Engineering Gender-Inclusivity into Software: Ten Teams' Tales from the Trenches

Claudia Hilderbrand[1,2], Christopher Perdriau[1], Lara Letaw[1], Jillian Emard[1], Zoe Steine-Hanson[1], Margaret Burnett[1], Anita Sarma[1]
[1]Oregon State University, Corvallis, OR 97330, USA· [2]Pacific Northwest National Laboratory, Richland, WA 99352, USA
[1]{minic,perdriac,letawl,emardj,steinehz,burnett,anita.sarma}@eecs.oregonstate.edu, [2]claudia.hilderbrand@pnnl.gov

ABSTRACT

Although the need for gender-inclusivity in software is gaining attention among SE researchers and SE practitioners, and at least one method (GenderMag) has been published to help, little has been reported on how to make such methods work in real-world settings. Real-world teams are ever-mindful of the practicalities of adding new methods on top of their existing processes. For example, how can they keep the time costs viable? How can they maximize impacts of using it? What about controversies that can arise in talking about gender? To find out how software teams "in the trenches" handle these and similar questions, we collected the GenderMag-based processes of 10 real-world software teams—more than 50 people—for periods ranging from 5 months to 3.5 years. We present these teams' insights and experiences in the form of 9 practices, 2 potential pitfalls, and 2 open issues, so as to provide their insights to other real-world software teams trying to engineer gender-inclusivity into their software products.

CCS CONCEPTS

• **Software and its engineering** • *Human-centered computing* → **Human-Computer Interaction (HCI)** → *HCI design and evaluation methods*

KEYWORDS

Inclusive software, software engineering practices, GenderMag

ACM Reference format:

Claudia Hilderbrand, Christopher Perdriau, Lara Letaw, Jillian Emard, Zoe Steine-Hanson, Margaret Burnett, Anita Sarma. 2020. Engineering Gender-Inclusivity into Software: Ten Teams' Tales from the Trenches. In *42nd International Conference on Software Engineering Proceedings (ICSE 2020), May 23-29, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA. 11 pages. https://doi.org/10.1145/3377811.3380371

1 Introduction

Software has repeatedly failed diverse populations, falling short of aiding their productivity or even being usable by some populations [7, 8, 14, 24, 25, 30, 38, 43]. Such failures are serious: they marginalize people who "don't fit"—where "don't fit" can simply mean being different from the people who wrote the software. Of the many forms of diversity for which this problem arises, its connection with gender diversity is particularly well documented [3, 5, 6, 7, 8, 9, 10, 12, 14, 19, 24, 30, 31, 37, 38, 43, 44, 46].

Making software products usable to people regardless of their gender has practical importance. If software teams fail to achieve inclusiveness, their market size shrinks. If a project's development tools or products fail to achieve inclusiveness, not only is product adoption reduced, but also the involvement of women and other underrepresented populations in the teams themselves [17, 30].

A few methods have emerged to help software teams engineer gender-inclusivity into their software. One of these is the GenderMag method (Gender-Inclusiveness Magnifier) [10]. GenderMag is a method for finding—and also fixing [43]—gender-inclusivity "bugs" in software. Empirical research reports that GenderMag is effective at helping software practitioners find and fix such inclusivity bugs in their teams [10, 43].

However, little is known about whether and how busy, real-world software teams can embed GenderMag into their development processes, given the many demands on their time and the practices they already have in place. To find out, we engaged with 10 software teams via Action Research.

Action Research is a type of longitudinal field study that involves "engaging with a community to address some problem… and through this problem solving to develop scholarly knowledge" [20]. It is done collaboratively with participants—not "to" or "for" or "focused on" them. Therefore, our study was a fully collaborative endeavor with 10 software teams who were working to engineer gender-inclusivity into their software. As per Action Research's longitudinal focus, our involvement spanned months to years. Specifically, we had consistent involvement over 9 months with four professional software teams at a university, and intermittent data collection over periods ranging from 5 months to 3.5 years with six teams based in industry.

The contribution of this paper is the first in-depth "*how* investigation" into GenderMag-based processes these teams worked out to make using GenderMag practical and viable in their real-world settings, as follows:

- *How* real-world software teams went about minimizing *time costs* of blending this method into their existing development processes.

ICSE '20, May 23–29, 2020, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7121-6/20/05…$15.00
https://doi.org/10.1145/3377811.3380371

- *How* real-world software teams went about maximizing the *benefits and impact* they gained for the time they spent using the GenderMag method; but also…
- *Real-world pitfalls* the software teams ran into (and sometimes averted), potentially sabotaging their benefits.
- Practices the software teams devised to *leverage portions* of GenderMag *beyond GenderMag evaluation sessions.*
- *Unresolved issues* for which real-world practices are still emerging.

2 Background

The practices we investigate are in the context of the GenderMag method. We begin by summarizing GenderMag, a software inspection method for finding and fixing inclusivity "bugs".

GenderMag starts by helping a software team find user-facing inclusivity bugs in their own UI, using five "facets" of individuals' cognitive styles for going about problem solving. These facets form the core of the GenderMag method—an individual's motivations, computer self-efficacy, attitude(s) toward risk, information processing style(s), and learning style(s).

GenderMag literature defines inclusivity bugs as issues tied to one or more of these cognitive facets. Such "bugs" are cognitive inclusivity bugs, but also gender-inclusivity bugs because the facets capture well-established (statistical) gender differences in how people problem-solve [2, 3, 5, 7, 12, 13, 15, 18, 19, 24, 32, 38]. For example, using these facets, a software team might discover an inclusivity bug if a feature is easily discoverable by people with a tinkering learning style, but not easily discoverable by people with a process-oriented learning style.

In essence, the diverse problem-solving styles represented by the facets capture cognitively diverse behaviors that occur both *within* a given gender as well as those with statistical differences *between* one gender and another. Thus, supporting multiple facet values in software tends to make software better for people of all genders, as illustrated in [43].

GenderMag makes the five facets concrete with a set of three faceted personas—"Abi", "Pat", and "Tim". Personas [1] are a widespread technique in industry. Each persona represents a subset of a system's target users—here, their purpose is to represent differences in the facet values. Abi's facet values represent the opposite end of the problem-solving style spectrum from Tim's, and Pat's facet values are a mixture of Abi's and Tim's. Tim's facet values are most often the ones software developers tend to design for, and Abi's facet values are often overlooked. Portions of the personas that are not about the facets (e.g., appearance, demographics, experience, job title, etc.) are customizable (Figure 1).

GenderMag sets these faceted personas into a systematic process via a specialized Cognitive Walkthrough (CW) [39, 45], as follows. Evaluators "walk through" each step of carrying out a scenario, and answer questions about subgoals and actions a user would need to accomplish those subgoals (italics added to show key differences from standard CWs):

SubgoalQ: Will <*Abi/Pat/Tim*> have formed this subgoal as a step to their overall goal? (Yes/no/maybe, why, *what facets are involved in your answer*).

ActionQ1: Will <*Abi/Pat/Tim*> know what to do at this step? (Yes/no/maybe, why, *what facets …*).

ActionQ2: If <*Abi/Pat/Tim*> does the right thing, will they know they did the right thing and are making progress toward their goal? (Yes/no/maybe, why, *what facets …*).

As these questions show, identifying issues using this process includes identifying the facets that are tied with each. These facets are often key to the fixes—an issue's fix is designed around the facet that raised the issue. For example, to fix an issue that was raised for a particular problem-solving style, a team would revise that part of the UI to support *multiple* problem-solving styles: the already supported one(s) and the unsupported one(s).

In one lab study, when user experience researchers used GenderMag to identify usability issues, over 90% of the issues were validated by other empirical results or field observations, and 81% aligned with gender distributions of those data [10]. More generally, previous empirical studies have found GenderMag to be effective at identifying issues, and at pointing toward effective fixes [6, 10, 14, 37, 43]. However, there is almost no research on how teams integrate it into their real-world environments. That is the gap this paper aims to help fill.

3 Methodology

To investigate the how's of integrating GenderMag into real-world teams' practices, we worked with 10 professional software teams, 4 from a university and 6 from five companies. Our methodology for this investigation was Action Research.

3.1 The Action Research Methodology

Action Research [40] is a type of long-term field research, common in the fields of medicine and education and now emerging in various computing disciplines. Action Research has three stages:

Figure 1: Key portions of the Abi persona. See the supplemental document for complete personas.

unfreezing, changing, and freezing [27]. In the unfreezing stage, an organization decides that a change is needed. In the changing stage, the organization experiments with new processes and creates variations with an eye toward producing the outcomes they want. The refreezing stage is when the new processes and changes become established as part of the organization's processes. The stages are not strictly linear; instead organizations often loop back to previous stages.

Action Research is unlike many types of field research in two primary ways. First, it is iterative and "hands-on". Researchers work together with a community—the researchers are also participants, and the participants are also researchers [20, 40]. Second, its purpose is to develop scholarly knowledge about a problem to be solved, and to iteratively solve it [27]. Thus, in contrast to other empirical methods, formative evaluations, summative evaluations, and treatment manipulations are intertwined within Action Research and cannot be separated.

Action Research emphasizes rigor by focusing on credibility and validity. Triangulation is widely used for this purpose; it reports phenomena only when multiple data sources, data instances, and/or investigators, etc., independently arrive at the same conclusions. Section 3.3 enumerates how our data facilitated triangulation, and Section 10 shows how triangulating these data cross-validated the practices and potential pitfalls we report.

3.2 Participants and Procedures

Our study included a diverse set of teams (Table 1). We did not collect demographics of team members, but we know that at least two genders participated in 9 of the teams. A mix of software developers, user-interface designers, site administrators, and marketing experts from University X (a public university) and five companies used the method on their own projects. All the teams had an interest in trying GenderMag (see Section 4 for more on this). About half the industry teams had previously used GenderMag, whereas all of the university teams were just starting.

Some teams new to GenderMag contacted us for help getting started; others used GenderMag on their own using materials from http://gendermag.org and/or the downloadable kit (the GenderMag "user manual" [9]). For teams who contacted us for help, we followed the same general process: a pre-GenderMag meeting to show a team member how to customize a persona and help identify suitable scenarios (use-case(s)) for analysis; and then a GenderMag session, which usually included time for debriefing. We started a team's first GenderMag session by briefly introducing the method's purpose, roles, and forms (see supplemental document); and reminded them of the team's scenario and customized persona. We then coached and worked hand-in-hand with the team members during the session to whatever extent they wanted. Likewise, some team members acted as researchers—as per Action Research—devising new GenderMag practices and collecting follow-up data. After the first GenderMag session, we participated in later sessions only if a team asked us to; otherwise, teams moved ahead as they saw fit. On the other hand, some participants were researchers—as per Action Research—they devised new GenderMag practices and collected data. We also answered email questions they sent and conducted 1-2 phone interviews. The materials available to the participants and our interview scripts are provided in the supplemental document.

3.3 Data Collected and Analyzed

Central to our methodology's validity is triangulation, a cornerstone of qualitative analysis—whether the same results manifest themselves multiple times from multiple sources of evidence [36]. Toward this end, we collected data of multiple types to triangulate both within and among the teams (Table 2).

The data we were able to collect were as follows. From the GenderMag sessions we attended, we collected filled-out GenderMag forms, audio-recordings of the session(s) (which we then transcribed), the teams' customized personas, and our observers' notes. We also collected any artifacts we could, such as the teams' screenshots and/or mock-ups. We then followed up the sessions with semi-structured interviews when possible, and in cases in which further data was offered (e.g., follow-up meetings, emails, public postings), we collected those too. When the collection of materials from a GenderMag session was not permitted or viable, we interviewed these teams (Table 2). The interview questions are enumerated in the supplemental document.

At the end of the data collection period, we offered a post-study interview and debriefing, both to update the data we had collected, and to see if teams had tried practices we had not witnessed in that team but had observed in other teams. At the end of the interviews, we shared study results with teams to let them see their contribution to the research and to show our appreciation of their work.

To analyze these data, we borrowed techniques from grounded

Team and timespan	Max # members at session(s) in our data	Applications these teams were working on
A: 1 year	6	Information for instructors and students about academic technologies
B: 5 months	Unknown	Interface for an AI product
C: 9 months	5	Analytics and reports for staff to gain insights into university trends
L: 1 year	7	Document technologies
M: 9 months	2	Education platform for instructors
N: 3.5 years	>12	An IT-support product for end users
O: 1 year	2	Search engine
P: 1 year	>7	Web based interface for visual sorting with a deep learning back end
W: 1 year	3	Web application for employees who manage web content
Y: 9 months	7	Application for customer communities

Table 1: The university and industry teams in our study, and the timespan over which we were able to intermittently collect data.

theory and used triangulation for validity. Specifically, two authors went through all data of all types (Table 2) and marked "ground-up" all entries about "process", which filtered out entries not relevant to "process". We then added memos to clarify the context of each, how it arose, and what it was trying to achieve/avoid, then thematically grouped them iteratively into practices/pitfalls. We then applied the inclusion criteria (below) to filter out practices/pitfalls lacking enough data and triangulated the rest as a final validity check (which we return to in Table 7).

We applied two sets of inclusion criteria. One inclusion criterion was that every practice/pitfall we report here had to have occurred in at least two independent occurrences or teams. Our purpose was to raise the likelihood that the practices/pitfalls would be applicable to other real-world teams looking for guidance on using GenderMag to make their software more inclusive. A second inclusion criterion was that every practice/pitfall included here had either not appeared in refereed publications, or had added new rationales/benefits/costs not previously reported (summarized later in Table 7).

4 From Unfreezing to Changing

In Action Research, the unfreezing stage is a necessary prerequisite to the changing stage. For University X, we were part of the unfreezing stage during the course of the investigation. University X had already reached Action Research's unfreezing stage and beyond from a *general* diversity and inclusion perspective, but not yet reached it from a gender-inclusive *technology* perspective.

At the time this study began, the CIO's office had just decided to explore the possibility of incorporating GenderMag into some of their IT processes. They funded a graduate student to help move it forward, began regular meetings, and arranged for the researchers to present the GenderMag method to a group of IT teams to see if any would want to step forward. We presented it at a campus IT meeting, and as Section 3 has mentioned, a number of teams expressed interest in trying it out. We report on those teams with whom we have the longest involvement.

Team	First GenderMag session				More ses-si-ons	Other mtgs	Inter-views	Emails, soc-media, shout-outs
	Form	Rec.	Pers.	Obs.				
A	✓		✓	✓	✓	✓	✓	✓
B						✓	✓	✓
C	✓	✓	✓	✓			✓	✓
L	✓	✓	✓	✓	✓		✓	
M							✓	
N		✓		✓	✓	✓	✓	✓
O	✓				✓	✓	✓	
P			✓	✓	✓		✓	✓
W	✓	✓	✓	✓	✓		✓	
Y	✓	✓	✓	✓	✓	✓		✓

Table 2: The multiple sources of data we were able to obtain from multiple teams enabled extensive triangulation. (Form=forms filled out by team during the session. Rec.=audio recording of session. Pers.=team's customized persona. Obs.=observers' field notes.)

The six industry teams in this paper were located in five companies at which the importance of diversity and inclusion had also been accepted. They had heard about GenderMag from presentations or papers and had expressed interest in trying it.

These events brought the teams to the outset of Action Research's change stage in a tentative way. Still, for busy software teams, changes in process can be expensive, so teams needed to work out whether the upfront *costs* (time) of changing their processes to engineer inclusiveness into their software would pay off in useful and impactful *benefits*, as the next sections consider.

5 Results: Minimizing Costs

To minimize their costs of running GenderMag sessions, teams worked out several practices—but also ran into two pitfalls. Table 3 summarizes, and we detail them in the next subsections.

5.1 Learning GenderMag vs. Doing GenderMag

Some teams wanted to get started with GenderMag immediately, but this sometimes led to incompatible goals for a GenderMag session—using the session to enable an entire team to learn GenderMag hands-on versus using a GenderMag session to do GenderMag evaluations to get the needed product fixes underway.

The incompatibility came from group sizes. Including many team members in a GenderMag session had at least two advantages consistent with those experienced by earlier teams [21]: (1) more of the team got (hands-on) experience with the method; and (2) more people in the room during the session brought diverse perspectives during the evaluation, which tended to increase the completeness of the evaluation.

Team A was one of the teams who decided to include a large group (seven team members) in their first GenderMag session. Their context was a website for instructors and students (Table 1), so they made the Abi persona an instructor (Figure 2) and evaluated the scenario: "<Abi wants to> find instructions to add a TA to a course site." At the time of that session, they had not differentiated the *learning* vs. *doing* goals.

For Team A, both advantages of having a large team materialized. Regarding hands-on learning of the method, all seven members actively engaged in the session. The team's designated recorder took detailed notes, with some other team members taking their own notes as well.

The second advantage materialized too: the relatively large size of the group helped bring out diverse perspectives, because the process captures the union of perspectives of everyone at the session—not just the more vocal people in the room.

However, the large group size slowed down the evaluation: the

	Practices or Potential Pitfalls	Team
1	Learning GenderMag hands-on vs. doing GenderMag	A, M
	Beyond our control	C, L, M
2	Multi-path evals	A, C, L
3	Abstracting beyond	A, C
	Evaluating a proxy	C, W

Table 3: The teams' practices, and two pitfalls (shaded in gray) teams ran into, in minimizing their costs.

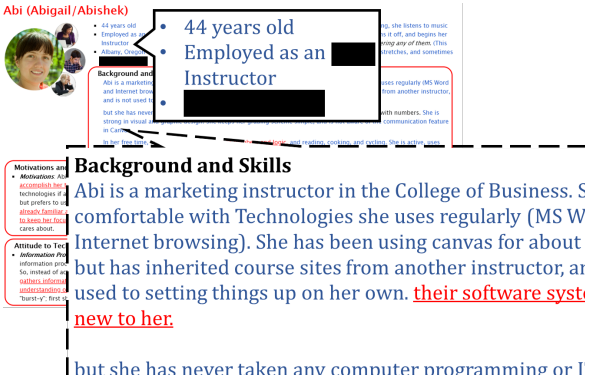

Figure 2: Team A customized Abi to be an instructor by filling in the customizable parts of Figure 1. Blue text was customization, red text was fixed (not customizable).

more people's opinions to capture, the more time was spent on each question. During the two-hour session they finished only one scenario (14 evaluation steps), not the two scenarios the team had planned to evaluate. The team decided that this pace was probably too time-costly to be viable.

During a follow up meeting, Team A decided that, to get enough GenderMag'ing done on their product, they needed to reduce the evaluation sub-team to just three members. This change also clarified who was accountable for follow-through on the issues they found.

*TA-2 *: "...we are ... going to pair up based on whose people's time and availability align with moving forward"*

* "T" teamname-datasource: e.g., TA-2=line 2 of a transcribed recording of Team A, and TA-Email means an email message from Team A.

TA-Email: "...we should be able to run through the full GenderMag process again with the two tasks above... it should provide a decent template for building a lot of the rest of the website."

This enabled Team A to proceed much more efficiently, and within a few months they released their redesigned product.

Practice 1: Learning GenderMag hands-on vs. doing GenderMag. Teams noticed that the goal of learning GenderMag hands-on had incompatibilities with doing GenderMag. Large groups seemed to be best for learning, small groups best for doing.

But small teams were not a panacea, as Team M found out. Team M started out trying to combine the "learning" with the "doing" in a single session, especially since Team M was unsure whether the method would even be useful:

TM-9: "...it was very complicated to explain to them why this was different and even though they were receptive to it, it was difficult to argue why a different type of persona was useful."

Thus, given this uncertainty about the method, Team M started with a small evaluation group: just two team members in their first session. The session went reasonably well, but after the session, they ran into a problem. They found themselves unable to communicate the need to fix the problems they had found to the team members who "owned" those parts of the system—who had not learned the method along with them:

TM-16: "...our supervisor <said> 'Why are you telling me all this?'"

This led them into the pitfall of being unable to change the aspects of the software they had evaluated, because they were not its decision makers—and had not involved the real decision makers in either the GenderMag learning or the doing. This pitfall arose here because of an "ownership" problem, but also arose in other situations that the evaluators did not really "own", such as software using third-party APIs or sub-systems not controlled by the team. All of these situations left one or more teams without the ability to act upon their results.

Potential Pitfall 1: Beyond our control.
Teams that tried to use GenderMag on (sub)systems for which they lacked decision-making power were less likely than they expected to fix the problems they found in the evaluation. Thus, the evaluation was either time wasted, or they had to spend extra time convincing the decision-makers to make the fixes.

5.2 Walking Multiple Paths "At Once"

A GenderMag walkthrough is designed to evaluate a single path (sequence of actions) through an interface—with no branching, because of the cognitive cost and group confusion of context switches between branches. However, Teams A, C, and L figured out how evaluating two small paths "at once" could increase their GenderMag method efficiency.

Figure 3 illustrates Team C's use of this practice. When evaluating their software's analytical reporting "dashboards", they ran across two different paths a user might take from a single starting place to achieve a single goal. The paths were short and diverged for only a short distance, so the team decided to evaluate both to compare them. Their multi-path evaluations paid off: they avoided re-evaluating in-common segments of the two paths. Their evaluation also revealed that the most straightforward path was not as discoverable as the alternative path, enabling the team to see why their users rarely chose the straightforward path:

TC-364: "... there's two modes of getting to the answer here, so the first mode, she'd hover on the <feature>; it doesn't tell you what to do ... She's not going to realize she has to click on the bar."

Similarly, Team L ran into multiple ways for their software to print a PDF. Comparing two possible paths with a multi-path evaluation like Team C's, Team L found an issue and a fix to make the most direct path discoverable to people with Abi's information processing style.

TL-634: "...the image isn't linked, that would be nice... also there is more than one way to download the pdf; this is the most direct way..."

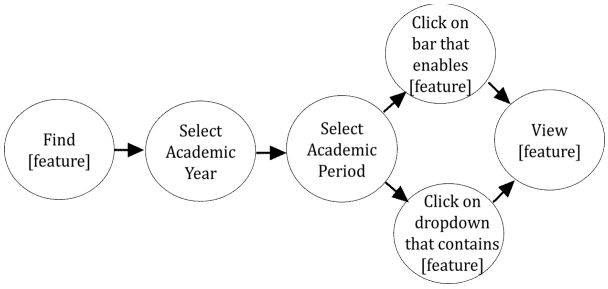

Figure 3: Team C evaluated both of the small paths that Abi could take to reach the same subgoal.

Practice 2: Multi-path evals.
Teams that did "simultaneous" evaluations of two small paths could reduce the number of sessions needed to evaluate both paths. This practice was viable when the actions started and ended at the same place and achieved the same subgoal, and also facilitated direct comparison between the paths.

5.3 Abstracting—With Discipline

A characteristic of the CW [45] family of which GenderMag is a member is that CWs are concrete, in a way analogous to testing. They take concrete inputs (for GenderMag, these are a particular customized persona, a particular scenario, a particular prototype) and produce concrete outputs for those inputs.

Despite this concreteness, Teams A and C both worked out a way to abstract beyond a session's concrete outputs. They did so by choosing for their evaluation a single instance of a UI pattern used in multiple places in their system. They then treated the single instance's evaluation as being applicable across all instances of that UI pattern, thereby eliminating the need to evaluate each instance in its own context. For example, Team C selected a "representative" analytical reporting dashboard to evaluate, with the idea of applying their results across all the instances of that dashboard in their application:

TC-3: "...it's not just for one dashboard even though we tackled just one dashboard … It's a good starting point for all our dashboards."

TC-6: "So some of the things we found in this session are definitely going to apply across the board..."

Practice 3: Abstracting beyond.
Teams abstracted beyond one session's concrete results to entire UI patterns, enabling them to reuse their findings and fixes.

On the other hand, it did not pay to be ad hoc about abstracting from one UI's evaluation to multiple instances of the UI. For example, Team C had hoped to evaluate an application that had recently been updated, but brought a machine to the session whose system was out of date. They tried to evaluate the new system using the older one as a "proxy", but this caused problems. It slowed them down, confused them, and lost the clear connection to what information Abi would and would not see:

TC-15: "...in the real environment, there wouldn't be…these other tabs."

TC-20: "So it might not have the styling …"

Even more problematic, the workflow and features that *were* available in the new interface to the users were *not* evaluated.

Thus, the practice of "abstracting beyond" paid off when it was used with discipline, i.e., only for multiple instantiations of a single pattern, but not when systems were merely "similar."

Potential Pitfall 2: Evaluating a proxy
Teams who tried to evaluate a "similar" system to the one they really cared about ended up evaluating things that were present in the proxy but not the real system, omitting things that were in the real system but not the proxy, and/or spending extra time during the evaluation trying to keep the differences straight.

6 Results: Maximizing Benefits

Teams worked out several ways to maximize the benefits they got from their GenderMag sessions. Table 4 summarizes these practices, which we detail next.

6.1 Abi's Powers

For our teams, the Abi persona was a powerful tool in two ways: (1) the strength of Abi's "inclusivity lens" brought their attention to users who were "not like we were imagining" and (2) the ways their Abi's empowered team members to talk about inclusivity issues in their software.

Previous studies have reported more inclusivity bugs when using Abi than when using the other GenderMag personas [8, 30]. Abi's lens strength may be because Abi-like populations tend to not be like the users developers were imagining when they made their design decisions. Under this hypothesis, the GenderMag kit [9] proposes that Abi offers the strongest lens, and all the teams decided for this reason to use Abi first. Some teams also used other personas: Team N also used Pat and Team O also used Tim.

Team M also had another "not like we imagined" reason for using GenderMag on their web application for Computer Science instructors. That team chose Abi to explore a user population who, despite being tech-savvy, had lower computer self-efficacy than their peers:

TM-14: "We chose to use Abi … because we wanted to explore a user with low self-efficacy with the technology, … it's hard to explain to our … team members why somebody with multiple PhD's … would blame themselves <for problems with the interfaces>"

Team N also chose Abi, but for an opposite "not like we imagined" user: to find inclusivity bugs for users who are *not* IT-savvy:

TN-21: "we primarily relied on the Abi persona … because we decided to err on the side of targeting… people who are expressly not IT people. <Abi's> attitude towards technology <risk> really tended to play a role."

Practice 4: Abi first.
All the teams used Abi as their first persona; some because Abi seemed to offer the most *powerful* inclusivity lens, and some to focus on a *particular* relevant but overlooked population.

Second, Abi empowered certain kinds of communication. Abi served as both an alibi and armor, such as by giving team members a way to provide design suggestions safely. For example, for Teams M and N, using Abi to communicate design problems averted implying that specific designers or developers had done "bad work", such as in the following examples pointing out places in the UI without enough information to satisfy comprehensive information gatherers like Abi (facet: Information Processing Style):

TN-190: "...we have the Devs who designed this UI and it was like once they were Abi they could let go of their ego. And they were really like, you're right Abi wouldn't understand this."

TM-17: "...when we brought it up to our operations lead...we kind of stressed...we feel that a professor who thinks and acts like Abi would have...been confused"

Abi also helped Team C talk about subsets of users while avoiding

	Practices	Team
4	Abi first	A, B, C, L, M, N, O, P, W, Y
5	Speaking through Abi	C, M, N
6	Calculating bias	L, N, W

Table 4: The teams' practices and potential pitfalls for maximizing their benefits.

potentially sensitive discussions about particular users or user populations:

TC-6: "…it was awesome that we had Abi to…be the user…Abi gave us the springboard to be able to talk about that and not necessarily feel bad…"

Practice 5: Speaking through Abi
Teams used Abi to ease potentially contentious or uncomfortable design discussions by framing critiques from Abi's perspective and talking about user groups by talking about Abi.

6.2 Calculating Your Software's Bias

At the end of their GenderMag sessions, three teams calculated bias by looking at the number of inclusivity bugs they found in their software. In doing so, teams followed the GenderMag convention of considering an issue an inclusivity bug if they had tied it to one or more of the facet values in the persona they had used—because issues tied to the facets disproportionately affect users who have those facets [8].

To make the calculations, the teams counted the number of evaluation questions they had answered; this became the denominator. They then counted the subset of "No" and/or "Maybe" responses tied to a GenderMag facet; this became the numerator. The resulting fraction is the percent of evaluation questions that revealed an inclusivity bug. For example, Figure 4 shows one team's bias calculations from one of their GenderMag sessions. In that session, 51% of the questions they answered showed the presence of inclusivity bugs.

These calculations turned out to be quite compelling to the team members and led to "big picture" discussions of three types. First, the teams began to realize how much they had been relying on assumptions about how their populations problem-solve.

TW-523: "... <Abi> violates a lot of our assumptions around…our tech."

TN-DebriefRecording: "Abi as defined probably would not do it <succeed>... if someone were <discussion of several facets> ... we need to accommodate that"

Second, the act of calculating bias generated considerable thoughtful discussion about the facets themselves and how they applied to different people. Team members started explaining the facets to each other, and even claiming some as their own:

TN-DebriefRecording: "My personality falls somewhere between Abi and Tim. I'm a read-the-manual kind of person, … I'm super risk-averse…"

TN-FieldNotes (not the same person as above): "I'm Abi!"

Third, they realized the importance of fixing the inclusivity bugs they had found, sometimes using the facets to categorize the bugs:

TW-506: "I would be interested in knowing more about how we can fix any of these problems. Let's pick any and let's go through whether this actually fix it."

TW-523 (not the same person as above): "... I think a lot of the failings were based on the fact that we assume users will explore the system…"

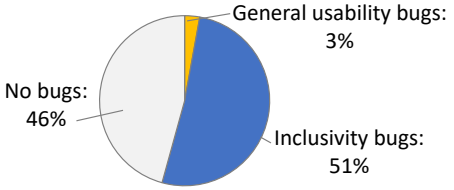

Figure 4: One team's bias calculations. (%'s are out of 35 total evaluation questions answered). If teams answered No and/or Maybe without marking a facet, the issue was noted as a general usability bug (yellow); if a facet was also marked, the issue was considered to disproportionately affect people whose cognitive styles match that facet (blue).

Practice 6: Calculating bias.
Calculating their "bias scores" from the GenderMag forms they had filled out led the teams into "big picture" reflections about their populations, the facets they were overlooking, and where to get started fixing the inclusivity bugs they had found.

7 Results: Beyond the Session

Teams also worked out practices that extended beyond GenderMag evaluation sessions, as Table 5 summarizes.

7.1 GenderMag'ing Beyond Products

Four teams surprised us by bringing GenderMag facets and personas beyond analytical evaluations. With these teams, GenderMag started influencing their user study recruitment, helped to bring inclusivity to the forefront of their workplace conversations, and even turned up in their daily lives.

For example, Team B and Team N decided to use GenderMag facets to pick the participants for their upcoming user study, to ensure representation of a diverse set of cognitive styles:

TB-Email: "…<add> facet-related questions for the screening document."

Team C reported that, once the method began to spread at University X, it increased awareness of and conversations about important gender issues:

TC-PI: "The suggestive emails from colleagues… got <us> thinking about an issue that's prevalent today."

Some even started to notice Abi's applicability far beyond the GenderMag endeavors in their workplace, noticing, for example, that the UI in gym equipment appeared to be optimized for people with a tinkering learning style:

TA-108: "I totally had an Abi moment at the gym!"

Practice 7: GenderMag'ing beyond products.
Teams brought aspects of GenderMag beyond their internal product evaluations, leveraging it for user studies, seeing it generate diversity/inclusion conversations at work, and noticing its applicability to other environments in their lives.

7.2 Analyzing Real Users' GenderMag Facets

For some teams, leveraging bits of GenderMag for use with real users went beyond simply recruiting for user studies. Four teams also worked out multiple ways to leverage GenderMag to analyze their real users as well.

	Practice	Team Name
7	GenderMag'ing beyond products	A, B, C, N
8	Facet survey	B, N, O, Y
9	GenderMag Moments	A, B, C, N, O, P, W

Table 5: Teams' "beyond the session" practices.

This practice started when Team N decided to do a survey to find out what facet values their own user populations had. Team N had a history of using surveys to categorize their user populations, so they merged portions of their existing surveys with questions like the one in Figure 5. Some of the questions they added (including the ones in Figure 5) came from literature searches for validated questionnaires, and others had to be worked out from scratch.

Team N later shared their facet questions, and Team B and Team O then started using the questions to help *analyze* data from their lab studies. For example, Team O grouped the inclusivity bugs they found by the facet values that had revealed them. This helped guide their work toward fixing these inclusivity bugs—and to then measure whether the fixes actually made their system more inclusive. Their lab study revealed that the resulting system was indeed more inclusive and was generally as good or better than the original across almost *all* of the facet values.

Practice 8: Facet survey.
Teams brought the facets into survey questions to measure their real users' facet values in multiple ways *besides* recruiting for user studies. They also used them to: (1) understand their user populations, (2) analyze their lab study data, and (3) measure the effectiveness of their fixes, facet by facet.

7.3 GenderMag'ing in a Moment

Team N was first to tell us about a practice we will term GenderMag Moments. They shared the practice, and seven teams ultimately used it. GenderMag Moments, tiny fragments of a GenderMag session, are triggered just-in-time by some kind of design question (e.g., "should we show the choices alphabetically or in sequence?") In a GenderMag Moment, team members already familiar with the full method, personas, and facets, answer the two GenderMag action questions in the context of the trigger:

ActionQ1: Will <Abi/Pat/Tim> know what to do at this step? (Yes/no/maybe, why, what facets ...).

ActionQ2: If <Abi/Pat/Tim> does the right thing, will they know they did the right thing and are making progress toward their goal? (Yes/no/maybe, why, what facets ...).

For example, Team A started blending GenderMag Moments into their design meetings to consider how to fix issues they had found by using the full method. At first, they did not realize they were even doing so, until one team member pointed out:

TA-31: "... we've just been doing Moments!"

Team A also used GenderMag Moments in a slightly different way. They expanded them to include referring back to the GenderMag forms they had filled out originally, to check the design fix would address all inclusivity bugs they had found.

Ultimately, seven teams used GenderMag Moments to save time and streamline the GenderMag process. This reduced the frequency of full GenderMag sessions needed, while allowing teams to continue assessing the inclusiveness of their designs.

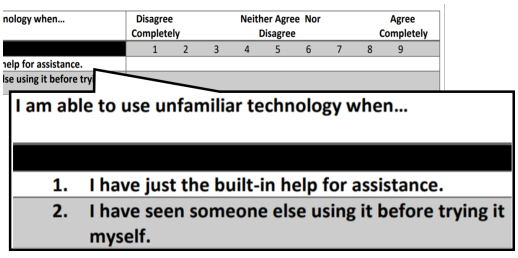

Figure 5: Facet survey. A portion of the facet survey used by some of the teams. This portion measures computer self-efficacy. The complete survey can be found in the supplemental document.

Practice 9: GenderMag Moments
Teams worked out two versions of GenderMag Moments: (1) using the GenderMag questions to guide the evaluation of design solutions just-in-time; (2) checking against the earlier sessions' filled-out forms to decide whether the fixes would address all the inclusivity bugs they had originally found.

8 The Practices Taking Hold

As the preceding sections have shown, teams had their own ways of integrating portions and variants of GenderMag into their existing practices. In this section, we consider what happened next.

We begin with the "bottom line": the products. All 10 teams decided to fix their products according to their GenderMag results. As Table 6 shows, 8 of the teams have already done so, and Team L's changes are in progress. The tenth team, Team M decided to also—but ran into the pitfall described in Section 5.1 and was not able to make the changes.

TY-Email: "...here are the changes we've made so far..." <enumerated resolved bugs related to their GenderMag analyses>.

TA-PublicPosting: "We used the GenderMag process and tools to completely redesign our website to make it easier to navigate, and to get answers quickly to commonly-asked questions."

which:

TA-PublicPosting: "...reduced help desk tickets on common questions."

Beyond specific product fixes, GenderMag also affected team practices. Some teams reported their GenderMag practices to be affecting team members' mindsets about their users, bringing diversity of cognitive styles to the forefront of their awareness:

Team	Changed the product?	Continued their GenderMag-based practices?
A	✓	✓
B	✓	✓
C	✓	✓
L	In progress	✓
M	*	?
N	✓	✓
O	✓	/
P	✓	✓
W	✓	?
Y	✓	✓

Table 6. How teams followed through. ✓: yes. ?: no evidence available. / (1/2 checkmark): team as.a whole did not reuse, but team *members* carried it to new teams. *: see text.

TC-PI: "<GenderMag> helped our team, by training people to realize that not everyone will click on stuff."

TA-PI: "was not something <we> even were aware of. <We were> not familiar with cog styles and how that might affect success when using the product."

As Table 6 shows, 7-8 of the teams showed evidence of using the GenderMag practices they had worked out (the practices reported here, and/or additional GenderMag-based practices enumerated in the supplemental document):

TC-Email: "The 'Practices / Pitfalls' handout is all around our office spaces, as a reminder…"

We do not have evidence to report on teams who did not provide explicit information one way or another on continuing their GenderMag practices. However, one phenomenon we've seen in the field is a lack of organizational ownership leading to this—employees not being empowered to spend further time on GenderMag. Organizational ownership seemed to make a difference in follow-up for some teams in our study. For example, by the time of this writing, the University X teams' practices had spread from the 4 university teams in this dataset to 11, perhaps due in part to a leadership group's report to University X's provost including a recommendation to:

X-leadership (memo): "… integrate GenderMag evaluations into <X's> regular IT practices."

9 Discussion: Heated Discussions in the Trenches

A few challenges did not produce enough independent data instances for us to include it in earlier sections as a practice/pitfall. Thus, we consider them to be issues that remain open.

9.1 Sometimes Talking about Gender is Hard

Gender bias can be a controversial topic, and some team members who were eager to fix their software's biases were less than eager to talk about them as gender biases. To those team members, the name "GenderMag" was uncomfortable:

TM-10: "I think the name GenderMag was kind of distracting. I had to clarify to people that it's about gender differences but that's not the only important part of it."

TB-64: "… I would be happier with a different name. But I didn't come up with one."

This discomfort echoes earlier reports of teams wanting to "talk about gender without talking about gender" [6, 34]. Although none of our teams reported that talking about gender took away the benefits of GenderMag, a previous group chose instead to use the vocabulary of the facets (e.g., different levels of risk tolerance, information processing styles, etc.) [6]. Another solution arose during the time of this investigation—referring to GenderMag's "family name" instead, InclusiveMag [29]. Early feedback has been encouraging, but we do not yet have field data.

9.2 Arguing over the Scenario Sequence

Earlier versions of the GenderMag method required a team member to pick an exact sequence of actions in advance as pre-work. This did not lead to arguments, but field studies showed that the pre-work was burdensome and potentially unnecessary [6, 8].

By the time of the current investigation, the GenderMag process had evolved so that the only pre-work required was to customize the persona (if desired) and name the scenario(s) being evaluated. The specific action path through the scenario was left to the team to choose just-in-time, one action at a time, as the session progressed.

This led to a new problem. For example, Team C members had different ideas about which action path to evaluate, debating at length each next step to evaluate. Such debates consumed valuable time and even led the team to try to backtrack—modify entire scenarios midstream—leading to ever more confusion.

To avoid this problem, we started coaching teams to leave deciding which step to evaluate next to a UI "driver", the person who does the actual clicking through the prototype during a session. So far, arguments over the next step in a sequence have not been reported or observed since we made this change.

10 Threats to Validity and Mitigations

No empirical study is perfect. One reason is the inherent trade-off among different types of validity [47]. Field studies, including Action Research studies, achieve real-world applicability, whereas controlled studies achieve isolation of variables.

External validity refers to the ability to generalize the findings of a study. We mitigated the risk of introducing threats to external validity by analyzing multiple teams at a university and in industry. Even so, the practices that we collected from the teams may limit our ability to generalize the use of these practices to teams outside these groups.

Internal validity refers to how the study design can influence conclusions of the study. Our study has several uncontrolled variables. For example, as an Action Research study, we did not attempt to control for teams' prior design practices or knowledge of gender issues; even had we wanted to, there is a lack of robust measurements for either. Teams and team members varied in the levels of insights they were able to gain from the method; some of these variations could have been due to the members' pre-existing ability to empathize with their users, and some could have been due to the project each was evaluating. There were also several factors that may have determined what we did and did not observe, such as team members' prior experience with inspection methods and the make-up of the teams and projects. Therefore, some interpretations we made from the data might be different had we studied different teams or projects. The practice/pitfall list is only what we observed and triangulated from 10 teams' data over this period of time. Finally, as in any Action Research study, we worked *with* the teams to help them develop solutions. As experts in this method, our contributions to the sessions we attended may have helped some teams to avoid pitfalls. Also, our mixed-gender research team's position is that methods for creating socially equitable software are critically important in Software Engineering practice, so our enthusiasm for the method may have caused us to not notice some potential pitfalls. Partial mitigations for these threats were that we were not present for all teams, that teams themselves (not us) collected some of the qualitative data, and that even teams we helped then decided alone (without us)

whether to continue.

To reduce effects of the threats above, we collected data from multiple teams and software projects and made extensive use of data triangulation, as detailed in Table 7.

11 Related Work

The most common type of SE research about gender inclusivity is in how inclusive software *communities* are [4, 22, 23, 26, 33, 41, 42, 44]. For example, researchers have shown that gender diversity within OSS communities, while limited, creates better communication structures [11, 26]. Ford et al. found that "peer parity" (having similar others for comparison) was an important factor in women's decision to engage in a software development community [16]. Mendez et al. found that gender biases in OSS tools and infrastructure can impact OSS newcomer success [30]. Lee and Carver found that some contributors used gender neutral profile names to avoid being judged because of their gender [26]. Paul et al. found that when reviewing pull requests men frequently wrote negative comments while withholding positive encouragements from women [35]. Terrell et al. found that, among new contributors (non-core members/outsiders), men's and women's pull request acceptance rate was similar when their profiles are gender-neutral but gender-biased when gender could be identified [41]. Such inclusivity bugs are problematic for both an organization's community and its productivity, as research across multiple fields has repeatedly shown. As a recent example in software engineering, Vasilescu et al.'s analysis of GitHub software projects and participant surveys found that gender and tenure diversity significantly increased productivity [42].

	First GM session	Multi GM sessions	Follow-up mtgs	Interviews	Emails	Evidence in prior lit.
Minimizing Costs						
1 Learning/doing	✓			✓		[6,21]
2 Multi-path evals	✓		✓	✓		
3 Abstracting			✓ ✓			
Beyond control	✓	✓				[8]
Eval'ing proxy	✓ ✓					
Maximizing Benefits						
4 Abi first	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓					[8,30]
5 Speak thru Abi			✓	✓ ✓		
6 Calculating bias	✓ ✓	✓ ✓				
Beyond the Session						
7 GenderMag'ing beyond products			✓✓	✓	✓ ✓	
8 Facet survey			✓	✓	✓✓✓	[43]
9 GenderMag Moments			✓	✓ ✓ ✓ ✓ ✓	✓	

Table 7: Evidence behind each practice/pitfall. The checkmarks are instances of the data sources (columns) providing the evidence. For example, we observed evidence of the "Facet survey" practice in 1 follow-up meeting, 1 interview, 3 emails, and in prior literature.

As to research in real-world practices for creating gender-inclusive *software*, there is only a little research. Williams created a collection of design process recommendations for including women in the decision-making that shapes software [46], but did not investigate them longitudinally. Also, there have been studies of GenderMag being used on real-world systems (e.g., [8, 14, 30, 31, 43]) that were not longitudinal. In these studies, teams investigating their own software have found gender-inclusivity bugs in surprisingly large fractions of their software features (reporting averages ranging from 25% and up). However, these studies did not investigate real-world teams' ways of *integrating* GenderMag into their existing practices. The only longitudinal investigation into teams' practices with GenderMag has been the short report about GenderMag at Microsoft [6]. That report covered only a few practices for integrating GenderMag into a real-world setting, and they do not overlap with the practices reported in this paper. (Additional practices our teams used, some of which are not novel, are in the supplemental document.)

Finally, research has investigated general usability inspection methods in real-world settings, such as heuristic evaluation and CWs; one notable example is [28]. However, these methods, and therefore investigations of their use, are not about how teams can engineer *inclusivity* into their software.

12 Conclusion

In this paper, we have presented a longitudinal field study in which ten real-world software teams at six institutions worked to "engineer inclusivity" into their software. The investigation spanned from 9 months to as long as 3.5 years in one team's case. The results revealed 9 practices, 2 potential pitfalls, and 2 open issues the teams worked on or encountered in combining the new method with their existing team practices and cultures. Some particularly interesting practices they worked out were:

- Even though GenderMag operates at the level of *concrete* UIs, teams *abstracted* them to UI patterns that were common in their applications (Practice 3).
- Even though GenderMag is an inspection method, teams used it to re-invent their ways of recruiting for and analyzing some of their user study methods—by leveraging the method's facets into survey and analysis instruments (Practice 8).
- Even though GenderMag is an evaluation process, teams also used it as a communication mechanism: speaking through Abi to gain both an alibi and armor (Practice 5).

This paper is the first investigation of its kind into practices of real-world teams who were exploring how to go beyond just making their software work, to making it work *inclusively* for different genders. Perhaps the central message behind these teams' experiences is that suspecting your software of gender bias and wanting to fix it are all very well and good—but integrating a systematic process can make all the difference:

TC-3: "I thought it was very, very informative … there are some things that we knew we had to change … This … gave us a process"

ACKNOWLEDGMENTS

We are very grateful to all the teams who contributed to this work. This work was partially supported by the National Science Foundation under Grant Numbers 1528061, 1815486, and 1901031.

REFERENCES

[1] Tamara Adlin and John Pruitt. 2010. *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*. Morgan Kaufmann/Elsevier, San Francisco, CA.

[2] Manon Arcand and Jacques Nantel. 2012. Uncovering the nature of information processing of men and women online: The comparison of two models using the think-aloud method. *Journal of theoretical and applied electronic commerce* 7, 2 (August 2012), 106-120.

[3] Laura Beckwith, Cory Kissinger, Margaret Burnett, Susan Wiedenbeck, Joseph Lawrance, Alan Blackwell, and Curtis Cook. 2006. Tinkering and gender in end user programmers' debugging. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM Press, New York, NY, 231-240.

[4] Amiangshe Bosu, Kazi Z. Sultana. 2019. Diversity and inclusion in open source software (OSS) projects: Where do we stand? In *Proceedings of the 2019 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM Press, New York, NY.

[5] Margaret M. Burnett, Laura Beckwith, Susan Wiedenbeck, Scott D. Fleming, Jill Cao, Thomas H. Park, Valentina Grigoreanu and Kyle Rector. 2011. Gender pluralism in problem-solving software. *Interacting with computers* 23, 5 (Sept. 2011), 450–460.

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

[7] Margaret Burnett, Scott D. Fleming, Shamsi Iqbal, Gina Venolia, Vidya Rajaram, Umer Farooq, Valentina Grigoreanu, and Mary Czerwinski. 2010. Gender differences and programming environments: Across programming populations. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM Press, New York, NY, 28.

[8] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender-inclusiveness software issues with GenderMag: A field investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM Press, New York, NY, 2586-2598.

[9] Margaret Burnett, Simone Stumpf, Laura Beckwith, and Anicia Peters. 2018. The GenderMag Kit: How to use the GenderMag method to find inclusiveness issues through a gender lens, http://gendermag.org, June 28, 2018.

[10] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness *Interacting with Computers* 28, 6-19 (Nov. 2016), 760-787.

[11] Gemma Catolino, Fabio Palomba, Damian A. Tamburri, Alexander Serebrenik, and Filomena Ferrucci. 2019. Gender diversity and women in software teams: how do they affect community smells? In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS '10)*. IEEE Press, Piscataway, NJ, USA, 11-20. DOI: https://doi.org/10.1109/ICSE-SEIS.2019.00010

[12] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G. Terveen. 2014. Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM Press, New York, NY, 674-686.

[13] Gary Charness and Uri Gneezy. 2012. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization* 83, 1 (June 2012), 50–58.

[14] Sally Jo Cunningham, Annika Hinze, and David M. Nichols. 2016. Supporting gender-neutral digital library design: A case study using the GenderMag toolkit. In *International Conference on Asian Digital Libraries*. Springer, 45-50.

[15] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association* 9, 3 (June 2011), 522–550.

[16] Denae Ford, Alisse Harkins, and Chris Parnin. 2017. Someone like me: How does peer parity influence participation of women on Stack Overflow? In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 239-243.

[17] Denae Ford, Justin Smith, Philip J. Guo, and Chris Parnin. 2016. Paradise unplugged: Identifying barriers for female participation on stack overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016). ACM, New York, NY, USA, 846-857. DOI:* https://doi.org/10.1145/2950290.2950331

[18] Catarina Gralha, Miguel Goulao, and Joao Araujo. 2019. Analysing gender differences in building social goal models: a quasi-experiment. *IEEE Intl. Requirements Engineering Conference* (12 pages).

[19] Jonas Hallström, Helene Elvstrand, and Kristina Hellberg. 2015. Gender and technology in free play in Swedish early childhood education. *International Journal of Technology and Design Education* 25, 2 (July 2014), 137-149.

[20] Gillian R. Hayes. 2014. Knowing by doing: Action Research as an approach to HCI. In *Ways of Knowing in HCI*, J. Olson and W. Kellogg (eds.). Springer, NY, 49-67.

[21] Charles Hill, Shannon Ernst, Alannah Oleson, Amber Horvath, and Margaret Burnett. 2016. GenderMag experiences in the field: The whole, the parts, and the workload. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC '16)*. 199-207.

[22] Nasif Imtiaz, Justin Middleton, Joymallya Chakraborty, Neill Robson, Gina Bai, and Emerson Murphy-Hill. 2019. Investigating the effects of gender bias on GitHub. In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*. IEEE Press, Piscataway, NJ, USA, 700-711. DOI: https://doi.org/10.1109/ICSE.2019.00079

[23] Daniel Izquierdo, Nicole Huesman, Alexander Serebrenik, and Gregorio Robles. 2019. OpenStack gender diversity report. *IEEE Softw.* 36, 1 (January 2019), 28-33. DOI: https://doi.org/10.1109/MS.2018.2874322

[24] Caitlin Kelleher. 2009. Barriers to programming engagement. *Advances in Gender and Education* 1, 1 (2009), 5-10.

[25] Andrew J. Ko and Richard E. Ladner. 2016. AccessComputing promotes teaching accessibility. *ACM Inroads* 7, 4 (Nov. 2016), 65–68.

[26] Amanda Lee and Jeffrey C. Carver. 2019. FLOSS participants' perceptions about gender and inclusiveness: a survey. In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*. IEEE Press, Piscataway, NJ, USA, 677-687. DOI: https://doi.org/10.1109/ICSE.2019.00077

[27] Kurt Lewin. 1952. Group decision and social change. In *Readings in Social Psychology*, Swanson, G.K., Newcome, T.M. and Hartley, K.L. Eds., Holt, New, York, 459-473.

[28] Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. 2010. State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction* 26, 8 (July 2010), 741-85.

[29] Christopher Mendez, Lara Letaw, Margaret Burnett, Simone Stumpf, Anita Sarma, Claudia Hilderbrand. 2019. From GenderMag to InclusiveMag: An inclusive design meta-method. *IEEE Symposium on Visual Languages and Human-Centric Computing*, (October 2019). Retrieved from https://arxiv.org/abs/1905.02812

[30] Christopher Mendez, Hema Susmita Padala, Zoe Steine-Hanson, Claudia Hilderbrand, Amber Horvath, Charles Hill, Logan Simpson, Nupoor Patil, Anita Sarma, and Margaret Burnett. 2018. Open Source barriers to entry, revisited: A sociotechnical perspective. In *Proceedings of the 40th. International Conference on Software Engineering (ICSE '18)*. IEEE, 1004-10015.

[31] Christopher Mendez, Anita Sarma, and Margaret Burnett. 2018. Gender in open source software: what the tools tell. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*. IEEE, 21-24.

[32] Joan Meyers-Levy and Barbara Loken. 2015. Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology* 25, 1 (Jan. 2015), 129-149.

[33] Dawn Nafus. 2012. 'Patches don't have gender': What is not open in open source software *New Media & Society* 14, 4 (June 2012), 669-683.

[34] Alannah Oleson, Christopher Mendez, Zoe Steine-Hanson, Claudia Hilderbrand, Christopher Perdriau, Margaret Burnett, and Andrew J. Ko. 2018. Pedagogical content knowledge for teaching inclusive design. In *Proceedings of the 2018 ACM Conference on International Computing Education Research*. ACM Press, New York, NY, 69-77.

[35] Rajshakhar Paul, Amiangshu Bosu, and Kazi Zakia Sultana. Expressions of sentiments during code reviews: Male vs. female. 2019. In Proceedings of the 26th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER '19). IEEE, Hangzhou, China.

[36] Jenny Preece, Yvonne Rogers, and Helen Sharp. 2015. *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

[37] Arun Shekhar and Nicola Marsden. 2018. Cognitive Walkthrough of a learning management system with gendered personas. 4th Gender & IT Conference (GenderIT'18), 191-198. DOI: https://10.1145/3196839.3196869

[38] Anil Singh, Vikram Bhadauria, Anurag Jain, and Anil Gurung. 2013. Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets *Computers in Human Behavior* 29, 3 (May 2013), 739–746.

[39] Rick Spencer. 2000. 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*. ACM Press, New York, NY, 353-359.

[40] Ernest T Stringer. 2007. *Action Research* (4th. Ed.). Sage, Newbury Park, CA.

[41] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, and Chris Parnin. 2016. Gender bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* 3 (Jan 2016).

[42] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd. Annual ACM*

Conference on Human Factors in Computing Systems (CHI '15). ACM Press, New York, NY, 3789-3798. DOI: https://doi.org/10.1145/2702123.270254

[43] Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 37th. Annual ACM Conference on Human Factors in Computing Systems (CHI '19)*. ACM Press, New York, NY.

[44] Zhendong Wang, Yi Wang, and David Redmiles. 2018. Competence-confidence gap: a threat to female developers' contribution on Github. In *2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. ACM Press, New York, NY, 81-90. DOI: https://doi.org/10.1145/3183428.3183437

[45] Cathleen Wharton, John Rieman, Clayton Lewis and Peter Polson. 1994. The cognitive walkthrough method: A practitioner's guide. Usability Inspection Methods. Wiley, NY, 105-140.

[46] Gayna Williams. 2014. Are you sure your software is gender neutral? *Interactions* 21, 1 (January 2014), 36–39.

[47] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. Kluwer.