Systematizing Inclusive Design in MOSIP: An Experience Report

Soumiki Chattopadhyay[1], Amreeta Chatterjee[1], Puja Agarwal[1], Bianca Trinkenreich[2], Swarathmika Kumar[3], Rohit Ranjan Rai[3], Resham Chugani[3], Pragya Kumari[3], Margaret Burnett[1], Anita Sarma[1]

[1]Oregon State University, United States, {chattoso, chattera, puja.agarwal, burnett, anita.sarma}@oregonstate.edu
[2]Colorado State University, United States, bianca.trinkenreich@colostate.edu
[3]MOSIP-IIIT Bangalore, India, {swarathmika, rohit, resham}@mosip.io, pragya.kumari@technoforte.co.in

***Abstract*—The GenderMag method has been successfully used by software teams to improve inclusivity in their software products across various domains. Given the success of this method, here we investigate how GenderMag can be systematically adopted in an organization. It is a conceptual replication of our prior work that identified a set of practices and pitfalls synthesized across different USA-based teams. Through Action Research, we trace the 3+ years long journey of GenderMag adoption in the MOSIP organization; starting from the initial 'unfreeze' stage to the institutionalization ('re-freeze') of GenderMag in the organization's processes. Our findings identify: (1) which practices from the prior work could be generalized and how some of them had to be modified to fit MOSIP organization's context (Digital Public Goods, open-source product, and fully remote work environment), and (2) the pitfalls that occurred.**

***Index Terms*—GenderMag, gender inclusivity, conceptual replication, inclusive design practices**

I. Introduction

Inclusive design has become a clarion call for action such that software tools provide fair and equitable experiences across all user demographics [1, 2, 3, 4]. Gender-inclusivity is one such type of inclusivity that has gained attention amongst industry practitioners [5, 6, 7]. One approach to gender-inclusive software design is the GenderMag method. This method is rooted in research that has found individual differences in preferred problem-solving styles to cluster by gender [8]. Developers risk embedding unintentional biases–inclusivity bugs—into their products if their tools do not accommodate the different problem-solving styles, since individuals whose styles are unsupported are disproportionately disadvantaged by the software.

The GenderMag method has been widely used across different domains such as CS courses [9, 10, 11, 12, 13, 14, 15], digital robotics [16], open-source projects [17, 18, 19], infrastructures [20], software engineering job advertisements [21], and so on. Field studies have shown the improvement in software inclusivity when teams fix the inclusivity bugs [22, 23]. Murphy-Hill et al.'s [5] found their Google code review tool to be more inclusive after they fixed the inclusivity bugs they found from GenderMag evaluations. Similarly, Guizani et al. [24] used the GenderMag 'Why/Where/Fix' debugging paradigm to find and fix inclusivity bugs in an open-source project, which led to improved inclusivity and better task completion.

Given the efficacy of GenderMag in designing gender-inclusive software, the question arises: *How can its adoption be systematized in software development teams*? Burnett et al. [22] in 2017 provides an early investigation of this question in their year-long study of five software development teams at Microsoft, USA, which provided insights into how GenderMag was introduced into Microsoft and early impacts of incorporating it on team dynamics. In 2020, Hilderbrand et al. [25] investigated the experiences of 10 different teams from different organizations to synthesize a set of 9 practices and 2 pitfalls of applying GenderMag to their products.

In this paper, we track the adoption of GenderMag through an Action Research study at MOSIP, a non-profit organization based in India. We frame our study as a conceptual replication of [25] to investigate the practices that help in the systematic adoption of GenderMag, and identify the pitfalls as well.

A conceptual replication allows us to assess which findings from [25] remain valid in a very different organization, product context, and as the teams get maturity in using GenderMag. (The teams in [25] were largely new to the method; 9 out of 10 had 1 year or less experience). Conceptual replications are important to perform as they ensure the insights from the study being replicated can be generalized or adapted to new contexts, while maintaining their utility.

A. Conceptual Replication Setting: MOSIP

The Modular Open Source Identity Platform (MOSIP) [26] was established in a university (IIIT Bangalore) in 2018, to allow affordable and accessible implementation of national ID systems. MOSIP helps governments implement effective Digital Public Infrastructure (DPI) in countries in Africa, South and Southeast Asia, and South America. They provide the infrastructure for generating IDs that are used by governments to collect user data including biometrics, as well as applications for end users for their ID management, connecting with government, and financial services.

Our study has two similarities to Hilderbrand et al.'s study. Both studies use Action Research as their field study methodology, and span multiple products and multiple teams. However, our study context is different in the following ways. (1) Unlike the Hilderbrand et al. study, the products in our study all relate to digital public goods, with a philosophy of

transparency, accessibility, and affordability. (2) The Hilderbrand et al. study had a mix of universities and private companies operating solely within their own institutions (not Open Source). In contrast, the organization in our study is a non-profit with fully open-sourced project code. This allows client countries to customize the applications according to their requirements. These differences could impact mindsets and interests in adopting GenderMag. (3) MOSIP applications are developed in the Global South and primarily for adoption by countries in the Global South, which impacts not only the software features, but also the work culture and organizational expectations. And (4), the MOSIP organization and employees are based in India, and operate in a completely distributed manner. Employees reside in different regions within the country, work independently, and meet virtually. The distributed nature of work setting also distinguishes us from the past GenderMag works.

These contextual differences enable investigating to what extent the practices from Hilderbrand et al. [25] generalize to this new context and which needed adaptation, and the pitfalls that were encountered.

II. Background

GenderMag (Gender-Inclusiveness Magnifier) is an inspection method to help software professionals evaluate the applications they are building from a gender-inclusiveness perspective and find gender-inclusivity bugs. It uses five core facet values to understand diverse problem-solving approaches among users, refer to Fig. 1. These facets are embodied in three distinct personas: "Abi", "Pat", "Tim", each representing unique combinations of these values. Abi is on one end of problem-solving spectra characterized by certain facet values as shown in Fig. 1, and Tim is positioned on the opposite end of these spectra. Pat occupies a different mix of facet values. Evaluating through the lens of these personas allows software professionals to not only gain inclusivity across problem-solving approaches, but also across gender. This is because the GenderMag problem-solving styles statistically cluster around genders; thus fixing a system's problem-solving biases found using GenderMag also fixes gender biases [22, 8, 24, 6].

Fig. 1. All three GenderMag Personas and their five facets.

Evaluators use these personas to do specialized specialized Cognitive Walkthrough (CW) [27][28] in order to evaluate the applications from a gender-inclusiveness perspective. In a GenderMag session, one participant has the Facilitator role, ensuring that the group stays on track, actively engages, and follows the rules, e.g., staying true to the persona. Another participant is the Driver, navigating through the application during the sessions, ensuring that the process is aligned with planned steps and actions. This is generally someone who has decision-making powers over the application. The participant acting as Recorder documents all the opinions of the participants on a specialized form called the GenderMag form [29].

A GenderMag form contains the scenario (overall goal), its sub-goals, and the actions. The form asks the following questions to the participants:

Subgoal: *Will <persona> have formed this subgoal as a step to their overall goal? (Yes/no/maybe, why)*

Action1: *Will <persona> know what to do at this step? (Yes/no/maybe, why)*

Action2: *If <persona> does the right thing, will s/he know s/he did the right thing and is making progress toward their goal? (Yes/no/maybe, why)*

Apart from the three main roles, a GenderMag session also has multiple evaluators, who each responds to these questions with "Yes/no/maybe" along with their reasoning, without any requirement for consensus among them. They also have the option to attach their answers to specific facet(s) that drove their reasoning. When an answer is associated with one or more facets, it is identified as an inclusivity bug.

In some instances, any UI element(s) which particularly helped / hindered the persona's course of action are documented in a designated section of the form titled "What in the UI helped/confused the persona?". The process concludes with a "Debrief" session, where the number of inclusivity and usability bugs are counted by tallying the total number of responses with or without facet(s) attached.

III. Methodology

Action Research is an iterative field research method that involves developing scholarly knowledge through engaging with communities seeking a change. It blurs the line between researchers and participants, hence allowing roles to interchange. This is because, in Action Research, research is done "with" the participants instead of "to" them. Additionally, unlike controlled studies, Action Research emphasizes on real-world applicability and aims for long-lasting impact [30].

Despite the fact that Action Research allows the intervention (GenderMag practices) to evolve to match the context, it still emphasizes rigor like other empirical methodologies. In Action Research, rigor focuses on credibility and validity, which are attained primarily through member checking (verifying interpretations of events by the participants themselves) and through triangulation (investigating whether multiple sources of evidence produce the same conclusion).

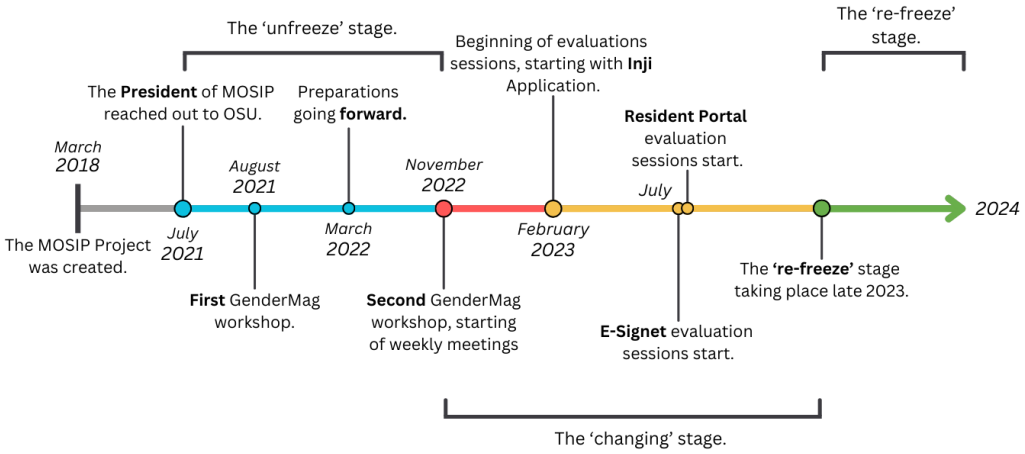

Fig. 2. Timeline of our study with the three stages of Action Research: 'unfreeze', 'changing', and 're-freeze'.

A. Participants and Applications

A key aspect of Action Research is that it is collaborative: those responsible for the actions are involved in deciding how to improve upon them. In our investigation, MOSIP employees were those responsible for the changes to their applications, and thus were participant-researchers. In total, 40 participant-researchers from different levels of the organization were part of the study, as shown in Fig. 3.

The MOSIP organization and its leadership is supported by three arms: the Office of President, Technology Development, and Dissemination. The Office of President supports the leadership through strategy, finance, legal, and communication functions. Technology Development is responsible for platform development, maintenance and innovation. Dissemination is mainly responsible for disseminating and supporting implementation of MOSIP's technologies in countries like Morocco, Ethiopia, Togo, Uganda, and many more.

Our GenderMag evaluations were conducted within the Technology Development arm, where stakeholders encouraged integrating GenderMag evaluations into the development process for all of MOSIP's products.

During the course of this study, we applied GenderMag to three applications, the fourth one (OpenG2P) is underway now. MOSIP has a wide variety of product applications: some are reference architecture, some are for operators making the digital IDs, and the rest are for end users. The MOSIP leadership made a conscious choice in selecting the following three applications because they all target end-user, and gender inclusivity would have the most widespread impact:

(1) Inji is a mobile app for digital ID and other vital documents issuance and verification for residents. It was the first choice because it has "*the most kind of varied audience interacting with this [ID software].*" [P8]

(2) E-Signet is a self-service portal that enhances digital identification in government services by offering secure, streamlined online authentication. It also supports governments to upgrade existing ID systems with digital authentication capabilities cost-effectively.

(3) Resident Portal is a web-based interface that allows residents, typically older individuals, to manage their Unique Identification Number (UIN)-related services, like updating personal details such as phone numbers and addresses.

B. Study Timeline

The Action Research process has three stages. Fig.2 refers to these stages and the timeline for our study, which spanned 38 months at the time of writing this paper. In the first stage, *Unfreeze*, the organization recognizes a problem and gets an impetus for change. In our study, this happened in 2020 which led MOSIP leadership to contact OSU to learn more about GenderMag in July 2021.

In the next stage, *Changing*, participants experiment with new processes to achieve desired results. This stage started with the evaluation of the Inji application on February 1st, 2023. Over seven months, we conducted 13 GenderMag sessions. Two employees at level L1 from the 'Office of President', were responsible for directing participation in the project and worked closely with the entire team. These evaluations included 9 other people from the 'Technology Department'. In the beginning, the university researchers led the evaluations and trained two advocates (product owners at level L2), who then took over the evaluations. Once Inji evaluations were over, the two product owners started the next set of evaluations in parallel. E-Signet evaluations began on July 20th, 2023 and included 11 sessions over 5 months. Resident Portal started on July 28th, 2023 and included 9 sessions spanning 3 months.

In the final, *Refreeze* stage, changes from the prior stage are integrated into the organization's standard processes. Here, MOSIP began standardizing the practices, fixing the inclusivity bugs, having conversations about inclusivity beyond the products being evaluated, and bringing the inclusivity lessons to their broader community.

C. Data Collection and Analysis

Our data sources included: (1) GenderMag evaluation session recordings (19 out of 33 sessions, the others were not

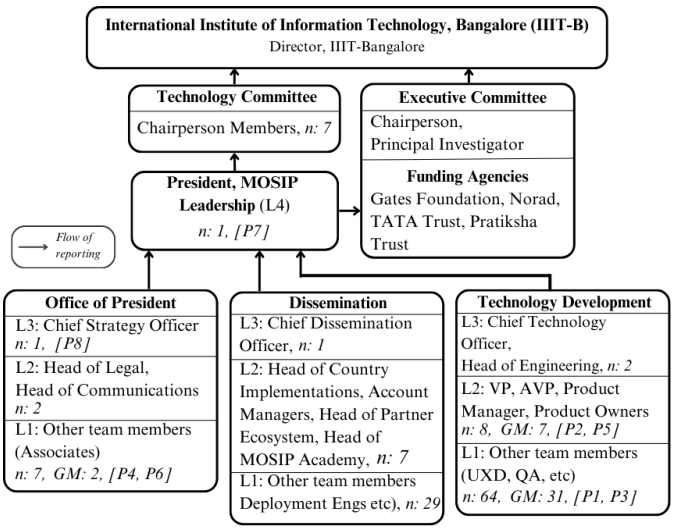

Fig. 3. MOSIP Organization Structure. Each level includes the number of employees in that level (n), the number of GenderMag participants (GM), and interview participants pseudonyms (Px).

saved because of technical errors), (2) the filled-out GenderMag forms, and (3) the inclusivity bug lists and the fixes. We also conducted 8 follow-up semi-structured interviews with employees in different roles, including leadership (discussed later). While the practices and pitfalls were derived largely from the evaluation sessions and interviews, the other pieces of data provided additional context. See supplementary [31] for an example GenderMag form, an example of an inclusivity bug fix, and the interview protocol.

GenderMag evaluation recordings: To identify the practices and pitfalls at MOSIP, we transcribed the session recordings. Two authors analyzed and coded around 20% of the transcripts using the practices and pitfalls from [25] as a code book, until they reached an Inter-rater Reliability [32] of 0.90 (Cohen's Kappa). This required two rounds. The first round included 10% of the data and resulted in an IRR of 0.39. After which, some of the practices in [25] were subdivided. For example, Calculating bias included 3 subcomponents. Following this subdivision, another 10% of the data was coded, and an IRR of 0.90 was achieved. One researcher then proceeded to code the remaining transcripts for Inji.

We validated the practices identified from Inji by *triangulating them with the other two applications*: E-Signet and Resident Portal (see supplemental [31]). (Quotes of MOSIP team members from these transcripts are referred as [MX-SX (Application)], where 'MX' is the speaker number, 'SX' the session number for a specific application (Inji/E-Signet /Resident Portal).

Interviews: We investigated the practices that were integrated as standard organizational procedures, and the experiences of the participants' using GenderMag through an analysis of the follow-up semi-structured interview. One author first transcribed the follow-up interviews and unitized the data points. The data points were divided into three types based on the interview questions: practices and pitfalls referenced from [25], and "organizational process". Two authors then independently analyzed the data points, applying deductive coding to the referenced practices and pitfalls, and inductive coding to identify the organizational processes. They then calculated the Inter-rater reliability on 10% data points from all three types, which gave a Cohen's Kappa value of 0.80. After discussing the discrepancies, the authors went for a second round of coding (10%), with a Cohen's Kappa value of 0.96.

TABLE I
INTERVIEW PARTICIPANTS' DEMOGRAPHICS.

ID	Role	Gender	MOSIP Tenure (yrs.)	Org. lvl.
P1	UX Designer	M	1.5	L1
P2	Product Manager	F	4.5	L2
P3	Technical Lead	M	1	L1
P4	Associate - Legal and Policy	F	3.5	L1
P5	Product Owner	F	3	L2
P6	Senior Associate - Policy and Outreach	M	3	L1
P7	President	M	6	L4
P8	Chief Strategy Officer	M	6	L3

IV. THE "UNFREEZE" STAGE

The Unfreeze stage in Action Research is about recognizing a problem and generating an impetus for change. The organization's "unfreeze" impetus came from the leadership, the President of the Executive Board (P7). Inclusivity is one of the core principles of MOSIP, as P7 stated: "*any identity system should be able to include everybody who has the right to be included.*" This goal was further driven home by the funding agencies. Specifically, "*the Gates Foundation was very actively asking us to be very vigilant on whether our systems are gender inclusive, it's a massive part of their agenda.*" [P8]. This urged P7 to review research on gender and software design, and found out about GenderMag. He reached out to us in July 2021 (Fig.2), which marked the beginning of our collaboration.

We conducted two workshops where OSU researchers provided GenderMag training. The first workshop (conducted on August 2021) introduced the method (1.5 hours), followed by a 2-hour hands-on training. Coming in to this workshop, the Engineering team was skeptical of gender biases in software; they were under the assumption that software is gender agnostic. Post workshop, there were several conversations among the engineering team members on the types and impact of the gender biases uncovered in the training session. The engineering leadership recognized the importance of "*training up our people to build experience on GenderMag framework would be very useful. I think that's where it took us a bit of time to figure out how to do this.*" [P8]

Once there was consensus that debiasing the software design was a valid use of the limited time of engineers, the head of engineering set up a meeting with OSU (Mar 2022) to discuss how and where gender biases may bleed into the different

parts of the MOSIP architecture, and which parts should be prioritized. After this meeting, once the head of engineering was convinced about the value of GenderMag, things started to progress forward. On identifying the Inji application as the first choice to be GenderMag'ed by MOSIP, we conducted another workshop (Nov 2022) with Inji developers and other relevant stakeholders of the application. This workshop followed the same format as the first workshop.

Note that the key participants in these conversations were individuals from the leadership team, including engineering leadership, as well as members from the Dissemination team. These discussions brought to focus the mandate set by the Gates Foundation on inclusivity of the software design, as P8 noted, "*there was an element of policy decision to this. But I think our engineering [team] after the workshop very quickly realized that this would be useful.*"

Additionally, to show (and get from us) commitment to GenderMag, MOSIP set up a student fellowship (which is continuing at the time of writing this paper). Once the fellowship was set up, the OSU student embedded with the team having full access to the team's communication channels, which they used to schedule GenderMag evaluations.

V. The "Changing" Stage

The "Changing" stage in Action Research is when the organization experiments with various methods and practices. A key aspect of Action-Research method is that it follows a highly iterative planning-acting-observing-reflecting spiral, the outcomes of which then feed into the next planning-acting-observing-reflecting spiral, and so on. Since the team had "bought in" to the concept of making their applications gender-inclusive, they engaged in iteratively refining and tailoring the GenderMag-based practices to suit the team's unique requirements and context.

Table II shows the practices that MOSIP followed. Some practices from [25] were used as is (marked in green), others were modified to fit the team context (highlighted in yellow). Only one practice (Facet survey) was not used at all. In the rest of the section, we focus our discussion on the practices that were *adapted* to the MOSIP context.

A. Large Group vs. Small Group

Previous study [25] reported that large groups are good for learning the GenderMag method, while small groups are best for hands-on evaluations. However, in our study, the teams preferred sessions with large groups for three reasons.

The first reason was largely logistical. They felt that recording all the information while simultaneously contributing their opinions during the sessions was cognitively taxing. Therefore, to reduce the cognitive load they suggested including more members in the evaluation sessions: "*...having a few more people generally would be helpful . . . Because we are also thinking and or writing . . . we would not be able to contribute so much.*" [M1-S2 (Inji)].

However, this decision has to be made carefully, as having too many members evaluating can lead to "*...too many conflicts of opinion*" [M1-S2 (Inji)]. Additionally, scheduling a common time across a large group of people was challenging. It took a lot of time and communication to set up these meetings, some of which had to be canceled when key members became unavailable. From the second year onwards, the teams set up regular product meetings for GenderMag evaluations that team members had to attend.

The other two reasons were due to the team dynamics - the second reason being individual personalities, and the third being work setting context. We found experienced employees (at the L1 levels) actively participated in the evaluation sessions, but newer hires showed a lack of interest and communication, as observed by one participant who was present across all three teams: "*the core team that works for E-Signet are senior developers, and they are older in age. But Resident Portal [members] are all kids, fresh out of college. They were like, 'I don't care what this portal looks like' they're not interested in making the UI good.*" [P5]

This disinterest combined with the distributed (online) settings resulted in minimal contributions in the evaluation sessions as P5 pointed out, "*I was handling Resident Portal. . . , I think 2 to 3 sessions. I literally got no bugs. . . it's just everything is perfect. This portal is perfect because of the (developers') biasness that came into picture.*" Individual characteristics and a lack of team bonding among the junior hires also played a role; "*All [Inji team members] are extroverts. They just want to contribute somehow as opposed to the Resident Portal team who are on mute 24/7. They are introverts in general. They wouldn't want to speak up.*" [P5]

The team lead explored different options to overcome this problem. They invited members from different applications, as stated by P1, "*there were a couple of different people on that call . . . there was a switching.*" Participant P5 also tried polling members to identify those who are interested in the sessions; "*If we can ask people, do you want to be a part of such evaluation? Probably that will help.*" The participant also explored hand-picking members based on personality, "*who are like extroverts and would probably talk a lot,*" as they are more likely to add to the evaluation. Eventually, the teams decided to continue with large groups (in future applications) as this proved to be beneficial as, "*. . . having a different set of people always help in bringing diverse opinion. . .* " [P3], making it the third reason to prefer large groups.

B. Multi-path Evaluation

Simultaneously evaluating two small paths (sequence of action items) that reach the same end goal facilitates the direct comparison between paths and helps reduce the number of sessions needed [25]. However, in our study, the teams found themselves deeply evaluating the alternate paths and losing the original flow. In earlier evaluations, we (OSU researcher) had intervened to bring back the focus: "*. . . we go back to that later, once we finish this sub-goal and the respective actions.*"

To overcome this problem, the teams started noting down the alternate subgoals or actions that they felt they need to evaluate in subsequent sessions, to not derail the current evaluation path. The team also introduced 'happy flow and

TABLE II
PRACTICES AND PITFALLS RELEVANT TO THE CHANGING STAGE.

Type	Practices / Pitfalls	Definition	Adopted / Occurred	How it was modified	Application
Practice	Learning vs. doing GenderMag	Large groups for learning, small groups for doing.	Modified	Large groups for both learning and doing.	Inji, E-Signet, Resident Portal
Practice	Multi-path evaluation	Simultaneous evaluations of two small paths.	Modified	Evaluation of two paths done serially.	Inji, E-Signet
Practice	Abstracting beyond	Abstracting findings of previous sessions to similar UI patterns.	Adopted as is	X	Inji, Resident Portal
Practice	Abi first	Choosing Abi as the first persona.	Adopted as is	X	Inji, E-Signet, Resident Portal
Practice	Speaking through Abi	Using Abi to criticize poor designs.	Adopted as is	X	Inji
Practice	Calculating bias: Understanding assumptions and the importance of fixing	Teams realize their biased assumptions about their population, and the gravity of fixing the inclusivity bugs, during 'debriefing'.	Modified	Teams realized their assumptions and the importance while doing the evaluations, not just after calculating bias.	Inji, E-Signet
	Calculating bias:Thoughtful discussions about facets	Calculating bias also stirs discussions about facets.	Adopted as is	X	Inji, E-Signet
Practice	Facet survey	Understand, analyze, and measure data, using facets.	Not used	X	X
Pitfall	Beyond our control (Pitfall 1)	Teams lacking decision-making power.	Did not occur	X	X
Pitfall	Evaluating a proxy (Pitfall 2)	Evaluating a "similar" system led to inaccurately assessing the real one.	Occurred as is	X	Inji

negative flow'. Happy flow was the ideal actions that would allow the user to accomplish the task. Interestingly, the teams also evaluated the 'negative flows', evaluating how a user stuck in their task would be able to get on the right path again. "*... In my experience, so far with GenderMag, I have not come across such a scenario where we are evaluating, there was this need to evaluate the negative flow as well.*" [OSU researcher].

In most cases, the happy and negative flows were evaluated sequentially, completing "happy flow" evaluations for all use cases and for both the personas before going to the negative flows. In very few cases, they evaluated both flows simultaneously, but these led to higher cognitive loads. When debriefing and discussing a lack of participation in the session, a potential problem was the additional cognitive load in keeping multiple pathways in the mind: "*we need to like literally explain the actual scenario, and then explain where is the negative scenario. We are switching contexts a lot...*" [M2-S6 (Inji)]

C. Calculating Bias

Hilderbrand et al. [25] reported that calculating the bias (number of inclusivity bugs found) when debriefing at the end of an evaluation, led teams to reflect their assumptions about the users and the importance of fixing the bugs, and spark thoughtful discussions about the facets. We found that the teams engaged in these practices during the evaluation session(s) and not just at the end, except for overarching discussions about the facets which happened during debriefing.

For example, during an evaluation session, the team realized their underlying assumptions: "*... when we develop a particular software... we are very biased because we think everybody is like us in terms of skills... we think everyone is tech-savvy...*" [M5-S1 (E-Signet)]. This became evident to them when they found very few diverse opinions when using the Tim persona: "*... the response was quite less today. Maybe it's because we don't have different opinions, like we used to have for Abi.*" [M1-S10 (Inji)]

TABLE III
PRACTICES STANDARDIZED IN THE 'RE-FREEZE' STAGE.

Practices	Definition
Customizing forms to add 'How to fix'	The inclusion of "how to fix" in GenderMag forms would help with design fixes.
Incorporate tracking of inclusivity bugs	Streamlining the process of addressing inclusivity bugs as part of regular bug triaging.
Training advocates	Training advocates in GenderMag to lead and train other members.
GenderMag moments	Applying GenderMag components just-in-time to quickly evaluate designs, instead of doing a full GenderMag session.
GenderMag'ing Beyond Product	Applying or thinking from the perspective of GenderMag outside of work.
GenderMag'ing Beyond Boundaries	Commitment to inclusivity thinking by raising awareness in the broader community

D. Practices Adopted As Is

Among the 9 practices in [25], 3 of them along with one subcomponent, were adopted as is (green rows Table II) in the 'changing' stage. As espoused in the practice 'Abstracting Beyond', MOSIP abstracted findings from previous session to similar UI patterns and did not re-evaluate similar UIs and use cases. For instance, "*I think [UI] we can skip because it's similar to UIN's [screen].*" [M1-S11 (Inji)]. Next, as in

any GenderMag evaluation, the MOSIP team also chose Abi as their first persona for almost every session, which refers to the practice 'Abi first'.

'Speaking through Abi' practice refers to evaluators using Abi as an armor to point out or criticize a teammate's design, was also adopted as is. However, over time, teammates got less defensive about their designs, and would openly discuss the issues as they understood why Abi-like users may face issues. Thus, this practice only occurred for the Inji evaluations (refer Table. II). With increasing GenderMag experience, people's mindset evolved and they got accustomed to thinking like Abi. Finally, recall, the subcomponent practice of thoughtful facet discussions when calculating bias also occurred here.

E. Pitfalls

Pitfall #1 from [25], lack of control over the applications, did not occur as the product owners (L2) were actively involved and supported by leadership (L4). Unfortunately, we found a similar pitfall to Pitfall #2 in [25]. In six of Inji's evaluation sessions, MOSIP team brought in UI versions which lacked and/or had extra features compared to the actual application. This discrepancy prevented them from evaluating the original system as pointed out by [M4-S2 (Inji)], "*... in the actual app, we will not have that [UI feature]... in case we are capturing feedback from that perspective in the actual app, we do not have that right now just for demonstration purpose.*" It also led them to skip a few use cases due to the lack of the appropriate screens, "*... that screen should've been here, it's not here in this prototype.*" [M3-S6 (Inji)]

VI. The "Refreezing" Stage

The "Refreezing" stage in Action Research occurs when methods and practices become a part of an organization's standard processes. We report on the 5 inclusivity practices that became integral to the organization's routines; 2 of which (marked in green in Table III) are common from [25].

A. Customizing Forms to Add 'How To Fix'

The MOSIP team, when answering 'why' Abi may not proceed further, also discussed potential fixes, which they sometimes noted in the GenderMag form. For instance, "*... in the UI elements right, like if we can note down about the warning message... so that later, when we come back, we can take that input for the revised UX...*" [M2-S6 (Inji)]. However, in some cases, we had to guide the discussion back to the evaluation so that the team didn't get too involved brainstorming design fixes and lose sight of the session goal.

We observed that the team started internalizing the facets, the Abi persona, and the potential fixes in the (hi-fidelity) prototypes they brought to the GenderMag evaluations. Table IV shows the % of inclusivity bugs going down over time.

When fixing the inclusivity bugs, the team referred to the completed GenderMag forms. While the GenderMag forms included details about 'why' [persona] would face a problem, it was insufficient after a large time gap between the evaluation and fixing. This was a particular problem with the Inji application. The Inji evaluations were conducted on hi-fidelity (Figma) prototypes from February to July. However, due to other business requirements, the application went through significant changes, and the development of the application based on the outcomes from the GenderMag evaluations did not begin until late 2023. By then, a significant time had elapsed since the initial GenderMag evaluations, which . This long time gap made it difficult for developers to "*even go back and recall what people had told us six months back.*" [P2]

TABLE IV
Decrease in the % of inclusivity bugs found in the debriefing stage after each GenderMag session of Inji.

Sessions	Persona	Inclusivity bugs %
S1	Abi	75%
S2	Abi	84%
S3	Abi	73%
S4	Abi	73%
S5	Abi	73%
S6	Abi	58%
S7	Abi	58%
S8	Abi	54%
S10	Tim	29%
S11	Tim	29%
S12	Abi	37%
S13	Abi	37%

A significant gap between the evaluation session and fixing the design is a pitfall (Pitfall #3) to avoid because (1) it makes it hard to recall the evaluation discussions, and (2) there is a chance that the application evolves, making the earlier evaluations ineffective: "*... by the time we collated everything, the app itself had gone through multiple iterations, and the feedback that we received was sort of irrelevant to the application itself.*" [P2] To mitigate this issue, the MOSIP team suggested that learning "*earlier in the process*" that the issues can be considered "*on piece meal basis*" and iteratively evaluating and fixing the application from the beginning would be helpful. Building on this, participant P5 proposed adding a 'how' section to the GenderMag forms (which will be used in the upcoming GenderMag evaluations of OpenG2P application). This would allow the teams to document the design fixes for the identified inclusivity bugs and address them strategically, "*... probably like we already talk about what is the problem if we can also talk about how it can be solved. Just one-liner will also suffice.*" [P5].

B. Systematizing Tracking of Inclusivity Bugs

To systematize the fixing of inclusivity bugs, MOSIP treated these bugs as a first order entity in their existing bug-tracking process. These bugs were then triaged, and were assigned priorities and developers in Jira [33] and tracked accordingly.

MOSIP used a script to automatically transfer the data from the GenderMag forms (which were treated as inclusivity bug reports) into a spreadsheet. The data from these evaluation reports was used for bug triaging. The MOSIP teams triaged the inclusivity bugs as follows. First, they identify blockers based on when the responses of the GenderMag questions is a "No" and how many evaluators found the bug to be a blocker.

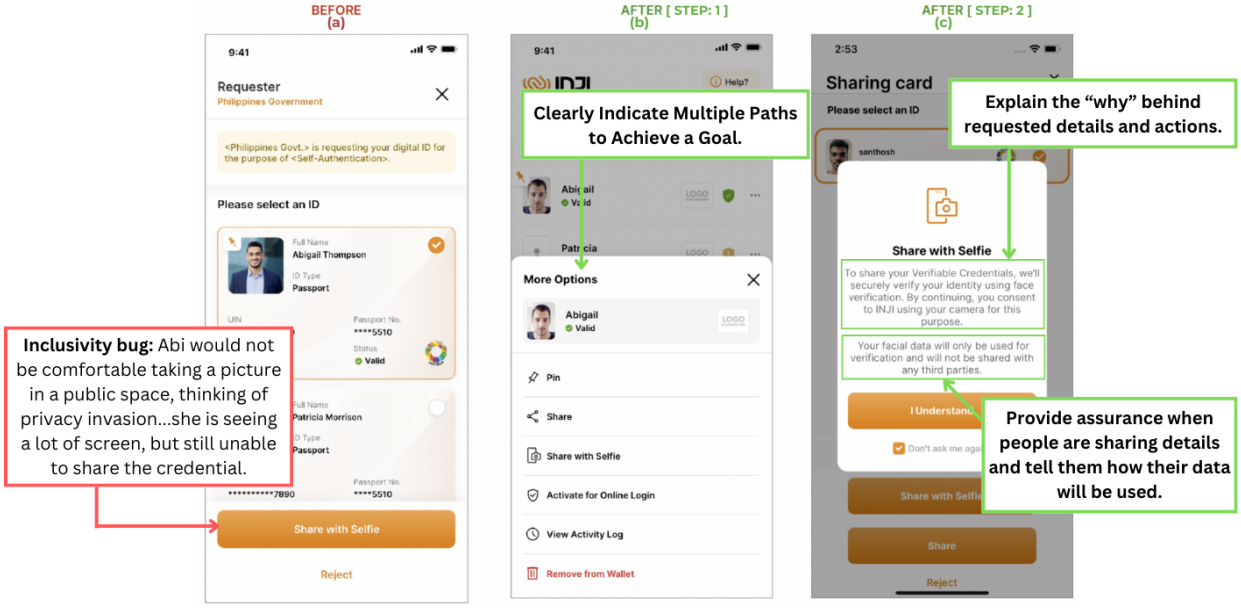

Fig. 4. **Before fix (a):** Inclusivity bug described *(from GenderMag forms)* will make Abi-like users feel vulnerable and prevent them from proceeding with the step, as she is risk-averse. **After fix (b):** Allowing Abi to choose from a list of options, and **(c)** clearly explaining the motive, and giving sense of security supports the facets: Motivation, Learning by Process vs. by Tinkering, and Attitude towards risk.

Next, they assigned the severity (impact level) of the bugs based on if the bug is part of a key vocabulary/terminology mismatch and therefore can create recurring issues.

For example, when evaluating one of the basic steps of downloading an ID card, the team found several issues with the UI. First, there was a problem with the workflow of getting the ID card, second there were key terminology mismatches, as evaluators identified: "*Terminology changed to get UIN/VID- what should Abi do with this? Should Abi go back and do something else? 'Download option' [for the ID card] still not available. Instead of Get UIN/VID, 'download card' might help*" [GenderMag form]. When discussing the fix for this bug, P5 commented: "*... even I will not click on that button... let's say there's a team of five people and all five of them or maybe four of them are saying that my Abi will not go ahead... I think that's the kind of bug that we would want to fix ...*" [P5]

Post prioritization, JIRA ticket(s) were opened and the team leads selected which bugs would be fixed in the two-week sprint cycles and assigned developers to fix the bug. Depending on the team's capacity (typically five developers), at least two high impact inclusivity bugs were selected for each sprint. The number of inclusivity bugs that could be added to a sprint depended on the urgency of other open issues (e.g., critical client request or security issues).

For each inclusivity bug, the UX designer updated the UI according to potential design fixes noted in the forms. The UI screens were then presented to the product team in a 1-2 hours sessions where different stakeholders gave feedback. As stated by P2, "*... we wear the GenderMag hat and we look at it from that lens and then we give the feedback, while our head of engineering will look at completeness of the feature...*"; and that the design catered to all users, i.e., Abi, Tim, and everybody, "*... OK, let's keep it for the Tims of the application. But let's also introduce a way that Abi can very prominently make out that 'I can actually do something over here'.*" [P2]

C. Training Advocates

A key part of integrating GenderMag into regular processes was identifying and training GenderMag advocates. This was a conscious decision made by the leadership to sustain gender-inclusive design in the organization over time. These advocates were product owners at the L2 level and were chosen based on three criteria: (1) they were proponents of gender inclusivity, (2) as product owners they had sufficient knowledge about the application and its features, and were committed to the success of the product, and (3) they had sufficient authority in the team to organize evaluation sessions and select inclusivity bugs when triaging bugs as per project demands. As P8 noted: "*product owners or product managers are the right choice to lead this kind of work, software architects or software engineers... they may not have full product perspective*." Both the product owners are now certified GenderMag trainers.

D. GenderMag Moments

One benefit of in-depth experience of performing GenderMag evaluations is the ability to apply its components just-in-time during design meetings for quick evaluations. This practice, which was adopted as is from [25], enabled the teams to significantly expedite their design process. As P2 stated: "*... will Abi know what to do? Would Abi have formed this goal in her mind?... those are the questions we started asking ourselves when we were not just re-evaluating Inji, but also designing the remaining applications parallelly.*"

They concluded by saying that they don't require full GenderMag sessions anymore, as 'thinking from Abi's perspective' has become a natural part of their thought process.

TABLE V
INCLUSIVE USER INTERFACE DESIGN GUIDELINES BY MOSIP.

Category	Design Guideline	Supported Facet(s)
Goal-Oriented Design	**Clearly Indicate Multiple Paths to Achieve a Goal.**	Motivation, Learning by Process vs. by Tinkering, Attitude Towards Risk
	If multiple user goals on the same screen have cascading relationships, clearly show how these goals depend on each other.	
	Explain where users are on their journey by tying actions back to user goals.	
	Use a linear flow and ensure steps build upon each other.	
Supporting diverse interaction styles	Provide adequate cues for user progression.	Information Processing Style, Learning by Process vs. by Tinkering
	Add clear cues to explain micro-interactions.	
	Consolidate all available user options with respect to a feature and provide clear visibility of all available user options.	
Supporting User motivation through trust	**Explain the "why" behind requested details and actions.**	Motivation, Attitude Towards Risk
	Provide assurance when people are sharing details and tell them how their data will be used.	
Support informed decision making	Provide detailed task information upfront.	Information Processing Style, Computer Self-Efficacy
	Explain warnings and negative outcomes.	
	Provide feedback adequately.	
Usability Guidelines	Maintain internal consistency.	
	Maintain external consistency.	
	Avoid cluttering and provide necessary information.	
	Use colloquial terms to explain technical jargon.	

E. GenderMag'ing Beyond Product

As also noted by Hilderbrand et al. [25], prolonged usage of GenderMag led to a mindset shift towards inclusive thinking. Their thoughts became attuned to Abi's facets and this alignment went beyond their current product development; it rewired their brains to consistently think from Abi's perspective. Team members began applying GenderMag components in their daily lives: "*I think every normal conversation in our office room will be like 'my Abi is not going to do that.*" [P5]

Similarly, another participant shared that they habitually observe their surroundings and assess how different individuals, whom they compare with various GenderMag personas, would interact in a situation as they went on saying, "*... when I have my parents using some technology versus when I have my grandparents using it versus when I'm using it versus when I have a small nephew, I tend to think about these different personas. I never used to think about it that way. Now they have suddenly become personas for me... just became a natural thing that's embedded in my mind.*" [P2]

F. GenderMag'ing Beyond Boundaries

This shift in mentality drove them to consider inclusivity in all aspects. Consequently, they aimed to make the entire application journey inclusive for client countries and their populations, by assisting them as well as other vendors in building inclusivity in their products. Therefore, they wanted to develop well-defined, practical, and user-focused guidelines organized to be as useful as possible to their target audience of developers and designers.

Thus, in collaboration with the two GenderMag advocates, the OSU researchers iteratively created the first set of design guidelines that were connected to the GenderMag facets. The guidelines were then shared with the rest of the team, including UX designers and the marketing team, for feedback. The guidelines were appropriately refined to eliminate redundancies, clarify language, and ensure each guideline added unique value. During this refinement, we categorized the guidelines in various ways, such as user needs and user journey, until the MOSIP team reached a consensus on their needs. The set of 16 guidelines were then grouped into four top-level categories, as shown in Table V).

Each guideline is accompanied by: (1) a concrete use case of 'Do's' (follow guidelines) and 'Dont's' by showing the 'before' and 'after' screenshots of design fixes, (2) a justification for the guideline based on GenderMag reasoning - why a persona would face a specific inclusivity bug in a given situation, and (3) the facet(s) implicated in the inclusivity bug. Fig. 4(a) gives an example of an inclusivity bug that the team found, and Fig. 4(b) and (c) providing design fixes. Each of the design fixes is annotated with the design guidelines (bolded items in Table V).

Note that some of these guidelines (6 out of 16) overlap with Nielsen's Heuristics [34], marked in gray in Table V. The team specifically wanted to create a *single set of comprehensive guidelines* that were relevant to their context (Digital Public Goods (DPG)) including those that overlapped with Neilsen's heuristics, so that their designers had a uniform source of guidelines to be used as a toolkit. The final set of guidelines [35], was presented at MOSIP Connect [36], their user-centered conference, where they highlighted the importance of inclusive design. These guidelines along with the recent advancements in GenAI can result in tools that can automate (parts of) the GenderMag evaluations [18, 11].

VII. LIMITATIONS

Like any empirical research, our study has limitations. The goal of a conceptual replication is to identify which methods generalize to other context(s). Here, we identified which practices from [25] generalized and which needed adaptation, and the pitfalls encountered. While we triangulated our findings across three different teams, further studies are needed to generalize across other kinds of organizational contexts. Since

national rollouts of the applications are still in progress, another limitation is that there was no post-fix user study. Hence, we could not independently confirm the effectiveness of the inclusivity fixes nor do we have data on user-reported bugs and their relation to inclusivity. However, we believe that the fixes should improve inclusivity as demonstrated in past works [24, 6, 11].

Being an Action Research study, it included some uncontrolled factors. We did not control the team dynamics or participants as they experimented with the GenderMag practices. Participants differed in their ability to understand and apply GenderMag: while some quickly grasped the concept and could easily channel the personas, some struggled, "*... as per Abi's perspective, I don't know how she will react after seeing this card. For me, it's okay... I'll be happy to see this.*" [M3-S7 (Resident Portal)].

Additionally, due to the longitudinal nature of the study results might have been influenced by potential turnover, which we did not explicitly investigate beyond the impact of newcomers discussed in Section V-A. Lastly, in line with the principles of Action Research, we worked closely with the teams, and observed how things unfolded without much intervention. As experts in this method, we guided the initial set of evaluations until the teams were confident and could proceed independently, potentially helping them avoid other pitfalls. However, several E-Signet and Resident Portal evaluations were performed without OSU researchers present, results from which we used to triangulate our findings.

VIII. Concluding Remarks

In this paper, we documented the process of integrating GenderMag into a large, distributed, open-source product development team. In total, there were 14 GenderMag practices (7 each in the Changing and Refreeze stages). In the Changing stage, 3 of the practices were used as is from [25], 3 were significantly adapted, and one was unused. In the Refreeze stage, 2 were used as is, and 4 were new. We also found two additional pitfalls. Here we highlight how the MOSIP context shaped the experience of adopting GenderMag.

In the MOSIP setting, an external impetus, a directive from the project's funding organization, played a crucial role in bringing gender inclusivity into focus. Even so, securing buy-in from key stakeholders, particularly engineering leaders, was critical for driving the adoption process. For engineering teams, which often faced resource constraints and tight deadlines, the perceived value of inclusive design had to outweigh the costs of conducting such evaluations.

Thus, identifying and training product owners as advocates was crucial for sustaining the initiative. These advocates were deeply invested in their products and committed to making them inclusive and truly useful for their end users. They ensured that inclusivity was a priority, consistently selecting one or two GenderMag-related bugs for each sprint, despite competing demands and pressures. Additionally, they played a role in training others to become advocates. This operational context was different from organizations in [25], which included a set of small teams and where the teams had buy-ins from both top-level management as well as developers.

Challenges to incorporating inclusive design arose when new hires or junior members didn't feel connected to the approach or the philosophy, treating it as mere lip service. This was further exacerbated in the remote, distributed team settings; another contextual difference from teams in [25] who mainly conducted in-person evaluations. This difference in context led to practice adaptations such as large evaluation teams, performing multi-path evaluations serially, and including inclusivity bug tracking as part of their development process. Additional strategies to overcome this challenge may include: (1) in-person training events to build a sense of value and connection to the team's mission, (2) incentives—whether monetary or tied to performance evaluations, and (3) more cost-effective methods for identifying inclusivity issues, such as the automated inclusivity detection (AID) tools [11].

Another contextual difference in our study was the fact that MOSIP creates digital public goods and includes clients who may customize the software for specific needs, this led them to customize the forms to include "how to fix" information and create the inclusivity design guidelines for broader awareness and adoption of inclusive design for their community.

One aspect where evaluating for inclusivity using GenderMag required a mindset shift was the following: Unlike running a test suite where all bugs are collected at once, triaged, and fixed, GenderMag evaluations require an iterative approach. Each evaluation takes time and this caused a significant time gap between evaluations and bug-fixing (another contextual difference that led to Pitfall #3). Additionally, addressing design fixes incrementally not only improved the evaluated features but also the UIs being developed.

Finally, as team members gained experience with GenderMag, we observed a notable shift in their engagement with the personas. Initially, they found it easier to discuss the bugs in terms of the facets instead of discussing them as gender-biases. In the earlier stages, even in the visual representation of Abi used the image of a man (Figure 4). But with time, as participants grew more comfortable discussing Abi and 'her' facets, they internalized the Abi persona. This shift in mindset occurred extended beyond the evaluation sessions and to other inclusivity dimensions (e.g., age, accessibility).

The MOSIP team started applying inclusive thinking not only to their products but also in everyday work situations, such as meetings. "*Earlier, I was not thinking from Abi's perspective. I was always thinking as [myself]... but when we started getting into the sessions more and more deeper, ... it started bringing in the change in our own mindset.*" [P2]

Acknowledgment

We thank all the MOSIP interviewees and participants. This work is supported by the MOSIP Inclusion Fellowship; the USDA National Institute of Food and Agriculture (2021-67021-35344); and NSF (1901031, 2042324, 2235601, 2303043, and 2345334).

References

[1] Anna Szlavi and Leandro S. Guedes. Gender inclusive design in technology: case studies and guidelines. In *International Conference on Human-Computer Interaction*, pages 343–354. Springer, 2023.

[2] Simeon Keates, P John Clarkson, Lee-Anne Harrison, and Peter Robinson. Towards a practical inclusive design approach. In *Proceedings on the 2000 conference on Universal Usability*, pages 45–52, 2000.

[3] Carlos Moreno Martínez, Joaquín Recas Piorno, Juan José Escribano Otero, and María Guijarro Mata-García. Responsive inclusive design (rid): a new model for inclusive software development. *Universal Access in the Information Society*, 22(3):893–902, 2023.

[4] Carina S González, Pedro Toledo, Vanesa Muñoz, María A Noda, Alicia Bruno, and Lorenzo Moreno. Inclusive educational software design with agile approach. In *Proceedings of The First international Conference on Technological Ecosystem for Enhancing Multiculturality*, pages 149–155, 2013.

[5] Emerson Murphy-Hill, Alberto Elizondo, Ambar Murillo, Marian Harbach, Bogdan Vasilescu, Delphine Carlson, and Florian Dessloch. Gendermag improves discoverability in the field, especially for women. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*, pages 2333–2344. IEEE Computer Society, 2024.

[6] 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 CHI Conference on human factors in computing systems*, pages 1–14, 2019.

[7] 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, 2017. doi: 10.1109/VLHCC.2017.8103461.

[8] 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*, 28(6):760–787, 2016.

[9] Lyndsey O'Brien, Tanjila Kanij, and John Grundy. Assessing gender bias in the software used in computer science and software engineering education. *Journal of Systems and Software*, page 112225, 2024.

[10] Amreeta Chatterjee, Lara Letaw, Rosalinda Garcia, Doshna Umma Reddy, Rudrajit Choudhuri, Sabyatha Sathish Kumar, Patricia Morreale, Anita Sarma, and Margaret Burnett. Inclusivity bugs in online courseware: A field study. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 356–372, 2022.

[11] Amreeta Chatterjee, Rudrajit Choudhuri, Mrinmoy Sarkar, Soumiki Chattopadhyay, Dylan Liu, Samarendra Hedaoo, Margaret Burnett, and Anita Sarma. Debugging for inclusivity in online cs courseware: Does it work? In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*, pages 419–433, 2024.

[12] Rosalinda Garcia, Patricia Morreale, Lara Letaw, Amreeta Chatterjee, Pankati Patel, Sarah Yang, Isaac Tijerina Escobar, Geraldine Jimena Noa, and Margaret Burnett. "regular" cs× inclusive design= smarter students and greater diversity. *ACM Transactions on Computing Education*, 23(3):1–35, 2023.

[13] Pankati Patel, Dahana Moz-Ruiz, Rosalinda Garcia, Amreeta Chatterjee, Patricia Morreale, and Margaret Burnett. From workshops to classrooms: Faculty experiences with implementing inclusive design principles. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 1035–1041, 2024.

[14] Pankati Patel, Jean Chu, Yulia Kumar, Daehan Kwak, Patricia Morreale, Rosalinda Garcia, and Margaret Burnett. Implementing inclusive software design in the cs curriculum. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, pages 1272–1272, 2023.

[15] Lara Letaw, Rosalinda Garcia, Patricia Morreale, Gail Verdi, Heather Garcia, Geraldine Jimena Noa, Spencer P Madsen, Maria Jesus Alzugaray-Orellana, and Margaret Burnett. Educating educators to integrate inclusive design across a 4-year cs degree program. *arXiv preprint arXiv:2209.02748*, 2022.

[16] Brett Stoddard, Theing Mwe Oo, and Heather Knight. A user interface for multi-robot furniture.

[17] Italo Santos, Katia Romero Felizardo, Marco A Gerosa, and Igor Steinmacher. Game elements to engage students learning the open source software contribution process. *arXiv preprint arXiv:2407.04674*, 2024.

[18] Amreeta Chatterjee, Mariam Guizani, Catherine Stevens, Jillian Emard, Mary Evelyn May, Margaret Burnett, and Iftekhar Ahmed. Aid: An automated detector for gender-inclusivity bugs in oss project pages. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1423–1435. IEEE, 2021.

[19] Christopher Mendez, Andrew Anderson, Brijesh Bhuva, and Margaret Burnett. The gendermag recorder's assistant. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 283–284. IEEE, 2018.

[20] 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: 18th International Conference on Asia-Pacific Digital Libraries, ICADL 2016, Tsukuba, Japan, December 7–9, 2016, Proceedings 18*, pages 45–50. Springer, 2016.

[21] Tanjila Kanij, John Grundy, Jennifer McIntosh, Anita Sarma, and Gayatri Aniruddha. A new approach towards ensuring gender inclusive se job advertisements. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, pages 1–11, 2022.

[22] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. Finding gender-inclusiveness software issues with gendermag: A field investigation. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2586–2598, 2016.

[23] Charles Hill, Shannon Ernst, Alannah Oleson, Amber Horvath, and Margaret Burnett. Gendermag experiences in the field: The whole, the parts, and the workload. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 199–207. IEEE, 2016.

[24] Mariam Guizani, Igor Steinmacher, Jillian Emard, Abrar Fallatah, Margaret Burnett, and Anita Sarma. How to debug inclusivity bugs? a debugging process with information architecture. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, pages 90–101, 2022.

[25] 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.

[26] MOSIP. People of MOSIP. https://mosip.io/people.

[27] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods*, pages 105–140. 1994.

[28] Rick Spencer. The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 353–359, 2000.

[29] Margaret Burnett and Anita Sarma. GenderMag. "https://gendermag.org/", 2015.

[30] Gillian R Hayes. Knowing by doing: action research as an approach to hci. In *Ways of Knowing in HCI*, pages 49–68. Springer, 2014.

[31] Supplemental package. https://zenodo.org/records/13911763.

[32] Jacob Cohen. Inter-rater Reliability. https://pubmed.ncbi.nlm.nih.gov/23092060/, 1960.

[33] Atlassian. JIRA. "https://www.atlassian.com/software/jira", 2002.

[34] Jakob Nielsen. 10 Usability Heuristics for User Interface Design. "https://www.nngroup.com/articles/ten-usability-heuristics/", 1994.

[35] MOSIP. Inclusive User Interface Design Guidelines. https://mosip.io/program_partner/User-Interface-Design-Guidelines.pdf, 2024.

[36] MOSIP. MOSIP Connect. https://mosip.io/news_events/mosip-connect-2024, 2024.