our customers enjoying watching the robot. I think this little one is helping develop a good mood in this cafe, and I wanted it to get help to keep amusing people."

In conclusion, this chapter showed that inclusivity is relevant to beyond-WIMP user interfaces. We investigated the inclusivity requirements in terms of culture and situational factors that predicted diverse peoples likelihood to help and cared for a robot in diverse locations. The next chapter examines the ability of an inclusive software design method for evaluating and redesigning a multiple robot controller.

Chapter 4: Examining the Capability of an Inclusive Software Design Method for Evaluating and Redesigning a Multiple Robots Controller

The previous chapter demonstrated the applicability of inclusive software design as an approach to evaluate the inclusivity of a beyond-WIMP robotic chair in the role of a help-seeking; however, what about the applicability of inclusive software design methods devised for WIMP UIs? Can an inclusive software design method be copy-cat to evaluate a beyond-WIMP UI?

Toward that end, this chapter evaluates the inclusivity of beyond-WIMP UIs that control multiple robots, empirically and analytically, using the GenderMag method (recall Chapter 2). We selected GenderMag as our inclusive software design method because previous work established the importance of accounting for gender differences in problem-solving styles when developing UI for interacting with robots [77]. More specifically, Showkat et al. and Balali et al. suggested that supporting some of the GenderMag facets could be valuable in creating inclusive robotic systems [77, 3].

We started with a touch-based (physical) controlling UI that allows multiple robots to move. We then ran an empirical lab study with 12 participants to collect data for two GenderMag facets and how they played out with participants' usage of the robotic UI. We then conducted an analytical evaluation to find the inclusivity issues for the complete set of GenderMag's facets. We finally redesigned the touch-based controlling interface and added an additional mobile UI that addresses the reported issues.

The context in this chapter is furniture arrangement with the robot from the previous chapter—ChairBot. We chose furniture arrangement because robotic chairs that automatically relocate could reduce the time and energy associated with organizing social events if they have a functional controlling interface for diverse end-users. The beyond-WIMP controlling interface here is fundamental to providing the initial instructions, setting operational timing, and teaching the robots. Designing beyond-WIMP UIs are challenging since users have to operate multiple robots based on knowledge similar but not equivalent to ordinary non-robotic chairs.

We chose furniture arrangement as a study context to evaluate the inclusivity of the beyond-WIMP controlling interface using GenderMag and framed the research questions around the inclusivity requirements as follows:

RQ_Issues: What issues must the ChairBots controlling interfaces eliminate to be inclusive?

RQ_Styles: What diverse arranging styles must the ChairBots controlling interfaces support to be inclusive?

We also added a research question to assess the functionality of the interface by comparing the performance of the ChairBots to no-robotic chairs as follows:

RQ_Performance: What do people think of the ChairBot in terms of Mobility, Ease of Use, and Enjoyability?

The utilization of GenderMag in this chapter revealed the capability of inclusive software design methods that have been devised for WIMP UIs to be used effectively to evaluate and redesign beyond-WIMP UIs. Our empirical results revealed that all usability issues with the touch-based (physical) UI must eliminate to ensure gender problem-solving inclusivity and the “Invisible Elements” usability issue had the most pressing bias percentage. Additionally, of the four arranging styles the participants used, “One-by-One” was the most utilized among gender problem-solving styles. These findings guided our redesign decision and steered our journey to inclusive software design for beyond-WIMP UIs.

4.1 Empirical Study

4.1.1 Methodology

To investigate the inclusivity of the ChairBot with end-users, we used the only version of the robot with attached touch sensors as an end-user UI. The touch sensors communicates with the chair base (gray base at the bottom of each chair in Figure 4.1), which is how users can make the chair moves. In this empirical study, the touch sensors enabled participants to instruct the chair by touching it to go forward, backward or rotate in place. Also, we sat up the ChairBots to allow people to send the same motion command to one or several robots.

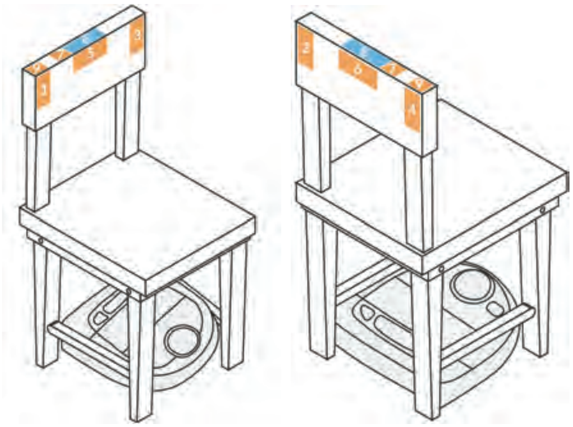
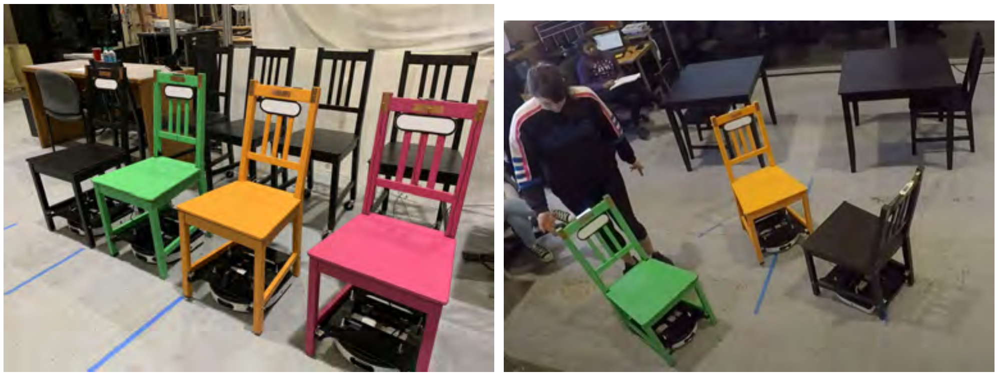


Figure 4.1: Schematic diagram of a ChairBot with 6 adhered touch sensors. **1,2:** Turn Left, **3,4:** Turn Right, **5:** Go Backward, **6:** Go Forward, **7:** Turn the robot On/Off, **8:** LED indicator, and 9: Turn All robots On/Off

We conducted a 2x2 mixed study for 12 participants to arrange both robotic and non-robotic chairs around either an empty space and/or a table. The participants were 18 - 35 years of age and their gender varied (six men, five women, and one gender non-conforming). We had two counterbalanced independent variables. The first independent variable was *chair type*, with ChairBots and non-robotic chair type being the same model of chairs on casters (Figure 4.2a). The other independent variable dictated the space around which the participants arranged the chairs: two preset tables or an empty space (Figure 4.2b). The order in which a user interacted with the chair types varied for each user such that half interacted with the Robotic chairs first(Table. 4.1). Using a 1-5 Likert scale (1=None and 5=Expert), only one participant identified themselves as an expert with robots while the majority described themselves as competent (M=2.58, SD=1.31). In this thesis we only analyzed the data around the first independent variable—*chair type* to answer **RQ_Performance**.

	Non-Robotic	**Robotics Chairs**
Empty Space	N=5	N=7
Around Table	N=7	N=5

Table 4.1: We recruited participants and divided them between two independent variables (chair type and space type)



(a) (b)

Figure 4.2: (a) A close-up on the ChairBot and non-robotic chairs participants were asked to arrange. (b) A participants arranging the chairs to face each other using touch sensors while the researcher at the back taking notes

The study procedure consisted of a consent form, orientation to a Think Aloud protocol[1], and two chair arrangement sessions corresponding to the study conditions. For each session, we gave participants a set of 3 chairs followed by a scripted tutorial to demonstrate the relevant chair type's functionality. Once the participants finished an arrangement session, they were asked to fill out a short survey about their experience. Since we asked participants to Think Aloud by verbally expressing their thoughts and reactions, before their first session, the participants practiced by talking through estimating how many windows were in their homes. At the end of the two chair arrangement sessions, we conducted semi-structured interviews with each participant focusing on their expectations from the robots and the issues they faced.

During the empirical study, we collected verbal and behavioral data. We collected behavioral data through our analysis of the video-recorded study area (i.e., participants' actions). We also collected verbal data from the Think Aloud, surveys, and interviews with the participants.

We analyzed the data in three steps. First, we transcribed the study sessions and segmented them by conversational turns (i.e., change of speech). Then, two researchers

[1] A user-testing method requires participants to verbally comment on what they are doing and thinking as they complete a task. The method is appropriate for revealing the cognitive processes in which participants perform tasks and aspects of the user interface that delight, confuse, and frustrate them [34, 30, 45].

independently coded 20% of the data. We selected the 20% data randomly from 4 different study sessions. Our three code sets covered evidence of two GenderMag facets (Computer Self-Efficacy and Learning Styles), usability issues faced by the participants, and chairs arranging styles. The two researchers reached an agreement of 98%(Jaccard index). Given this reliability, one of the researchers coded the rest of the data as the last step. Table 4.2 lists these three coded sets and the sources for each.

We used participants' responses from the survey to analyze the performance of the robotic and non-robotic chairs based on three factors: Mobility, Ease of Use, and Enjoyability. Participants evaluated each factor using a 1-5 Likert scale (1=Strongly Disagree and 5=Strongly Agree). The first factor, Mobility, referred to the average of scores based on how *expected, appropriate, and natural* the motions were as perceived by participants. The second factor, Ease of Use, referred to the average of scores based on how *obvious, easy to use, and convenient* the chairs were as perceived by participants. and the third factor, Enjoyability, referred to the average scores based on how *likeable, pleasant and simple* the chairs were as perceived by participants.

Table 4.2: The three code sets from the empirical study.

Code Set	**Data Source**	**Code**
Evidence of GenderMag Facets	Interview, Think Aloud	Low Computer Self-Efficacy, high Computer Self-Efficacy, Learning without Tinkering, Learning by Tinkering,
Usability Issues	Interview, Think Aloud, Video	Memory Challenges, Invisible Elements, Accidental Activation, Motion Inaccuracy
Arranging Styles	Interview, Think Aloud, Video	One by One, Trying Things, Clear The Stage, Clustering

4.1.2 Results

For the first code set, *Evidence of GenderMag Facets*, we deductively tied the facets (Computer Self-Efficacy and Learning Styles) to participates' verbal data as follows:

- *Low Computer Self-Efficacy*: If the participant blamed technology failures on themselves, or mentioned not being good or familiar with technology/robots, or said it's hard for them. An example of this code is when we asked P8 about their overall experience with the ChairBots in the post sessions interview and stated: "It was kind of hard . . . Since these are robots, I was afraid I was going to break them. So, that was kind of like a big worry factor". 7/12 of the participants in this study showed low level of Computer Self-Efficacy (confidence) when interacting with the ChairBots.

- *High Computer Self-Efficacy*: If the participant blamed technology failures on the robot or had positive self assessment. An example of this code is when P1 accidentally turned all the ChairBots On instead of a single ChairBot, and was surprised when all of them moved. P1 startled when all the chairs moved and stated: "The chairs don't obey me!" when in fact P1 was the one who accidentally accidentally activated the Move All feature. 5/12 of the participants in this study showed high level of Computer Self-Efficacy (confidence) when interacting with the ChairBots.

- *Learning without Tinkering*: If the participant adjusted the chairs following a clear plan or idea. An example of this code can be seen when P4 was arranging the ChairBots stating: "I'm just going to start doing something simple to see if I can get both robotic chairs to get tucked in to the other sides of the tables." Here P4 verbalized a clear plan and walked us through it. 9/12 of the participants in this study did not tinker around to learn how to interact with the ChairBots.

- *Learning by Tinkering*: If the participant experimentally adjusted the chairs without following a clear plan or idea. An example of this code is when we asked P2 about their intentions in the first arrangement session, P2 stated: "I don't know, when I have a robot I tend to be more playful with it because I like tinkering with stuff." This example showed that the participant's was tinkering around without following a clear idea about where each robot should be. 3/12 of the participants in this study tinkered around to learn how to interact with the ChairBots.

We identified a second code set, *Usability Issues*, independently of the first code set so as to later look for co-occurrences. *Usability Issues* referred to the 44 issues the participants faced when arranging the ChairBots. We inductively grouped these issues into 4 codes and labeled them using the UI Tenets and Traps cards [49] (Figure 4.3) as follows:

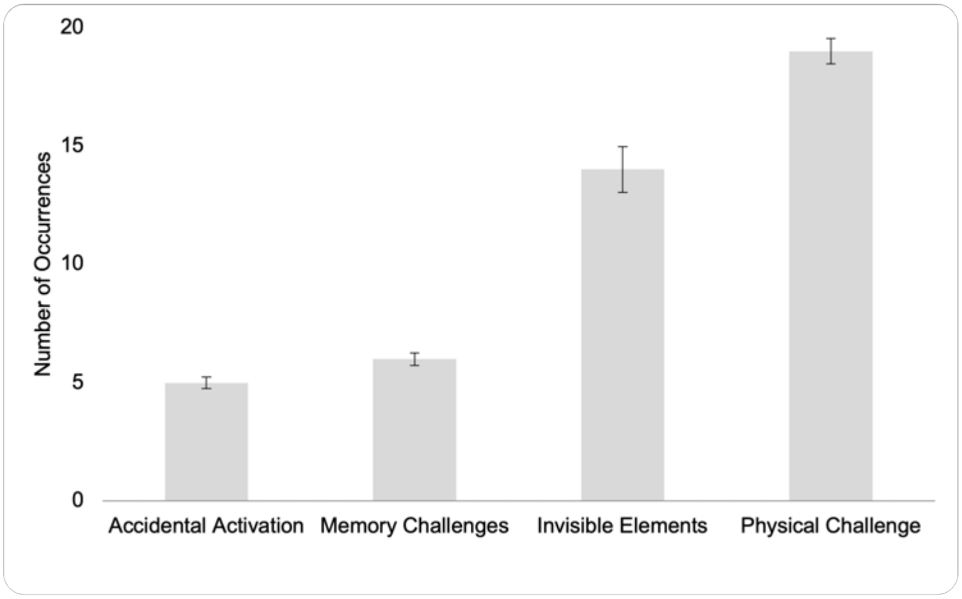


Figure 4.3: The distribution of the issues participants faced

- *Accidental Activation*: If the participants were confused about activating the single or multiple robots. The robotic chairs can be switched to a group mood that allows participants to move them together, yet (5/44) 11% of the issues occurred due to accidental activation of these two sensors (7 and 9 in Figure 4.1). For instance, P1 accidentally turned all the ChaiBots On and was startled when they moved. We considered this instance and other similar ones a usability issue because the two touch sensors were placed too close to each other, making participants prone to accidental activation and resulting in an unintended robot motion.

- *Memory Challenges*: If the participants had to recall turning the ChairBots On/Off. P11, for instance, had finished orienting a ChairBot, turned it Off, then tried moving to the second ChairBot. When the second ChairBot did not move, P11 said, "I'm not thinking of turning [the robots] On and Off. I remember the Off but not On." We considered this instance a usability issue since the robots required the participants to recall toggling between On/Off, challenging how they are used to interacting with everyday objects that do not move after placing them. This usability issue occurred only (6/44) 14% of the time.

- *Invisible Elements*: If the participants touched/triggered the wrong sensor. Since the sensors did not have any visible functionality cues (labels) or physical affordability, participants often forgot the directional movement attached to each sensor. P5, for instance, stated: "I think before when I pressed on this one [sensor#4], it went to the right instead of this way [left]." Here P5 thought the sensors worked differently in different ChairBots, forgetting that they were using two different sensors. P5's confusion showed that participants needed visual cues to signal the directional movements of each sensor rather than learning to overcome its absence. This usability issue occurred (14/44) 32% of the time.

- *Physical Challenge*: If the participants wrestled to place the chair accurately because of the predefined motions. To increase the speed of the robots, the ChairBots had a predefined set of motions that allowed it to move forward/backward 100 mm and rotate right/left 45 degrees per touch with a 5 degrees margin of error, yet (19/44) 43% of the issues faced by the participants had to deal with this error. P9, for instance, stated: " . . . I had a job for six years arranging tables and chairs, and my boss was very serious about [accurate placements]. [The boss] would come in, and all the chairs must be in a perfect line . . . , and if it wasn't right, [The boss] would fix it." We considered this instance and similar ones a usability issue because it was difficult for P9 to ultimately place the robots with 100% accuracy, as they wanted.

To answer **RQ_Issues** we segmented the data from the second code set, *Usability Issues*, by the first code set, *Evidence of GenderMag Facets*, such that the issues are grouped by participants' GenderMag Facets (Table 4.3 and Table 4.4). The segmentation showed the UI must eliminate all 4 issues to be inclusive and the Invisible Elements issue had the highest percentage of bias compared to the the other issues.

The last column (Issues%) in Table 4.3 shows all four *Usability Issues* were encountered by at least one participant's with high-self efficacy (blue) and one participant's with low-self efficacy(orange). Out of the four *Usability Issues* {Accidental Activation, Memory Challenges, Invisible Elements, and Physical Challenge}, (8/12) participants faced the Invisible Elements issue, and participants with high self-efficacy (blue) were less likely to face that issue than participants with low self-efficacy (orange). An example of the Invisible Elements issue is a participant touching the backward sensor instead of the forward leading to triggering unintended backward motion. The fact that a given chair might not move as a participant expected was not comfortable for those with low self-efficacy and affected their confidence in performing the task of making an arrangement. This observation is consistent with Showkat et al. [77]'s high self-efficacy resulted in high task success and lower task completion time when compared to low self-efficacy.

Table 4.3: The usability issues grouped by evidence of participant's Computer Self-Efficacy: High and Low.

	Computer Self-Efficacy												Issues%
	High					Low							
	P2	P1	P6	P9	P7	P10	P5	P3	P11	P8	P4	P12	
Accidental Activation	*	*				*	*		*				40%, 43%
Memory Challenges	*		*	*		*		*		*			69%, 43%
Invisible Elements			*		*	*	*	*	*	*	*		40%, 86%
Physical Challenge	*	*		*	*	*	*	*			*	*	80%, 71%

Additionally, Table 4.4 shows participants who did not tinker around with technology (orange) were likely to encounter the Invisible Elements issue. None of the tinkering participants (blue) faced the Invisible Elements issue, suggesting that they regarded it as an anticipated motion and adapted the arrangement around it. For example, P5, who was not a tinkerer, asked the facilitator to guide them through which sensor to use even after the initial tutorial: " Which one should I press? So that the robot can move forward . . . [after moving the ChairBots forward] So these two [sensor#3 and sensor#4 in Figure 4.1] are doing the same thing? These two corners are doing the same thing? If I push on this side it does the same thing?"

For the third and last code set, *Arranging Styles*, we inductively derived 4 unique codes from a total of 74 instances of arrangement styles with both types of chairs (Figure 4.5). Theses code sets were: One By One, Trying Things, Clear The Stage, and Clustering. The most used arrangement style was *One By One* and the least used was *Clustering*. All 12 participants used the *One By One* style for at least a part of their session.

The first and most used arrangement style *One By One* refers to instances where participants moved chairs one by one to the final position (Figure 4.5d). While this

Table 4.4: The usability issues grouped by evidence of participant's learning Styles: With Tinkering and Without Tinkering.

	Learning Styles												**Issues%**
	With Tinkering			**Without Tinkering**									
	P2	**P1**	**P9**	**P10**	**P5**	**P3**	**P11**	**P6**	**P8**	**P7**	**P4**	**P12**	
Accidental Activation	*	*		*	*		*						67%, 33%
Memory Challenges	*		*	*		*		*	*				67%, 44%
Invisible Elements				*	*	*	*	*	*	*	*		0%, 77%
Physical Challenge	*	*	*	*	*	*				*	*	*	100%, 67%

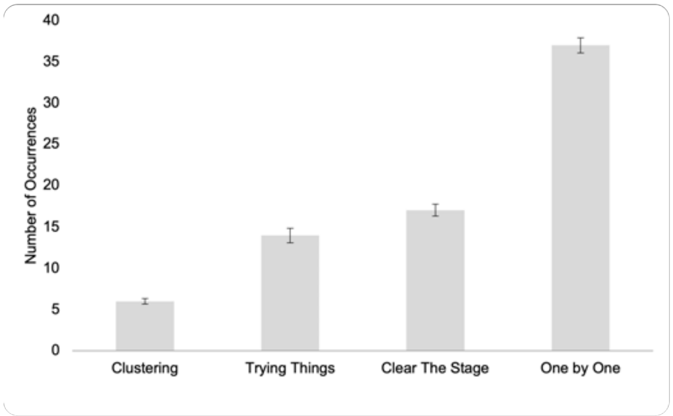
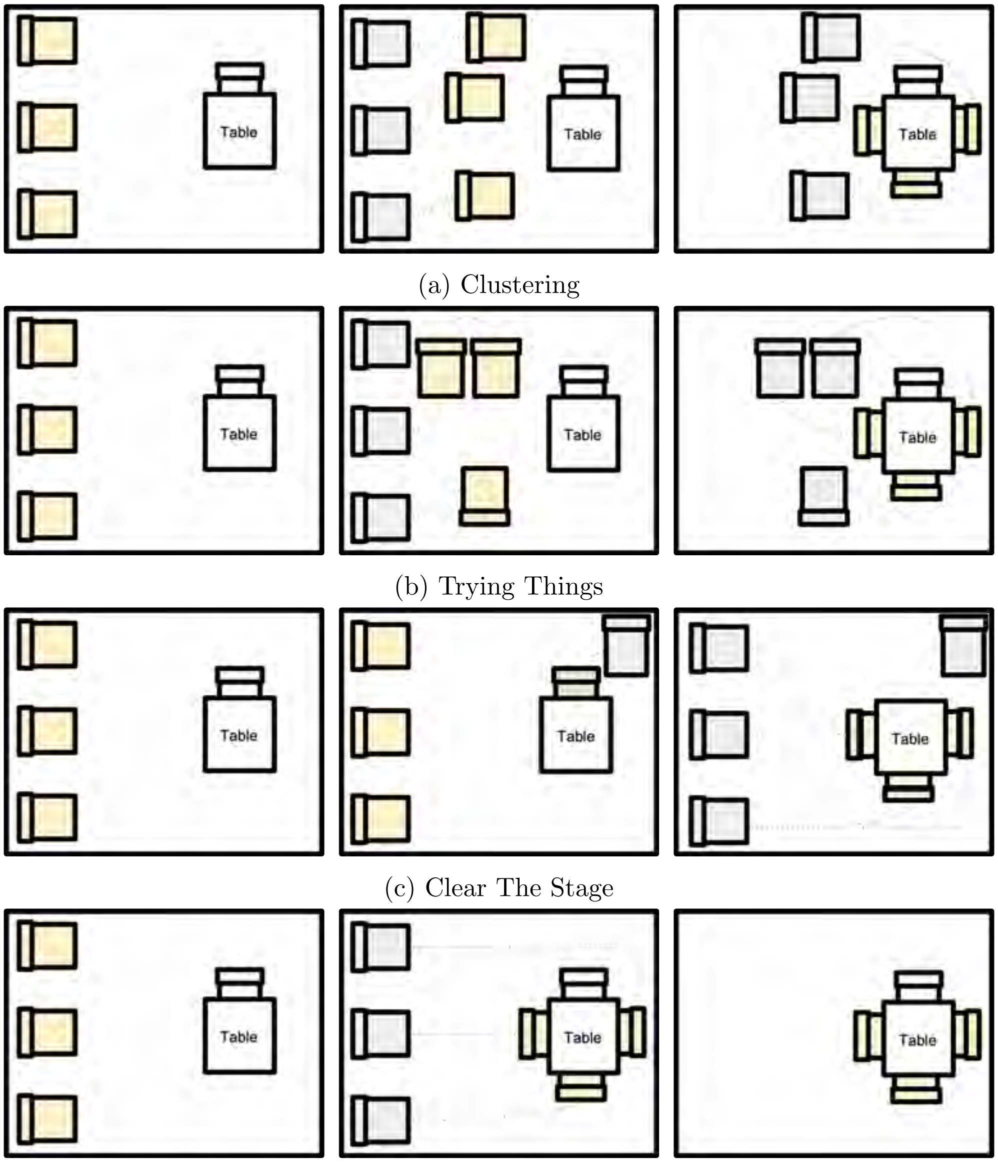


Figure 4.4: The distribution of the arranging styles participants used

style might be time-consuming with more chairs, all participants found the style handy, especially when they already knew how they wanted the chairs to be oriented or when they had moved the chairs to their final destination. Participants used this arrangement style (37/74) 50% of the times.

The second style, *Clear The Stage* refers to instances where participants cleared other furniture objects that might be hazardous (i.e., tables and other chairs) before attempting to move the desired chair. This style was beneficial to navigate around tighter areas or when colliding with other furniture was inevitable (Figure 4.5c). P1, for instance, explained their tendency to use Clear The Stage style to avoid "hitting other expensive or sentimentally valued furniture objects". This perhaps explains the higher



(a) Clustering

(b) Trying Things

(c) Clear The Stage

(d) One by One

Figure 4.5: Visual description of the four Arranging Styles participants used.

occurrences (17/74) of using the Clear The Stage style when arranging chairs around tables compared to the open space.

Trying Things was the third arranging style and related to instances where participants moved chairs to several positions before settling on a final position (Figure 4.5b). As the name suggested, participants used this style when the overall arrangement was not precisely pictured. P5, for example, used this style five times in the open space to organize what he referred to as an office-like group activity with a leader and 2 members. While P5 verbalised their intentions for the arrangement, P5 spent extra time trying out several positions and orientations of a specific chair—the group's leader chair. Participants used this style (14/74) 19% of the times.

Clustering, the least used style (6/74 8%), refers to instances where participants moved chairs closer to the final desired position (Figure 4.5a). This arrangement style is very similar to how people sort objects, yet the lower number of chairs and the tight lab space did not encourage users to use this arranging style much compared to the prior three.

To answer **RQ_Styles** we segmented the data from the previous code set, *Arranging Styles*, by the first code set, *Evidence of GenderMag Facets*, such that the styles are grouped by participants' GenderMag Facets (Table 4.5 and Table 4.6). The segmentation showed the UI must support all 4 arranging styles to be inclusive and One-by-One was the most utilized among gender problem-solving styles.

The last column (Styles%) in Table 4.5 shows all four *Arranging Styles* were used by at least one participant's with high-self efficacy (blue) and one participant's with low-self efficacy(orange). Out of the four *Arranging Styles* {One By One, Clear The Stage, Trying Things, and Clustering}, all participants used the One by One style, and participants with high self-efficacy (blue) were more likely to use several styles compared to participants with low self-efficacy (orange). Additionally, the table shows (4/5) participants with high self-efficacy (blue) were more likely to use the Clustering style compared to participants with low self-efficacy (orange). This results show that to maintain the inclusivity of the UI, all 4 arranging styles must be supported especially the one by one and clustering styles.

This trend is also noticeable with tinkering (blue) participants who used almost all arranging styles compared to participants who did not tinker (orange). Table 4.6 shows (2/3) participants who tinkered (blue) used all arranging styles, suggesting that the robot controlling interface must provide both low and high levels of controlling functionalities to be usable for diverse users. Additionally, the robot controlling UI must continue supporting two arranging styles, Clear The Stage and Trying Things, since

Table 4.5: The arranging styles grouped by evidence of participant's facet values: High and Low Computer Self-Efficacy

	Computer Self-Efficacy												Styles%
	High					Low							
	P1	**P9**	**P6**	**P2**	**P7**	**P8**	**P3**	**P4**	**P5**	**P12**	**P10**	**P11**	
One by One	*	*	*	*	*	*	*	*	*	*	*	*	100%, 100%
Clear The Stage	*	*	*	*			*	*	*	*			80%, 57%
Trying Things	*	*		*		*	*				*	*	60%, 57%
Clustering	*	*	*		*	*							80%, 14%

participants who tinkered (blue) were likelier to use them than participants who do not tinker (orange). This result echoes the description of GenderMag's Learning facet (recall Chapter 2), suggesting tinkering tendencies accompany exploring and experimenting with the technology.

Table 4.6: The arranging styles grouped by evidence of participant's facet values: With Tinkering and Without Tinkering learning Styles

	Learning Styles												Styles%
	With Tinkering			Without Tinkering									
	P1	**P9**	**P2**	**P8**	**P6**	**P3**	**P7**	**P5**	**P4**	**P12**	**P10**	**P11**	
One by One	*	*	*	*	*	*	*	*	*	*	*	*	100%, 100%
Clear The Stage	*	*	*		*	*		*	*	*			100%, 55%
Trying Things	*	*	*	*		*					*	*	100%, 44%
Clustering	*	*		*	*		*						66%, 33%

To answer **RQ_Performance** we looked at how the participants ranked the robotic and non-robotic chairs. Overall participants ranked the non-robotic chairs higher in terms of mobility, ease of use, and enjoyability (Figure 4.6). The lowest ratings of the robotic chairs were for mobility and usability. For example, P7 stated, *"The non-robotic chairs felt natural to push and pull. The robots moved successfully but required a bit of patience."* Perhaps the learning curve and 180 degree robotics wheels reduced the ranking of the robotic chairs compared to the non-robotic ones. While this result has nothing to do with inclusive software design (center of the thesis), it gave us ideas what needs to be fixed to make the robots more usable.

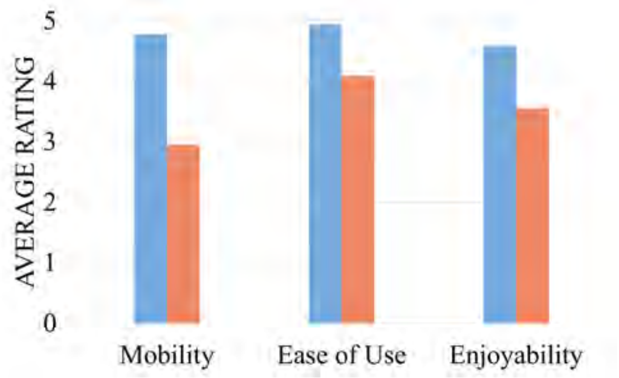
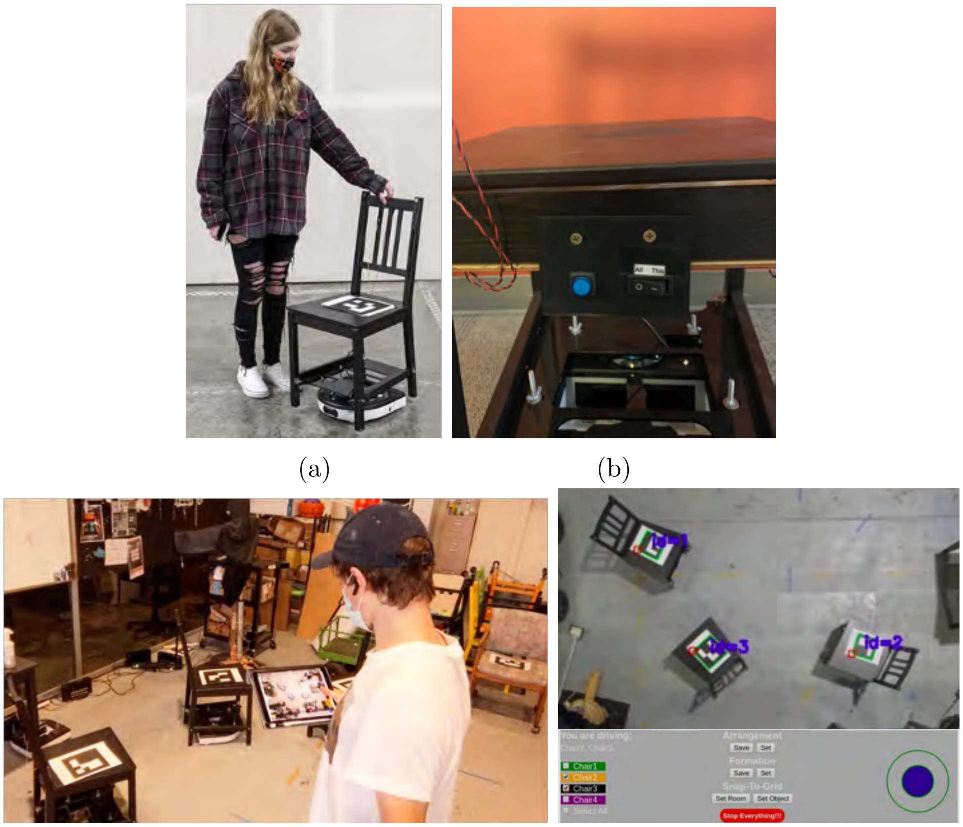


Figure 4.6: The ratings of the Non-Robotic vs. Robotic chairs in terms of Mobility, Ease of Use, and enjoyability. These ratings correspond to a 5-point Likert Scale averaged (mean) answers from all participants.

4.2 Design Remedies

To address the usability issues from the empirical study, we redesigned the physical interface (Figure 4.7a and Figure 4.7b) and added an additional mobile one (Figure 4.7c and Figure 4.7d) . In this section we will cover the design changes based on the usability issues and arranging styles.

For the new physical interface (Figure 4.7a and Figure 4.7b), we addressed the *physical challenge* issue by replacing the copper-based sensors with padded force sensors, providing realistic interactions and precise motions. Additionally, and to partially address the *memory challenges* issue, we removed sensor#7 and the LED light, and replaced them with a push button indicating if the chair is On or Off (Figure 4.7b). While this new button did not prevent users from recalling if the chair was On or Off, the button served as a visual cue and were externally consistent with home appliances and technology gadgets. We also added a switch button for users to toggle, but not that easily, between the move all or one feature (Figure 4.7b). The switch was intentionally different from the prior On/Off one to avoid *accidental activation*. Finally, we decided to address



(a) (b)

(c) (d)

Figure 4.7: (a)A lab member testing the updated physical interface (b) A close up of the updated physical interface showing the On/Off and toggling buttons. (c) A lab member testing the new digital interface. (d) A close up of the new mobile interface showing the controlling features for 3 ChairBots

the *invisible elements* issue by labeling each with a directional arrow. As for the new mobile interface, we addressed the *physical challenge* and *invisible elements* issues with a live overhead video stream and a visual joystick (Figure 4.7d). As for the *memory challenges* and *invisible elements* issues, we added toggling buttons at the bottom-left section to switch between chairs.

Considering the findings from the empirical study, we decided to keep the *One by One* arranging style in both interfaces (Figure 4.7). As for the remaining 3 arranging styles we added the ability to integrate geometric knowledge of the space to the mobile interface(Figure 4.7(d)). This new snap-to-geometry feature allowed users to command the ChairBots movement relative to the geometry of the room or its objects (e.g., parallel to a table). snap-to-geometry can be defined for room-centric geometries relative to the walls of the room, or furniture-centric geometries relative to an object in the scene. This feature would also help solving the *physical challenge* issue from the second code set.

4.3 Analytical Study

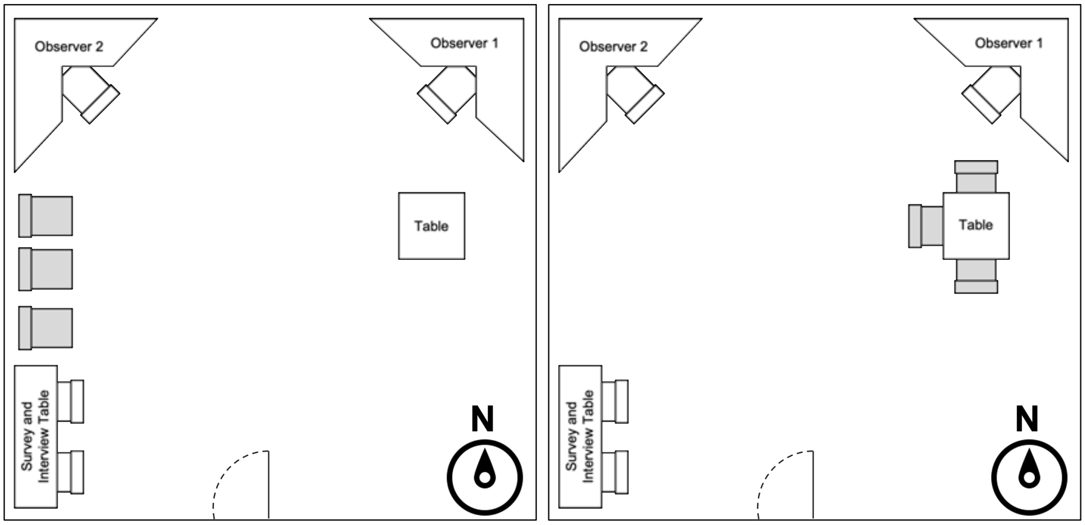
4.3.1 Methodology

We followed the empirical study with an analytical study using GenderMag's walkthrough to evaluate the inclusivity of the same robotic UI in Figure 4.1. We chose this sequence of studies because we noticed that we can't find all gender

We intentionally used the UI in Figure 4.1 instead of the updated one to validate the inclusivity biases. Additionally, in the previous empirical study, we used a subset of GenderMag's facets similar to [77] while in this analytical study, we run a GenderMag walkthrough using all facets with a team of three HCI researchers and a software engineer. All 3 HCI researchers were actively working with the ChairBots, and only one was familiar with using GenderMag's walkthrough to evaluate the inclusivity of WIMP UIs. The software engineer had HCI experience but was not experienced in programming or building robots.

As in all GenderMag walkthroughs, we (1) chose a persona and customized it, (2) chose a scenario and listed its subgoals, (3) walked through each subgoal and action from the perspective of the persona, a new user, and (4) reported the usability issues and inclusivity bugs. The persona we chose was *Abi* (recall Figure 2.2) since it represents cognitive problem-solving styles that are often overlooked by technology creators and we customized it to be a *business student without HCI or robotics experience.* The scenario plotted Abi as *a participant in an empirical study who wants to arrange a set of three ChairBots around a table to match a provided sketch* (Figure 4.8). This scenario had 2 subgoals and 8 actions, and covered all the features including moving all ChairBots at the same time.

We walkedthrough each subgoal and action by answering the standard GenderMag questions. Each subgoal had one question—*Will Abi have formed this subgoal as a step*



(a) (b)

Figure 4.8: Overhead view of a study area mimics the empirical study, showing the starting position (a) and the requested arrangement (b) form the Abi persona.

to their overall goal? (Yes/ Maybe/No, What Facets, and why?) The actions, on the other hand, had two paired questions. The first question for each action assesses if the Abi persona would understand if the action is needed to achieve the larger goal of arranging the ChairBots and was *Will Abi know what to do at this step? (Yes/ Maybe/No, What Facets, and why?)* The second question for each action assesses if the Abi persona would understand that they made progress toward achieving the larger goal (arranging the ChairBots) after they took action, and was *If Abi does the right thing, will Abi know that they did the right thing and is making progress toward their goal? (Yes/ Maybe/No, What Facets, and why?).* By answering these questions (i.e., running GenderMag's walkthrough), we identified a total of seven inclusivity bugs and one usability issue (Table 4.7).

4.3.2 Results

The analytical study using GenderMag's walkthrough revealed seven inclusivity bugs and a usability issue with the UI. The gender problem-solving bias percentage for this

Table 4.7: The subgoals and actions discussed in the analytical study, and the inclusivity issues that arose in each.

Item	Question ID	Revealed
Subgoal#1: Abi wants to move all the ChairBots close to the table	S1	Bug 1
Action#1.1: Abi wants to turn all the ChairBots ON	A1a	Bug 2
	A1b	–
Action#1.2: Abi wants to move the ChairBots forward as close as she can to the table	A2a	Bug 3
	A2b	–
Subgoal#2 Abi wants to move each ChairBot individually to the correct position	S2	Bug 4
Action#2.1: Abi wants to turn OFF Two ChairBots	A3a	Bug 5
	A3b	–
Action#2.2: Abi wants to adjust the position of the ChairBot	A4a	Bug 6
	A4b	–
Action#2.3: Abi wants to turn the ChairBot OFF	A5a	Bug 7
	A5b	–
Action#2.4: Abi wants to turn the reminding ChairBots ON	A6a	–
	A6b	–
Action#2.5: Abi wants to adjust the position of the ChairBot	A6a	–
	A6b	–
Action#2.6: Abi wants to turn the ChairBots OFF	A6a	Issue 1
	A6b	–

ChairBot's controlling UI was 38.9% (7/18) with an additional 5.6% (1/18) due to usability only. These percentages of gender and usability biases [36] can be described with the following equations:

$$GenderBias = \frac{\textit{Number of questions with Maybe \& No responses } \textbf{\textit{with}} \textit{ GM facet}}{\textit{Number of questions answered}}$$

$$UsabilityBias = \frac{\textit{Number of questions with Maybe \& No responses } \textbf{\textit{without}} \textit{ GM facet}}{\textit{Number of questions answered}}$$

As Table 4.8 shows, the two Abi facet values that were less supported in the physical interface were *computer self-efficacy* and *learning: by process vs. by tinkering*. Specifically, 75%(6/8) the inclusivity bugs were due to lack of supporting Abi's low *computer self-efficacy* and 62.5% (5/8) Abi's *process oriented learning* style. For example, the team reported that Abi's low self-efficacy and process oriented learning style might result in Bug 5 as Abi tries to turn two ChariBots Off. The team justified the answer stating that the there are plenty of information for Abi to process and recall, which might make the persona blame herself:

> “Abi may not remember everything from the tutorial and be confused about what the next step is. . . . Abi needs to process a lot of information about where to touch the chair for it to move which might make Abi blame herself.”

Table 4.8: The found inclusivity bugs segmented by GenderMag’s facet values for Abi. ∗: The team reported a bug tied to this facet.

Facet	Inclusivity Bugs							Bias%
	Bug 1	**Bug 2**	**Bug 3**	**Bug 4**	**Bug 5**	**Bug 6**	**Bug 7**	
Motivations	∗				∗		∗	37.5%
Computer Self-Efficacy	∗	∗	∗	∗	∗		∗	75%
Attitude Toward Risk		∗	∗				∗	37.5%
Information Processing Style								0%
Learning: by Process vs. by Tinkering	∗	∗			∗	∗	∗	62.5%

As Table 4.8 shows, the Abi facet values *motivation* and *attitude towards risk* were also not supported in the physical interface. Specifically, 37.5%(3/8) the inclusivity bugs were due to not of supporting Abi’s task oriented , *motivation* and risk-aversion about using unfamiliar technologies (*attitude towards risk*). For example, Bug 3 represents the state at which Abi wants to move the ChariBots forward closer to the table. The team reported that while motivated, Abi’s risk-aversion and low self-efficiency about operating the unfamiliar robots might prevent the persona from moving all 3 robot as the first action:

> “Abi have the motivation (wants to move the robots), but till this point Abi did not control any robots and unfamiliar with this ChairBots . . . This is the first time that Abi has to move them and Abi needs to move all three. As such, Abi may have a problem managing 3 chairs at once as a first step ”

The only usability issue the team reported represent the state at which Abi wants to turn a ChariBot Off as a last action. Here the persona, Abi, had finished the arrangement around the table and the last chair need to be turned Off. The team labeled this a usability bug since it depends on the persona’s ability to remember having to do it.

> "Abi may think the task is already done and forgets to turn the chair OFF.
> It all depends on her memory."

The outcomes of the GenderMag walkthrogh helped us identify not only *where* a bug can arias, but *why* that bug might arise—what specific problem solving facet(s) are not supported in the physical interface. Similar to the initial empirical study, the inclusivity bugs were due to lack of supporting divers levels of self-efficacy and/or learning styles. The next subsection cover the design remedies we developed to address the inclusivity bugs. Farther work on validating the design remedies is published in [79] as part of Stoddard's Masters thesis. We hope that these findings inform the design of similar robots and encourage other researchers to use GenderMag as an inclusive design method to evaluate beyond-WIMP UIs.

In conclusion, this chapter showed GenderMag's capabilities as an inclusive software design method in evaluating and redesigning two beyond-WIMP UIs. We investigated the inclusivity requirements for the UIs in the context of robotic furniture arrangements. The next chapter investigates whether and how a family of inclusive software design methods, similar to GenderMag, can be used analytically to evaluate the multidimensionality of a Hands-Free IDE.

Chapter 5: Investigating Whether and How Analytical Inclusive Software Design Methods Can Be Used to Evaluate a Hands-Free IDE for Humans Multidimensional Identities

5.1 Introduction

The previous chapter demonstrated the capability of an inclusive software design method to evaluate and redesign a beyond-WIMP UI; however, inclusivity researchers have been voicing their concerns about the penalties of designing for a single demotion of human ever intersecting identities [7, 74, 64]. A 2018 meta-review by Schlesinger et al. [74] found that most literature about inclusive design considers only a single user identity, such as gender or socio-economic status (SES). Such one-dimensional approaches (using GenderMag in the previous Chapter) have been impactful, but they have yet to be able to serve users multiple and intersecting identities [20, 63, 62]. A well-known example is the face recognition failure rate for Black women, in which facial recognition systems achieved reasonable accuracy when predicting for men and for women, and for people with both darker and lighter skins—but these one-dimensional evaluations did not reveal the disproportionate lack of support for darker-skinned women [7, 6]. To expand behind robotics and generalize our inclusivity lens behind gender to humans multidimensional identities, we run an additional study. In this study, we investigated whether and how practitioners can use InclusiveMag's generated analytical inclusive software design methods to evaluate the multidimensionality of a Hands-Free IDE.

We selected the InclusiveMag meta-method (recall Chapter 2) to generate new analytical methods for the multidimensional population—low-SES immigrant women. Since InclusiveMag is about supporting *both* endpoints of facet value spectra, we were also interested in this population's multidimensional "opposite": high-SES nonimmigrant men. These multidimensional populations were at the intersection of three diversity dimensions: SES, Immigration, and Gender.

$$\mathcal{D}im = \mathcal{SES} \qquad \mathcal{D}im' = \mathcal{I}mmigration \qquad \mathcal{D}im'' = \mathcal{G}ender$$

Using these three dimensions and their intersection, we investigated:

$$\text{SESImmigrationGenderMag} \subseteq \text{SESMag} \cup \text{ImmigrationMag} \cup \text{GenderMag}$$

To perform this investigation, we needed four InclusiveMag-derived methods—one for each of the SES, immigration, and gender dimensions and one for the multidimensional SES+immigration+gender dimension. We needed software to evaluate, which in our case was a prototype of a Hands-Free IDE. We also needed evaluations of this prototype based on the products of InclusiveMag-derived methods. Finally, we needed HCI researchers, designers, and practitioners to create and use these items.

To this end, our specific research questions were:

RQ1_Bugs: Can the bugs found analytically by a multidimensional HCI practitioner team also be found analytically by an HCI practitioner team for a component diversity dimension?

RQ2_Reasoning: If a multidimensional HCI practitioner team analytically finds a bug, will the facets they use also be used by an HCI practitioner team for a component diversity dimension, to find that same bug?

The utilization of InclusiveMag-derived methods in this chapter revealed that as teams of HCI practitioners evaluated the multidimensionality of the Hands-Free IDE, the inclusivity bugs that the single-dimension teams found were "mostly" a subset of the ones found by the multidimensional teams. Additionally, in considering the reasoning behind the teams' bug-finding, a subset relation was "almost always" satisfied, showing that at least one single-dimension team used the same facet value as the corresponding multidimensional team. Our findings show that HCI practitioners can indeed use the analytical inclusive design methods generated using InclusiveMag to evaluate the multidimensionality of beyond-WIMP UIs.

5.2 Methodology

We conducted a mixed-method empirical study with 10 HCI teams (24 participants). We recruited the 24 participants from current and former offerings of a 10-week, advanced HCI (Inclusive Design) course for graduate and 4^{th} year undergraduate HCI students. Being in a course gave participants a pre-existing reason to work for weeks using InclusiveMag, namely, getting a "good enough" course grade per their standards. We chose this particular course because in the course, students learn inclusive design skills hands-on using InclusiveMag (recall Figure 2.1). Specifically, over the 10-week course, they use InclusiveMag's Steps 1–2 to create their own inclusive design methods and then use the methods they created to evaluate the inclusivity of prototypes they are designing (InclusiveMag's Step 3). Thus, all participants were familiar with the InclusiveMag meta-method.

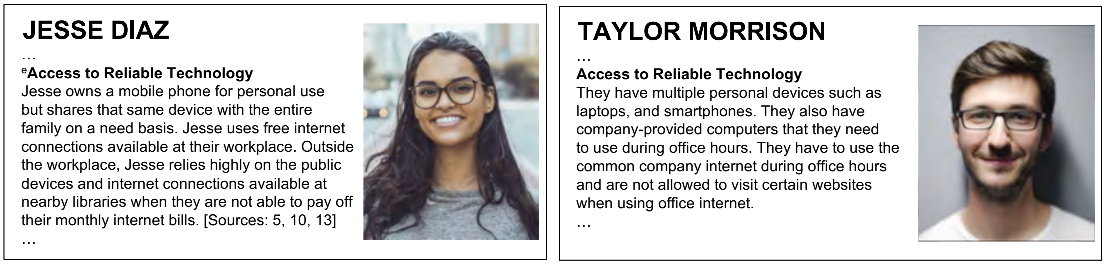
The 24 participants acted in three roles: *researcher-participants* to create InclusiveMag-derived methods for different diversity dimensions, *practitioner-participants* to use those methods to evaluate a prototyped Hands-Free IDE, and a *designer-participant* to provide that prototype and an appropriate workflow for using it. For the *researcher-participants* we followed a classic case study methodology [70], in which there were no controls—participants did whatever they did in their own context to research the diversity dimensions. The *practitioner-participants* used the products of the *researcher-participants* following specialized cognitive walkthroughs to reason their way through an evaluation of the hands-free IDE.

5.2.1 The researcher-participants' work

The researcher-participants' mission was not to reveal answers to the research questions. Rather, it was to provide, in an ecologically valid manner, the methods by which the practitioner-participants would search for inclusivity bugs in the IDE, to enable us to answer RQ1_Bugs and RQ2_Reasoning. Toward that end, the researcher-participants (9/24 of the participants) worked in two teams (Immigrant-R and Intersect-R) of 4-5 people each, in which they followed InclusiveMag's Steps 1–2 (recall Figure 2.1) to construct analytical methods with facets and research-based personas for the diversity dimensions of Immigration and SESImmigrantGender, respectively. The teams spent

8-10 weeks of the course doing this work, which is as much time or more than many UX researchers can spend on population research [28, 80].

To carry out InclusiveMag's Steps 1–2 (recall Figure 2.1), Teams Immigrant-R and Intersect-R performed extensive research (however they saw fit, per case study methodology [70]) so as to create whatever facet types emerged from their research. Teams Immigrant-R and Intersect-R used various research methods, including interviews; published blogs and youtube-based interviews/documentaries with/about their populations; reviews of academic literature; and drawing upon the lived experiences of team members who self-identified as members of the population they were investigating. In addition, since Team Intersect-R was researching a multidimensional population, they informed their research with Team Immigrant-R's findings and with the GenderMag and SES-Mag foundational research. Team Immigrant-R's work resulted in an ImmigrationMag method whose facets formed the core of personas Ahava and Bernadette; likewise, Team Intersect-R's work resulted in an SESImmigrationGenderMag method with personas Jesse and Taylor (Figure 5.1).



JESSE DIAZ

…

[e]**Access to Reliable Technology**

Jesse owns a mobile phone for personal use but shares that same device with the entire family on a need basis. Jesse uses free internet connections available at their workplace. Outside the workplace, Jesse relies highly on the public devices and internet connections available at nearby libraries when they are not able to pay off their monthly internet bills. [Sources: 5, 10, 13]

…

(a)

TAYLOR MORRISON

…

Access to Reliable Technology

They have multiple personal devices such as laptops, and smartphones. They also have company-provided computers that they need to use during office hours. They have to use the common company internet during office hours and are not allowed to visit certain websites when using office internet.

…

(b)

Figure 5.1: Excerpts from personas that Team Intersect-R created to represent (a) Low-SES Immigrant Women and (b) High-SES Nonimmigrant Men. (The sources and footnotes refer to that team's internal documents, not sources/footnotes in this work.)

The other two diversity dimensions in the study were gender and SES. For the gender dimension, we used GenderMag [9], whose ecological validity stems from its use by practitioners at several organizations (e.g., [8, 35, 84]). For the SES dimension, we used Hu et al.'s facets [39], which had been created by a team of HCI researchers that included several with professional HCI experience; we then created personas based on those facets for purposes of this study.

Researcher Team	**Diversity Dimension**	**Facets**	**Personas**	**Practitioner Team**
Hu et al. [39]	SES	Access to Reliable Technology (Access); Technology Self-Efficacy (SE); Technology Risks (Risks); Technology Privacy/ Security(Priv.); Perceived Control & Attitude Toward Authority (Control); Communication Literacy/Education/Culture (Commun.)	Low-SES: Dav, High-SES: Fee	SES-P1, SES-P2
Immigrant-R	Immigration	Level of English Language Proficiency (Commun.); Willingness to Accept Help (Accept Help); Mental Health/Past Trauma (Mental Health); Comfort using Technology (SE)	Immigrant: Ahava, Nonimmigrant: Bernadette	Immigrant-P1, Immigrant-P2
Burnett et al. [26]	Gender	Motivations (Mot.); Computer Self-Efficacy (SE); Attitude Towards Risk (Risks); Information Processing Style (Info. Proc.); Learning: by Process vs. by Tinkering (Learn)	Women: Abi, Men: Tim	Gender-P1, Gender-P2
Intersect-R	Immigration, SES, Gender	Communication Literacy and Culture (Commun.); Access to Reliable Technology (Access); Risks, Privacy, Security (Risks); Perceived Control and Attitude Toward (Control); Information Processing Strategies (Info. Proc.)	Immigrant Low-SES Women: Jesse, Nonimmigrant High-SES Men: Taylor	Intersect-P1, Intersect-P2

Table 5.1: Which participants did what. The first four columns list the research-participants' teams and the products they produced. The last column lists the practitioner-participants teams that used the InclusiveMag-derived products.

All four analytical methods set the facets and personas into a specialized cognitive walkthrough process. Table 5.1 summarizes all four diversity dimensions used in the study, which participant teams created them, what facets they included, and which personas brought those facets to life.

The particular definitions of the facets frequently overlapped. Table 5.1 and Table 5.2 together show all the facets the Intersect-R team ultimately created corresponded to a facet for at least one single-dimension population, although their terminology sometimes varied. For example, the Intersect-R facet *Communication Literacy and Culture (Commun.)* was similar to the SES facet *Communication Literacy/Education/Culture (Commun.)* and the Immigrant-R facet *Level of English Language Proficiency (Commun.)*. All three of these *Commun.* facets (light-blue cells in Table 5.2) covered the persona's ability to communicate using cultural references and jargon, read comprehensively, and speak English as a primary or second language.

Researcher Team	**Facets**						
Hu et al. [39]	Commun.	Access	Risks Priv.	SE Control			
Immigrant-R	Commun.			Accept Help SE			Mental Health
Burnett et al. [26]			Risks	SE	Info. Proc. Mot.	Learn	
Intersect-R	Commun.	Access	Risks	Control	Info. Proc.		

Table 5.2: The facets the researcher Teams created for each diversity dimension. Each row shows the matched facets across dimensions.

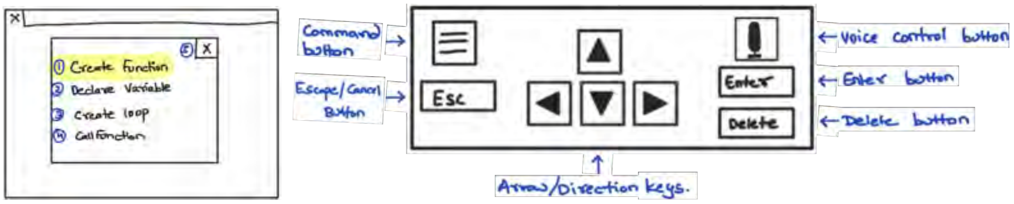
5.2.2 The practitioner-participants' and designer-participant's work

Given the analytical methods they inherited, the mission of the 17 practitioner-participants (three of whom had also been researcher-participants) was to apply their respective methods and evaluate the inclusivity of the IDE. Their work processes and products were the data we used to answer RQ1_Bugs and RQ2_Reasoning. Thus, the practitioner-participants worked in 8 teams (2-3 people per team), using the researcher-participants'

InclusiveMag-derived products to analytically evaluate the personas' user experiences with the hands-free IDE prototype. The designer-participant had been a UI designer of that hands-free IDE prototype, and served as the expert on the prototype and its intended workflow. Table 5.1's rightmost column lists the teams of practitioner-participants with the persona they worked with to perform their evaluations.

The practitioner-participants teams' evaluations of the hands-free IDE prototype (recall Step 3 in Figure 2.1) were cognitive walkthroughs specialized to the facets of the diversity dimension. For these cognitive walkthroughs, they walked through an action sequence of the prototype from the perspective of their personas and facets, asking before each question whether their persona would do that action, and after each action whether their persona would feel like they were making progress.

The action sequence, which the designer-participant had provided, was: (1) press the "command" button on the foot-keyboard, (2) press the "voice" button on the keyboard, (3) say "1," (4) press the "enter" button on the keyboard, (5) say "1," (6) say "AddFunction" to name the function, and (7) press the "enter" button on the keyboard. This sequence of seven actions allowed users to create a function using voice commands. Figure 5.2a shows the IDE's screen early in this sequence, after a user says "1" to select the "create function" option. Figure 5.2b shows the foot-keyboard input device.



(a) (b)

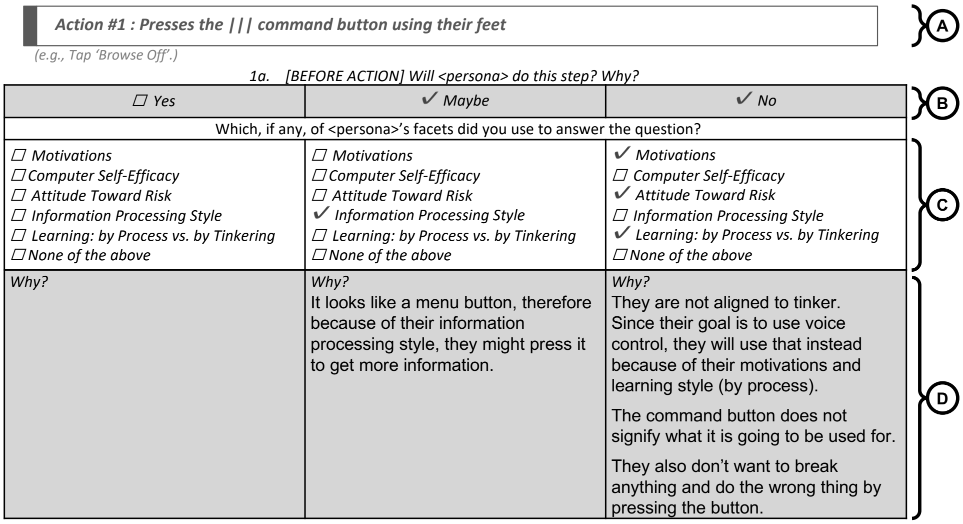
Figure 5.2: Excerpts of the Hands-Free IDE. (a) The hands-free IDE screen displaying a menu with the "Create Function" option selected. (b) The foot-keyboard which enabled users to navigate cursor position, push buttons, tell the system to listen for voice commands, etc.

The designer-participant also customized all personas' background information to ensure appropriateness for IDE usage, as follows:

> "[Persona] is 17 years old. [Persona] is in their last year of high school. [Persona] is living with their parents. [Persona] is comfortable with technology,

and their hobby of coding has led them to want to study computer science in their dream college."

The teams of practitioner-participates wrote down their evaluations for each action in the sequence, using a walkthrough form that consisted of 7 pre-action and 7 post-action questions. The pre-action questions were "Will [persona] do this step? (yes/maybe/no, what facets, and why)." The post-action questions were "If [persona] does the right thing, will they know that they did the right thing and is making progress toward their goal? (yes/maybe/no, what facets, and why)." For example, Figure 5.3 shows how Team Gender-P1 answered Action 1's pre-action question.



Action #1 : Presses the ||| command button using their feet

(e.g., Tap 'Browse Off'.)

1a. [BEFORE ACTION] Will <persona> do this step? Why?

☐ *Yes*	✓ *Maybe*	✓ *No*
Which, if any, of <persona>'s facets did you use to answer the question?		
☐ *Motivations* ☐ *Computer Self-Efficacy* ☐ *Attitude Toward Risk* ☐ *Information Processing Style* ☐ *Learning: by Process vs. by Tinkering* ☐ *None of the above*	☐ *Motivations* ☐ *Computer Self-Efficacy* ☐ *Attitude Toward Risk* ✓ *Information Processing Style* ☐ *Learning: by Process vs. by Tinkering* ☐ *None of the above*	✓ *Motivations* ☐ *Computer Self-Efficacy* ✓ *Attitude Toward Risk* ☐ *Information Processing Style* ✓ *Learning: by Process vs. by Tinkering* ☐ *None of the above*
Why?	*Why?* It looks like a menu button, therefore because of their information processing style, they might press it to get more information.	*Why?* They are not aligned to tinker. Since their goal is to use voice control, they will use that instead because of their motivations and learning style (by process). The command button does not signify what it is going to be used for. They also don't want to break anything and do the wrong thing by pressing the button.

Figure 5.3: Team Gender-P1's Walkthrough Form for Action 1's pre-action question. **A**: The action the team is evaluating. **B**: The answers (yes/maybe/no) to the pre-action-1 question. (Since not all members of Team Gender-P1 agreed on the answer, they selected both Maybe and No.) **C**: The facets the team members used to decide their answers. **D**: What Team Gender-P1 wrote about their reasoning.

5.2.3 What counted a bug

We declared a pre- or post-action to be a *bug* if anyone on the team identified a problem— i.e., if Maybe or No had been checked off (Figure 5.3), even if Yes had also been checked off. As in other works (e.g., [10] and the previous Chapter), we defined a bug as also being an *inclusivity bug* if the team wrote that the bug was tied to one of the persona's facet values, because that would suggest that the bug would arise disproportionately often for people with that facet value. Note that there could only be one bug per pre- or post-action; multiple explanations or difficulties surrounding that pre- or post-action were considered part of the same bug.

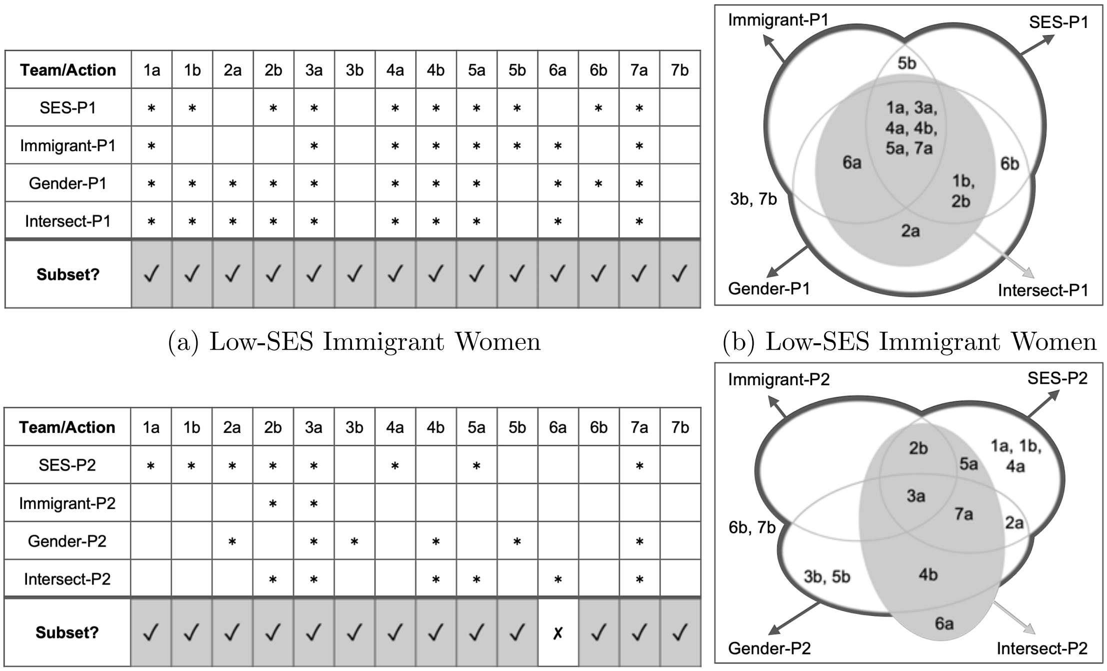
The teams had two ways they could mention facet values during their walkthroughs. First, they could simply check off a facet (Figure 5.3C). Alternatively, they could write about it in the free-text part of the form (Figure 5.3D). To identify facet values mentioned in the 109 free-form text entries the teams had made, we used qualitative coding. Two researchers independently coded 20% of the free-form text data. Their agreement level was 97.78% (Jaccard index), indicating a very high level of agreement. Given this agreement level, one researcher coded the remaining data.

5.3 Results

Given the use of InclusiveMag-derived methods, **RQ1_Bugs** asks whether the bugs the multidimensional practitioner-participants teams found analytically were the same bugs that at least one of the single-dimensional teams found analytically. The answer in practice was "mostly." The single-dimension practitioner-participants teams together were able to analytically find all except one of the bugs that the multidimensional team found analytically (Figure 5.4).

Building off the previous subset relation (SESImmigrationGenderMag $\subseteq$ SESMag $\cup$ ImmigrationMag $\cup$ GenderMag), we used three criteria to determine whether a bug the multidimensional team found was an element of the union of the bugs the single-dimension teams had found for that pre- or post action question.

Criterion 1: If at least one single-dimension team found the same bug as the corresponding multidimensional team, then the subset relationship holds. 15 of the total 28 analytical questions satisfied Criterion 1 (Figure 5.4). For example in column 2a of Figure 5.4a, both Teams Gender-P1 and Intersect-P1 reported a bug here; since Intersect-P1 found a bug that at least one of the other teams found, Criterion-1 is satisfied. Column 2a represents the state of the prototype just before the user needs to press the "voice" button. Team Gender-P1 anticipated that their persona, Abi, would need to tinker around to find this voice button; but tinkering is not in line with Abi's process-oriented learning



Team/Action	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b
SES-P1	*	*		*	*		*	*	*	*		*	*	
Immigrant-P1	*				*		*	*	*	*	*		*	
Gender-P1	*	*	*	*	*		*	*	*		*	*	*	
Intersect-P1	*	*	*	*	*		*	*	*		*		*	
Subset?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(a) Low-SES Immigrant Women

Immigrant-P1
SES-P1
5b
1a, 3a, 4a, 4b, 5a, 7a
6a
6b
1b, 2b
3b, 7b
2a
Gender-P1
Intersect-P1

(b) Low-SES Immigrant Women

Team/Action	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b
SES-P2	*	*	*	*	*		*		*				*	
Immigrant-P2				*	*									
Gender-P2			*		*	*		*		*			*	
Intersect-P2				*	*			*	*		*		*	
Subset?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓

(c) High-SES Nonimmigrant Men

Immigrant-P2
SES-P2
2b
5a
1a, 1b, 4a
3a
7a
2a
6b, 7b
3b, 5b
4b
6a
Gender-P2
Intersect-P2

(d) High-SES Nonimmigrant Men

Figure 5.4: RQ1 results: (Top): For all 14 analytical questions, the multidimensional Immigrant Low-SES Women's team found bugs whenever the single-dimensional teams did. (Bottom): This was also true for 13/14 of the Nonimmigrant High-SES Men's analytical questions.
(Left): Bugs each team reported for each of the 7 pre-action questions (a's in the table's columns) and post-actions questions (b's in the table's columns). $*$: The team reported a bug in the analytical question. (Right): The extent to which the multidimensional teams' findings (shaded) were a subset of the union of the bugs the single-dimensional teams found (thick black outline).

style:

> Gender-P1: . . . [Abi] is not a tinkerer, would not like to press the button. [Abi] might want to press "ESC" to go back and look for more information/help.

Team Intersect-P1 also anticipated a bug; they were not sure if their persona, Jesse, would be confident in associating the microphone icon with voice commands:

> Intersect-P1: This button looks familiar with other popular applications. But not entirely sure, if [Jesse] may have the idea about a feature that has something to do with voice.

Criterion 2: If at least one single-dimension team found a bug but the corresponding multidimensional team missed it, the subset relationship still holds. 8 of the total 28 analytical questions fulfilled Criterion 2. In column 5b in Figure 5.4c, for example, Team Gender-P2 reported a bug, but Team Intersect-P2 did not. This meets Criterion 2 because here the union of the three single-dimension teams have done at least as well at bug-finding as the multidimensional team. Column 5b represents the state of the prototype after the user has said "1." Team Gender-P2 found that the persona, Tim, would face an issue understanding the IDE's response:

> Gender-P2: . . . It is not certain that [Tim] will know [they are making progress toward their goal] because [Tim] needs to try other options to understand the [IDE].

Team Intersect-P2 did not find a bug; they believed the screen would be self-explanatory for their persona, Taylor:

> Intersect-P2: The options displayed on the screen are self-explanatory for [Taylor] as they are comfortable with using technology and technological terms. . .

Team Intersect-P2 and Team Gender-P2 findings still satisfy the subset relationship: Gender-P2 did "*at least* as well" as Intersect-P2 at bug-finding.

Criterion 3: If all teams agreed that there was no bug with a pre- or post action question, the subset relationship holds. 4 of the 28 analytical questions fulfilled Criterion 3 (Figure 5.4). Column 7b in Figure 5.4c is one example: neither Team Immigrant-P2 nor Intersect-P2 found a bug. 7b represented the state of the prototype just after the user had pressed enter to create the function. Immigrant-P2 felt their persona, Bernadette, would know they had been successful:

> Immigrant-P2: [Bernadette]'s initial goal was to create a function with voice control. Once [Bernadette] sees this screen, [Bernadette] will know the function has been created successfully.

Team Intersect-P2 came to the same conclusion, since their persona, Taylor, is used to working with IDEs:

> Intersect-P2: As [Taylor] is comfortable with using technology and have high perceived control over technology, [Taylor] will feel good about [accomplishing the task].

The team's agreement about the absence of a bug with 7b satisfies the subset relationship.

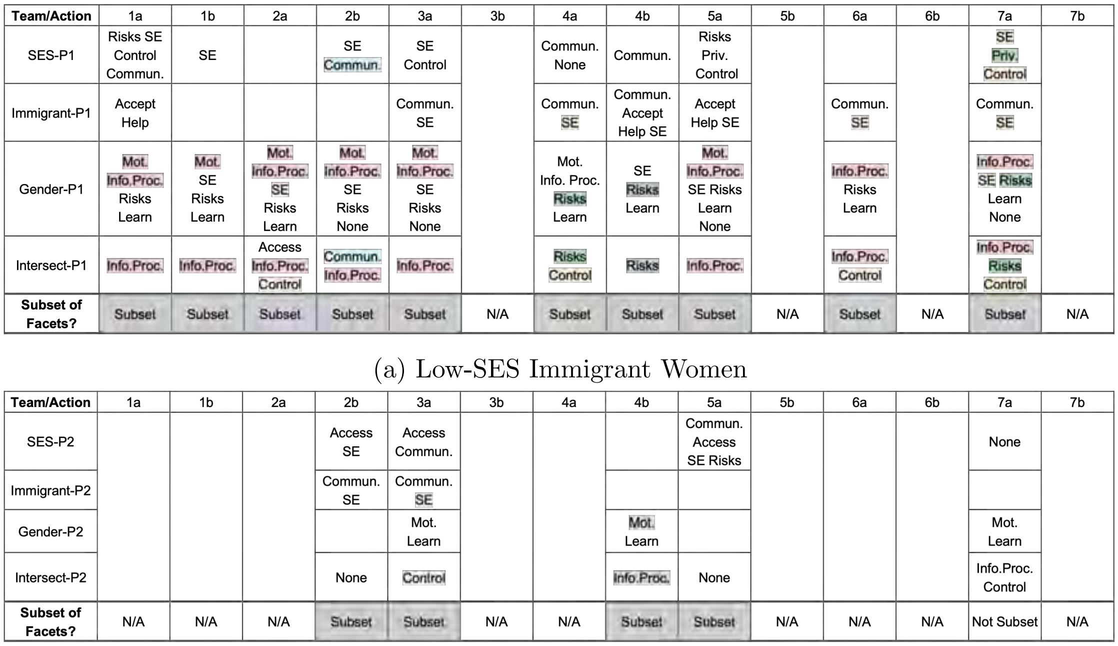
If none of the these criteria were met, the subset relationship did *not* hold. Only one of the 28 analytical questions, column 6a in Figure 5.4c did not fulfill the criteria. 6a represented the state of the prototype just before the user had to say "AddFunction" to name the function. Team Intersect-P2 thought their persona would use other terms instead:

> Intersect-P2: maybe, [Taylor] might even say different words related to this like 'create function', 'begin function', . . . etc because as [Taylor] use different technologies, [Taylor] are used to seeing different tech words across platforms

In considering the reasoning behind the teams' bug-finding, **RQ2_Reasoning**, we considered whether the multidimensional teams' use of *facet values* were a subset of the facets used by the combined related single-dimension teams. Here we only used the previous Criterion 1 aspect of the subset relation. The teams who considered the underrepresented populations (P1 teams) almost always satisfied this subset relation (10/10 times) in their use of the facet values (Figure 5.5a). The P2 teams did so the majority of the time too (4/5 times) in their use of the facet values (Figure 5.5b).

For example, consider column 7a in Figure 5.5a, for which all P1 teams had identified an inclusivity bug. Column 7a represents the state of the prototype just before pressing the "enter" button to create the addition function. The three single-dimension teams together associated the bug with a total of five facets: *Learn*, *Commun.*, *Info.Proc.*, *Risks*, and *Control*. The Intersect-P1 team associated the bug with a subset of these, *Info.Proc.*, *Risks*, and *Control*. Team Intersect-P1 used these three facets to reason that the persona, Jesse, had performed several actions but was still uncertain they were making progress toward creating the function:

> Intersect-P1: [Jesse] made a lot of progress. Without concrete clue it will be very tough to make [Jesse] confident.



Team/Action	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b
SES-P1	Risks SE Control Commun.	SE		SE Commun.	SE Control		Commun. None	Commun.	Risks Priv. Control				SE Priv. Control	
Immigrant-P1	Accept Help				Commun. SE		Commun. SE	Commun. Accept Help SE	Accept Help SE		Commun. SE		Commun. SE	
Gender-P1	Mot. Info.Proc. Risks Learn	Mot. SE Risks Learn	Mot. Info.Proc. SE Risks Learn	Mot. Info.Proc. SE Risks None	Mot. Info.Proc. SE Risks None		Mot. Info. Proc. Risks Learn	SE Risks Learn	Mot. Info.Proc. SE Risks Learn None		Info.Proc. Risks Learn		Info.Proc. SE Risks Learn None	
Intersect-P1	Info.Proc.	Info.Proc.	Access Info.Proc. Control	Commun. Info.Proc.	Info.Proc.		Risks Control	Risks	Info.Proc.		Info.Proc. Control		Info.Proc. Risks Control	
Subset of Facets?	Subset	Subset	Subset	Subset	Subset	N/A	Subset	Subset	Subset	N/A	Subset	N/A	Subset	N/A

(a) Low-SES Immigrant Women

Team/Action	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b
SES-P2				Access SE	Access Commun.				Commun. Access SE Risks				None	
Immigrant-P2				Commun. SE	Commun. SE									
Gender-P2					Mot. Learn			Mot. Learn					Mot. Learn	
Intersect-P2				None	Control			Info.Proc.	None				Info.Proc. Control	
Subset of Facets?	N/A	N/A	N/A	Subset	Subset	N/A	N/A	Subset	Subset	N/A	N/A	N/A	Not Subset	N/A

(b) High-SES Nonimmigrant Men

Figure 5.5: RQ2 results: Whenever an multidimensional team found an inclusivity bug, were the facets they used the same as those the single-dimension team used? (Facet colors show facets similar to the multidimensional team's, as per Table 5.2.) (Top): Yes for Low-SES Immigrant Women in 10/10 cases. (Bottom): Yes for High-SES Nonimmigrant Men in 4/5 cases.

Team Immigrant-P1 attributed the bug to two facets, *Commun.* and *SE*, the latter similar to Team Intersect-P1's *Control* facet. They reasoned that Ahava's low self-efficacy might make the persona hesitant about switching input modalities:

> Immigrant-P1: ...if [Ahava] are using voice control, [Ahava] might be hesitant to go back and forth between the foot and voice commands.

Team Gender-P1 used three similar facets to Team Intersect-P1, (*Info. Proc.*, *SE*, and *Risks*), as well as an additional facet, *Learn*, to highlight the issue of frequent switching between the input modalities:

> Gender-P1: [Abi] will use VOICE, and probably say ”2” instead of going back to the keyboard. [Abi] will use the same process as before.

A similar facet-subset relation held for the P2 teams. In 4 out of 5 cases, the facets Team Intersect-P2 used for bugs were a subset of the union of facets the other teams used (SES-P2, Immigrant-P2, and Gender-P2) (Figure 5.5b). For example, Teams Intersect-P2 and SES-P2 found a bug in column 2b, which represents the state of the prototype after the user pressed the "voice" button on the keyboard. Team Intersect-P2 found the display screen confusing to the persona but did not associate this bug with any facets:

> Intersect-P2: The numbers against each option might be confusing for [Taylor]. [Taylor] might not know how to proceed next.

Team SES-P2 agreed with Team Intersect-P2 about the bug—but used facets to find it. Specifically, Team SES-P2 used *Access* and *SE* to reason that the persona, Fee, would find the display screen confusing:

> SES-P2: The number label might help [Fee] to guess that they have to say the number to take the action. However, it's still not clear since [Fee] might expect to see an audio icon or a keyboard prompt or even a sound (similar to Alexa listening)

Teams Intersect-P2 and SES-P2 findings satisfy the subset relation because Team Intersect-P2's judgment of "no facets" (i.e. the empty set) is a subset of SES-P2's use of *Access* and *SE*.

In conclusion, this chapter showed that as teams of HCI practitioners evaluated the multidimensionality of the prototyped Hands-Free IDE, the inclusivity bugs that the single-dimension practitioner-participants teams found were "mostly" a subset of the ones found by the multidimensional teams. Additionally, in considering the reasoning behind the teams' bug-finding, a subset relation was "almost always" satisfied, showing that at least one single-dimension team used the same facet value as the corresponding multidimensional team. Our findings show that HCI practitioners can indeed use the analytical inclusive design methods generated using InclusiveMag to evaluate the multidimensionality of beyond-WIMP UIs.

Chapter 6: Concluding Remarks

This dissertation investigates if and how inclusive software design can be borrowed from Human-Computer Interaction to design and evaluate beyond-WIMP (Windows, Icons, Menus, and Pointers) user interfaces. We choose beyond-WIMP because ubiquitous computing systems that are part-WIMP or beyond-WIMP, such as autonomous delivery robots, are becoming a reality and replacing keyboard-plus-mouse user interfaces; however, little has been done to address their software inclusivity. We also choose inclusive software design because it addresses digital exclusion biases and advocates for supporting a more comprehensive range of users. Toward that end, this dissertation proposes the following thesis: *Inclusive software design and its methods that have been devised for WIMP UIs can be used effectively to design and evaluate beyond-WIMP UIs.*

This dissertation comprises a three-stage investigation of inclusive software design for three beyond-WIMP UIs: A social robot, a multiple robots controller, and a Hands-Free Integrated Development Environment (IDE).

The first stage (Chapter 3) explored the applicability of inclusive software design as an evaluation approach with a social robot that interacts with diverse people in diverse places. More specifically, we studied the inclusivity requirements in terms of culture and situational factors that predicted diverse peoples' likelihood to help and care for a robot in 6 cafes around a University Campus. The results from the study indicated that cultural and situational factors such as the overall mood (aka social atmosphere) of a cafe and the robot approaching styles predicted participants' likelihood to help and care for the robot. The stage showed that software inclusivity is relevant to beyond-WIMP user interfaces and that future help-seeking can benefits from our four inclusive design implications.

In the second stage (Chapter 4), we examined the use of a particular inclusive software design method to evaluate and redesign a multiple robots controller. The method we used was GenderMag, an evaluation method that reports the percentage of biases and inclusivity issues of user interfaces. Similar to [77, 3], in this stage, we used GenderMag empirically in a lab study where we asked participants to use the touch-based multiple robots controller to arrange robotic chairs. We then used GenderMag analytically with design experts to evaluate the inclusivity of the controller. Based on GenderMag empirical and analytical evaluations, we redesigned the touch-based controller and developed an additional mobile UI. The results from the stage indicated that inclusive software design methods, such as GenderMag, can be used analytically to evaluate and redesign beyond-WIMP UIs.

In the third stage(Chapter 5), we investigated if and how inclusive software design methods can be used analytically to evaluate the inclusivity of a Hands-Free IDE. Unlike the previous two stages, this stage recognized users' multidimensional identities and evaluated the Hands-Free ID from the perspective of three dimensions: socioeconomic status (SES), Immigration Status, and Gender. In this stage, we used the products of the InclusveMag family of inclusive software design methods: SESMag, ImmigrationMag, GenderMag, and SESImmigrationGenderMag. More specifically, we asked eight teams of practitioners to evaluate the inclusivity of the Hands-Free IDE analytically and compared whether the inclusivity bugs the multidimensional practitioner-participants teams found analytically using SESImmigrationGenderMag were the same bugs that at least one of the single-dimensional teams found using SESMag, ImmigrationMag, or GenderMag. The results from the stage indicated that inclusive software design methods, such as the family of InclusveMag, can be used analytically to evaluate the inclusivity of beyond-WIMP UIs and that the inclusivity bugs found by using single-dimensional methods are a subset of the bugs found by a multidimensional one.

This dissertation contributes several design implications, new technology development, and empirical contributions to inclusive design, Human-Computer Interaction, and Human-Robot Interaction. In addition to the four inclusive design implications, the first stage (Chapter 3) reveals how inclusive design might be applied to evaluate social robots' inclusivity empirically. The second stage of this dissertation contributes two novel multiple robot controllers touch-based UI and a mobile one (Chapter 4). Stages 2 and 3 (Chapter 4 and Chapter 5) also contribute empirical evidence regarding how inclusive software design methods such as GenderMag can be used analytically to evaluate and (re)design beyond-WIMP UIs for diverse users.

Bibliography

[1] Nils Backhaus, Patricia H Rosen, Andrea Scheidig, Horst-Michael Gross, and Sascha Wischniewski. “somebody help me, please?!” interaction design framework for needy mobile service robots. In *Proceedings of the 2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, volume 2018, pages 54–61, Genoa, Italy, 2018. IEEE.

[2] Markus Bajones, Astrid Weiss, and Markus Vincze. Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user. *arXiv preprint arXiv:1606.02547*, 1(AISB-NFHRI/2016/02), 2016.

[3] Sogol Balali, Ross T. Sowell, William D. Smart, and Cindy M. Grimm. Privacy concerns in robot teleoperation: Does personality influence what should be hidden? In *Proceedings of the 2019 International Conference on Social Robotics*, ICSR ’19, page 719–729, Berlin, Heidelberg, 2019. Springer-Verlag.

[4] Jeanne H Ballantine, Keith A Roberts, and Kathleen Odell Korgen. *Our social world: Condensed: An introduction to sociology.* SAGE Publications, 2017.

[5] Joey Benedek and Trish Miner. Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, 2002.

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, USA, 2018. PMLR.

[7] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers.* PhD thesis, Massachusetts Institute of Technology, 2017.

[8] Margaret Burnett, Robin Counts, Ronette Lawrence, and Hannah Hanson. Gender hcl and microsoft: Highlights from a longitudinal study. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 139–143. IEEE, 2017.

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

[10] Margaret M. Burnett, Anicia Peters, Charles Hill, and Noha Elarief. Finding gender-inclusiveness software issues with gendermag: A field investigation. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade, editors, *Proceedings of the 2016 ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 2586–2598, New York, NY, USA, 2016. ACM.

[11] Elizabeth Cha and Maja Matarić. Using nonverbal signals to request help during human-robot collaboration. In *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 2016, pages 5070–5076, Daejeon, Korea, 2016. IEEE.

[12] Amreeta Chatterjee, Mariam Guizani, Catherine Stevens, Jillian Emard, Mary Evelyn May, Margaret Burnett, Iftekhar Ahmed, and Anita Sarma. AID: an automated detector for gender-inclusivity bugs in OSS project pages. In *Proceedings of the 2021 IEEE/ACM International Conference on Software Engineering*, ICSE '21, pages 1423–1435, New York, NY, USA, 2021. IEEE.

[13] Bohkyung Chun and Heather Knight. The robot makers: an ethnography of anthropomorphism at a robotics company. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(3):1–36, 2020.

[14] P John Clarkson and Roger Coleman. History of inclusive design in the UK. *Applied ergonomics*, 46:235–247, 2015.

[15] Roger Coleman. About: Inclusive design. *Design Council*, 1999.

[16] Dermot Crowley. *Smart Teams: How to Work Better Together*. Milton, Queensland: John Wiley & Sons, Thousand Oaks, CA, United State, 2018.

[17] Sally Jo Cunningham, Annika Hinze, and David M. Nichols. Supporting gender-neutral digital library creation: A case study using the gendermag toolkit. In *Digital Libraries: Knowledge, Information, and Data in an Open Access Society*, pages 45–50, Cham, Switzerland, 2016. Springer International Publishing.

[18] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of oz studies—why and how. *Knowledge-based systems*, 6(4):258–266, 1993.

[19] Tawanna R. Dillahunt, Vaishnav Kameswaran, Linfeng Li, and Tanya Rosenblat. Uncovering the values and constraints of real-time ridesharing for low-resource populations. In *Proceedings of the 2017 ACM Conference on Human Factors in Computing Systems*, CHI '17, page 2757–2769, New York, NY, USA, 2017. ACM.

[20] Sheena Erete, Yolanda A Rankin, and Jakita O Thomas. I can't breathe: Reflections from black women in cscw and hci. *Proceedings of the 2020 ACM on Human-Computer Interaction*, 4(CSCW3), 2020.

[21] Sheena Erete, Yolanda A Rankin, and Jakita O Thomas. I can't breathe: Reflections from black women in CSCW and HCI. *Proceedings of the 2021 ACM journal on Human-Computer Interaction*, 4(CSCW3):1–23, 2021.

[22] Abrar Fallatah, Bohkyung Chun, Sogol Balali, and Heather Knight. Semi-ethnographic study on human responses to a help-seeker robot. In *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI '2020)*, pages 640–640, March 2020.

[23] Kerstin Fischer, Bianca Soto, Caroline Pantofaru, and Leila Takayama. Initiating interactions in order to get help: Effects of social framing on people's responses to robots' requests for assistance. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 999–1005, Edinburgh, Scotland, UK, 2014. IEEE, IEEE.

[24] Denae Ford, Mahnaz Behroozi, Alexander Serebrenik, and Chris Parnin. Beyond the code itself: how programmers really look at pull requests. In *Proceedings of the 2019 IEEE/ACM International Conference on Software Engineering: Software Engineering in Society*, ICSE '19, pages 51–60, New York, NY, USA, 2019. IEEE.

[25] Jodi Forlizzi and Carl DiSalvo. Service robots in the domestic environment: A study of the roomba vacuum in the home. In *Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, page 258–265, New York, NY, USA, 2006. ACM.

[26] GenderMag. The gendermag project. Website, 2018. Retrieved July 26, 2022 from https://gendermag.org/.

[27] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. Interactions with a moody robot. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 186–193, 2006.

[28] Christian A Gonzalez, Melissa A Smith, and Robert J Youmans. Are human factors students prepared for careers in user experience research? a survey of predicted and

actual skill utilization. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61 of *HFES '17*, pages 1101–1105, Los Angeles, CA, USA, 2017. SAGE Publications.

[29] Catarina Gralha, Miguel Goulao, and Joao Araujo. Analysing gender differences in building social goal models: a quasi-experiment. In *Proceedings of the 2019 IEEE International Requirements Engineering Conference*, RE' 19, pages 165–176, New York, NY, USA, 2019. IEEE.

[30] Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 1253–1262, New York, NY, USA, 2006. Association for Computing Machinery.

[31] Philip J. Guo. Non-native english speakers learning computer programming: Barriers, desires, and design opportunities. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. ACM.

[32] JoAnn T Hackos and Dana Chisnell. Handbook of usability testing: How to plan, design, and conduct effective tests, 1995.

[33] Martyn Hammersley. What is ethnography? can it survive? should it? *Ethnography and Education*, 13(1):1–17, 2018.

[34] Bruce Hanington and Bella Martin. *Universal methods of design expanded and revised: 125 Ways to research complex problems, develop innovative ideas, and design effective solutions.* Rockport publishers, Beverly, United States, 2019.

[35] Claudia Hilderbrand, Christopher Perdriau, Lara Letaw, Jillian Emard, Zoe Steine-Hanson, Margaret Burnett, and Anita Sarma. Engineering gender-inclusivity into software: Ten teams' tales from the trenches. In *Proceedings of the 2020 IEEE/ACM International Conference on Software Engineering*, ICSE '20, page 433–444, New York, NY, USA, 2020. ACM.

[36] Claudia Hilderbrand, Christopher Perdriau, Lara Letaw, Jillian Emard, Zoe Steine-Hanson, Margaret Burnett, and Anita Sarma. Engineering gender-inclusivity into software: ten teams' tales from the trenches. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 433–444, 2020.

[37] Charles G. Hill, Maren Haag, Alannah Oleson, Christopher J. Mendez, Nicola Marsden, Anita Sarma, and Margaret M. Burnett. Gender-inclusiveness personas vs.

stereotyping: Can we have it both ways? In *Proceedings of the 2017 ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 6658–6671, New York, NY, USA, 2017. ACM.

[38] Kat Holmes. *Mismatch: How inclusion shapes design.* MIT Press, 2020.

[39] Catherine Hu, Christopher Perdriau, Christopher J. Mendez, Caroline Gao, Abrar Fallatah, and Margaret Burnett. Toward a socioeconomic-aware HCI: five facets. *CoRR*, abs/2108.13477, 2021.

[40] Niels Ebbe Jacobsen, Morten Hertzum, and Bonnie E. John. The evaluator effect in usability tests. In *Proceedings of the 1998 ACM Conference on Human Factors in Computing Systems*, page 255–256, New York, NY, USA, 1998. ACM.

[41] Simeon Keates, P John Clarkson, Lee-Anne Harrison, and Peter Robinson. Towards a practical inclusive design approach. In *Proceedings on the 2000 ACM Conference on Universal Usability*, CUU '00, pages 45–52, New York, NY, USA, 2000. ACM.

[42] Matthias Kranz and Albrecht Schmidt. Prototyping smart objects for ubiquitous computing. In *Proceedings of the International Workshop on Smart Object Systems in Conjunction with the 7th International Conference on Ubiquitous Computing*, Tokyo, Japan, 2005. Springer.

[43] Christopher A. Le Dantec and W. Keith Edwards. Designs on dignity: Perceptions of technology among the homeless. In *Proceedings of the 2008 ACM Conference on Human Factors in Computing Systems*, CHI '08, page 627–636, New York, NY, USA, 2008. ACM.

[44] Lara Letaw, Rosalinda Garcia, Heather Garcia, Christopher Perdriau, and Margaret Burnett. Changing the online climate via the online students: Effects of three curricular interventions on online CS students' inclusivity. In *Proceedings of the 2021 ACM Conference on International Computing Education Research*, ICER' 21, pages 42–59, Virtual Event, USA, 2021. ACM.

[45] Clayton Lewis. *Using the" thinking-aloud" method in cognitive interface design.* IBM TJ Watson Research Center Yorktown, Heights, NY, 1982.

[46] Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human–Computer Interaction*, 26(8):741–785, 2010.

[47] Jennifer Mankoff, Anind K Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. Heuristic evaluation of ambient displays. In *Proceedings of the 2003 ACM*

Conference on Human Factors in Computing Systems, CHI '03, pages 169–176, New York, NY, USA, 2003. ACM.

[48] Jennifer McIntosh, Xiaojiao Du, Zexian Wu, Giahuy Truong, Quang Ly, Richard How, Sriram Viswanathan, and Tanjila Kanij. Evaluating age bias in e-commerce. In *Proceedings of the 2021 IEEE/ACM International Workshop on Cooperative and Human Aspects of Software Engineering*, CHASE@ICSE 2021, pages 31–40, New York, NY, USA, 2021. IEEE.

[49] Michael Medlock and Herbst Steve. UI tenets and traps cards, 2017.

[50] Christopher J. Mendez, Lara Letaw, Margaret Burnett, Simone Stumpf, Anita Sarma, and Claudia Hilderbrand. From gendermag to inclusivemag: An inclusive design meta-method. In *Proceedings of the 2019 IEEE Symposium on Visual Languages and Human-Centric Computing*, VL/HCC '19, pages 97–106, New York, NY, USA, 2019. IEEE.

[51] Rolf Molich and Robin Jeffries. Comparative expert reviews. In *Proceedings of the 2003 ACM Conference on Human Factors in Computing Systems Extended Abstracts*, CHI '03, pages 1060–1061, New York, NY, USA, 2003. ACM.

[52] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *Proceedings of the 2011 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '11, page 331–338, New York, NY, USA, 2011. IEEE.

[53] Kathleen Musante and Billie R DeWalt. *Participant observation: A guide for fieldworkers.* Rowman Altamira, 2010.

[54] Bilge Mutlu and Jodi Forlizzi. Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. In *Proceedings of the 2008 ACM/IEEE International Conference on Human Robot Interaction*, HRI '08, pages 287–294, New York, NY, USA, 2008. IEEE.

[55] Alan F. Newell, Peter Gregor, Maggie Morgan, Graham Pullin, and Catriona Macaulay. User-sensitive inclusive design. *Universal Access in the Information Society*, 10(3):235–243, 2011.

[56] Jakob Nielsen. Usability inspection methods. In *Proceedings of the 1994 ACM Conference on Human Factors in Computing Systems*, CHI '94, pages 413–414, New York, NY, USA, 1994. ACM.

[57] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the 1990 ACM Conference on Human Factors in Computing Systems*, CHI '90, pages 249–256, New York, NY, USA, 1990. ACM.

[58] Susmita Hema Padala, Christopher John Mendez, Luiz Felipe Dias, Igor Steinmacher, Zoe Steine Hanson, Claudia Hilderbrand, Amber Horvath, Charles Hill, Logan Dale Simpson, Margaret Burnett, Marco Gerosa, and Anita Sarma. How gender-biased tools shape newcomer experiences in oss projects. *IEEE Transactions on Software Engineering*, (01):1–1, apr 5555.

[59] Sora Park and Justine Humphry. Exclusion by design: intersections of social, digital and data exclusion. *Information, Communication & Society*, 22(7):934–953, 2019.

[60] Andreas F Phelps and Michael J Horman. Ethnographic theory-building research in construction. *Journal of construction engineering and management*, 136(1):58–65, 2010.

[61] Peter G Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies*, 36(5):741–773, 1992.

[62] Gede Artha Azriadi Prana, Denae Ford, Ayushi Rastogi, David Lo, Rahul Purandare, and Nachiappan Nagappan. Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in OSS. *CoRR*, abs/2010.00822:1–1, 2020.

[63] Yolanda A Rankin and India Irish. A seat at the table: Black feminist thought as a critical framework for inclusive game design. *Proceedings of the 2020 ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[64] Yolanda A. Rankin, Jakita O. Thomas, and Sheena Erete. Real talk: Saturated sites of violence in cs education. *ACM Inroads*, 12(2):30–37, may 2021.

[65] Raquel Ros, Marco Nalin, Rachel Wood, Paul Baxter, Rosemarijn Looije, Yannis Demiris, Tony Belpaeme, Alessio Giusti, and Clara Pozzi. Child-robot interaction in the wild: advice to the aspiring experimenter. In *Proceedings of the 13th international conference on multimodal interfaces (ICMI*, pages 335–342, Alicante, Spain, 2011. ACM.

[66] Eric Rose. Approaching humans for help: A study of human-robot proxemics. Union College Honors Thesis, 2016.

[67] Stephanie Rosenthal and Manuela Veloso. Mobile robot planning to seek help with spatially-situated tasks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI '2012)*, volume 26, pages 2067 – 2073, Toronto, Canada, 2012. AAAI.

[68] Dirk Rothenbücher, Jamy Li, David Sirkin, Brian Mok, and Wendy Ju. Ghost driver: A platform for investigating interactions between pedestrians and driverless vehicles. In *Adjunct Proceedings of the 2015 ACM International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '15, page 44–49, New York, NY, USA, 2015. ACM.

[69] Karina A. Roundtree, Steven Hattrup, Janani Swaminathan, Nicholas Zerbel, Jeffrey Klow, Vivswan Shitole, Abrar Fallatah, Roli Khanna, and Julie A. Adams. Inclusive design guidance: External autonomous vehicle interfaces. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 1054–1058, Los Angeles, CA, USA, 2020. SAGE.

[70] Per Runeson, Martin Höst, Rainer Austen, and Björn Regnell. *Case study research in software engineering–guidelines and examples*. John Wiley & Sons Inc., Hoboken, NJ , USA, 2012.

[71] Selma Sabanovic, Marek P Michalowski, and Reid Simmons. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 596–601, Istanbul, Turkey, 2006. IEEE.

[72] Elaheh Sanoubari and James E Young. Hi human can we talk? an in-the-wild study template for robots approaching unsuspecting participants. In *Workshop on the Social Robots in the Wild at the ACM/IEEE 2018 International Conference on Human-Robot Interaction (HRI)*, Chicago, United State, 2018. ACM/IEEE.

[73] Allison Sauppé and Bilge Mutlu. The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3613–3622, Seoul, Korea, 2015. ACM.

[74] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. Intersectional HCI: engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 5412–5427, Denver, USA, 2017. ACM.

[75] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–36, 2020.

[76] Arun Shekhar and Nicola Marsden. Cognitive walkthrough of a learning management system with gendered personas. In *Proceedings of the 2015 Conference on Gender & IT*, GenderIT '15, pages 191–198, New York, NY, USA, 2018. ACM.

[77] Dilruba Showkat and Cindy Grimm. Identifying gender differences in information processing style, self-efficacy, and tinkering for robot tele-operation. In *Processdings of the 2018 IEEE International Conference on Ubiquitous Robots*, UR '18, pages 443–448, New York, NY, USA, 2018. IEEE.

[78] David Sirkin, Brian Mok, Stephen Yang, and Wendy Ju. Mechanical ottoman: how robotic furniture offers and withdraws support. In *Proceedings of the 2015 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 11–18, New York, NY, USA, 2015. ACM.

[79] Brett Stoddard. Designing and evaluating a user interface for multi-robot furniture. 2022.

[80] Maria Stone, Frank Bentley, Brooke White, and Mike Shebanek. Embedding user understanding in the corporate culture: Ux research and accessibility at yahoo. In *Extended Abstracts of the 2016 SIGCHI Conference on Human Factors in Computing Systems*, CHI EA '16, pages 823–832, New York, NY, USA, 2016. ACM.

[81] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction,*, 24(3), 2017.

[82] Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. Avoiding the south side and the suburbs: The geography of mobile crowdsourcing markets. In *Proceedings of the 2015 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, page 265–275, New York, NY, USA, 2015. ACM.

[83] Gustavo F. Tondello, Dennis L. Kappen, Elisa D. Mekler, Marim Ganaba, and Lennart E. Nacke. Heuristic evaluation for gameful design. In *Proceedings of the 2016 ACM Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY Companion '16, pages 315–323, New York, NY, USA, 2016. ACM.

[84] Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 ACM Conference on Human Factors in Computing Systems*, CHI' 19, New York, NY, USA, 2019. ACM.

[85] Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühnlenz, Dirk Wollherr, and Martin Buss. Robots asking for directions: The willingness of passers-by to support robots. In *Proceedings of the 2010 ACM/IEEE international conference on Human-Robot Interaction (HRI)*, volume 5, pages 23–30, Osaka, Japan, 2010. ACM/IEEE.

[86] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. The cognitive walkthrough method: A practitioner's guide. In *Usability Inspection Methods*, page 105–140. John Wiley & Sons, Inc., Hoboken, NJ , USA, 1994.

[87] Stephen Yang, Brian Ka-Jun Mok, David Sirkin, Hillary Page Ive, Rohan Maheshwari, Kerstin Fischer, and Wendy Ju. Experiences developing socially acceptable interactions for a robotic trash barrel. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 277–284, Kobe, Japan, 2015. IEEE.