GenderMag Experiences in the Field: The Whole, the Parts, and the Workload

Charles Hill, Shannon Ernst, Alannah Oleson, Amber Horvath, and Margaret Burnett
School of EECS
Oregon State University
Corvallis, OR, USA
{hillc, ernstsh, olesona, horvatha, burnett}@eecs.oregonstate.edu

***Abstract*—Recent research has reported numerous studies bringing into question the gender inclusiveness of many kinds of software. Inclusiveness of software (gender or otherwise) matters because supporting diversity matters—it is well-known that the more diverse a group of problem-solvers, the higher the quality of the solution. To help software creators identify features within their software that are not gender-inclusive, we recently created a method known as GenderMag. In this paper, we investigate the experience of teams of software professionals using GenderMag to find problems with software they are building. Our results show a high engagement with GenderMag personas—more than twice that of other personas research—and a very high degree of accuracy (93%) most of the time. Finally, our results pinpointed situations that we term "detours" that were especially prone to errors, with teams 6 times more likely to make errors in detours than they did otherwise.**

Keywords—GenderMag, Gender inclusiveness, Field study, Diversity, Cognitive walkthrough, Personas

I. Introduction

Over the past decade, research has emerged showing that individual differences often cluster by gender in the ways that men and women go about computing tasks [2, 3, 4, 5, 8, 9, 10, 11, 13, 15, 16, 17, 23, 24, 26, 27, 29, 32, 36, 37, 39, 40, 42, 43, 46, 47, 50]. Further, although many features of software tend to support problem-solving styles favored by males, far fewer support those favored by females [2, 3, 5, 9, 11, 17, 23, 24, 27, 38, 42, 46, 50, 51].

To help software creators address this issue, we devised the GenderMag method [7]. GenderMag is a method that helps software creators detect such issues in software that they are building. We recently conducted a study of software teams in the field using GenderMag [6]. Using the GenderMag method to evaluate their own software, those teams found a total of 25 gender-inclusiveness issues in the 99 user actions and subgoals they evaluated (Figure 1). In addition, they found another 30 usability issues unrelated to gender-inclusiveness, for a total of 55 usability issues.

Perhaps because of a recent media awareness of the lack of inclusiveness in the technology industry, GenderMag is already attracting interest from a number of software teams. Although the method is still in its first year and at a beta stage, it has been used by more than 20 software teams in 5 countries: the US, Canada, Denmark, Germany, and the U.K. Among these teams are the four who were included in the above field study.

This paper brings a magnifying glass to the experiences of teams who have used GenderMag. GenderMag is in some ways cognitively taxing, and we consider where the cognitive load may have tripped up the software teams or caused them to introduce errors. If they did make errors, how pervasive were these errors, and how did they impact the teams' results? On the other hand, what worked better than expected: i.e., what are the potential pitfalls we might expect them to trip them up that did *not* trip them up after all? Answers to questions like these are needed to point the way forward for works aiming to promote inclusiveness in software.

II. Background and Related Work

A. Background: The GenderMag Method

GenderMag (Gender-Inclusiveness Magnifier) is an inspection method to enable software practitioners to evaluate software they are creating from a gender-inclusiveness perspective. GenderMag has been piloted by (at least) 20 software teams across the world so far.

We have detailed elsewhere [7] the formation of GenderMag, its formative empirical work, a controlled lab study, and an early field study [6]. Thus, here we present only as much of the method as needed for interpreting the empirical results

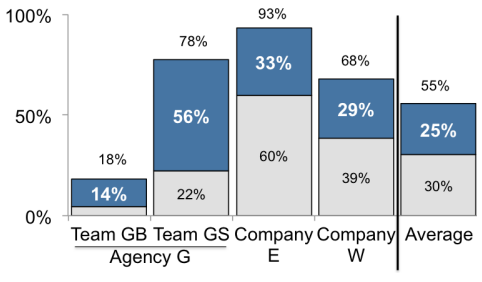


Figure 1: Issues from an early field study [6] that each team found as a percentage of the number of user actions and subgoals evaluated. Above bars: total issues. Dark blue: gender-inclusiveness issues. Light gray: other issues.

978-1-5090-0252-8/16/$31.00 ©2016 IEEE

presented in upcoming sections.

GenderMag's foundations are in people's problem-solving approaches that tend to cluster by gender, and thus its scope is software used in problem-solving situations. Toward this end, GenderMag focuses on five facets of problem-solving that have been extensively investigated in the literature. It brings these five facets to life via a set of faceted personas, and embeds their use into a systematic process based on a facet-focused specialization of the Cognitive Walkthrough (CW) [49, 51]. The five facets are:

Motivations: Research spanning over a decade has found that females are more likely than males to be motivated to use technology for what they can accomplish with it, whereas males are more often than females motivated by their enjoyment of technology per se [4, 5, 10, 23, 26, 29, 32, 46].

Information processing styles: Females are statistically more likely to gather information comprehensively—gathering fairly complete information before proceeding—but males are more likely to use selective styles—following the first promising information, then backtracking if needed [8, 15, 36, 37, 42]. Each style has advantages, but either is at a disadvantage when not supported by the software.

Computer self-efficacy: Empirical data have shown that females often have lower computer self-efficacy (confidence) than their male peers, and this can affect their behavior with technology [2, 3, 4, 5, 17, 24, 27, 32, 39, 40, 47].

Risk aversion: Research shows that females tend statistically to be more risk-averse than males [16], surveyed in [50], and meta-analyzed in [13]. Risk aversion with software usage can impact users' decisions as to which features to use.

Tinkering: Research across ages and professions reports females being statistically less likely to playfully experiment ("tinker") with software features new to them, compared to males. However, when females do tinker, they tend to be more likely to reflect during the process and thereby sometimes profit from it more than males do [3, 5, 9, 11, 26, 43].

GenderMag brings these facets to life with a set of four faceted personas—"Abby", "Pat(ricia)", "Pat(rick)" and "Tim". Each persona's mission is to represent a subset of a system's target users as they relate to these five facets. Thus, Abby, Patricia, Patrick and Tim are identical in several ways: all have the same job, live in the same place, and all are equally comfortable with mathematics and with the technology they regularly use. Their differences are strictly derived from the gender research on the five facets. Tim's facet values are those most frequently seen in males (e.g., Figure 2), Abby's facet values are those frequently seen in females that are the most different from Tim's, and the two Pats' (identical) facet values add coverage of a large fraction of females and males different from both Abby and Tim. The two Pats' identical facet values are to raise awareness that differences relevant to inclusiveness lie not in a person's gender identity, but in the facet values themselves.

GenderMag intertwines these personas with a specialized Cognitive Walkthrough (CW). The CW is a long-standing inspection method for identifying usability issues for users new to a system or feature [51]. In a GenderMag CW, evaluators answer the CW questions for each step of a detailed use case (a goal and list of actions) with respect to the five facets, from the perspective of one of the above personas.

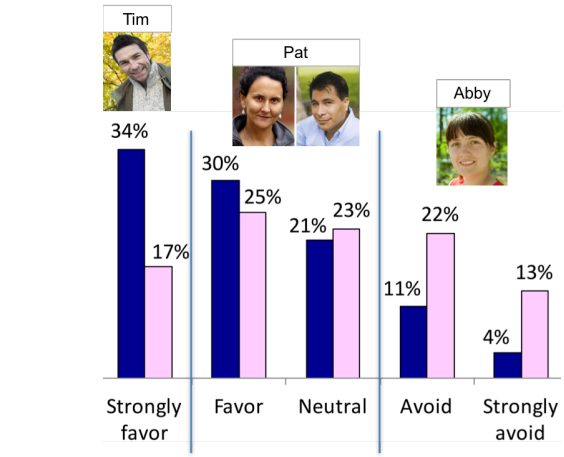
GenderMag has been evaluated in two ways so far: in a lab study [7] and in a field study [6]. Both of these studies produced encouraging results, but they leave unanswered the question of the process itself: where its strengths lie as a process, and where serious weaknesses may lie that are likely to produce faulty results. The purpose of this paper is to help fill this gap, by evaluating the process of "GenderMag'ing".

B. Related Work

1) Evaluations of Cognitive Walkthroughs

We have mentioned that a specialized Cognitive Walkthrough (CW) drives the process of GenderMag'ing. The most recent comprehensive study of cognitive walkthroughs that we have been able to locate is the 2010 survey by Mahatody et al. [31]. Their survey describes many variations of the cognitive walkthrough (CW) introduced by Lewis [30] and updated by Wharton et al. [51]. Later adaptations to the CW include having users in the CW [21] or incorporating theories of cognition [18, 44]. Other variations of the CW focus on solving problems with the classic CW [45, 49].

One of the earliest responses to CW issues other than revisions to the original method was that of Spencer [49]. That work identified constraints of CWs that reduced their usefulness. After identifying these issues, Spencer made changes to the CW method to help fix them. Streamlined CWs reduced



• ***Way of Learning Technology***: When learning new technology, Abby leans toward process-oriented learning, e.g., tutorials, step-by-step processes, wizards, online how-to videos, etc. She doesn't particularly like learning by tinkering with software …

Figure 2: (Top): A portion of the empirical foundations behind the personas' tinkering facet [5]. Percentages show the proportion of males (blue) and females (pink) represented by each value. (Bottom): A portion of the Abby persona's tinkering facet.

the amount of questions in the CW in order to relieve the issues Spencer found.

More recently, Grigoreanu et al. [22] presented a variant of the CW called the Informal Cognitive walkthrough. This method helps reduce the amount of time necessary for the CW, and helps the reliability of the CW method by including representative users. However, the method heavily relies on the skill of a researcher, which limits its usefulness in companies without research staff.

2) *Evaluations of Personas*

Personas are widely used in industry, sometimes simply to communicate about user needs during software design, such as via ideation and role-playing during informal tests, and sometimes for much more [19, 35, 38, 41]. Personas were developed by Cooper as a way to focus, clarify and understand user goals and needs [14]. Among the benefits recounted for personas is that they induce empathy towards users [1] and facilitate communication about design decisions [41]. Some reasons cited behind these benefits are that personas focus issues [25], provide common language to talk about the user [34], reduce conflict over what the goals are [1], and summarize data on users in a palatable format [20].

However, researchers have also reported drawbacks and controversy with personas. Creating personas takes significant time and effort, only too often to then be largely ignored. For example, Friess reports personas being referenced only 2% of the time in product decision-making conversations [19]. Even when personas are used with CWs as focal points [19, 28], Friess found personas to be used only 10% of the time [19].

Some of these issues with personas arise from practitioners not believing the persona is credible; practitioners finding personas to be abstract, misleading, impersonal; and practitioners finding the personas' personifying details to be distracting [12, 35]. Further, when personas are used, research suggests they are used most often by the select group of people who created the persona and have formalized training on personas, in part because they have more firsthand knowledge to the intent of the persona [35]. Those who do not aid in the creation of personas are less likely to use them in design decisions, instead preferring the data behind the persona [35].

Consistent with these findings, Marsden and Haag have reported a tension between UX designers and software developers, in which designers feel the need to argue the validity of their personas [33]. Their findings also suggest that many software developers have difficulties empathizing with personas and that, in order for software developers to accept personas, the personas had to either be grounded in empirical work, or be palatable, mainstream stereotypes.

Given this degree of controversy about personas—with as many reports of their failures as of their successes—one aspect of our paper's investigation is how the GenderMag personas contribute to or detract from its effectiveness.

III. METHODOLOGY

A. *The Field Study*

The data upon which we primarily base this investigation came from a previous field study [6]. In that study, 4 software teams used GenderMag in the wild to evaluate their own software—two teams from **g**overnment agency G, one **w**est-coast-based team of a multi-national hardware/software company (W), and one **e**ast-coast-based team at another multi-national hardware/software company (E). However, we have partial data from 16 additional teams who have used GenderMag; and when we illustrate with examples from those additional teams, we refer to those teams merely as Team Xs.

The teams learned about GenderMag from our website or from talks at conferences and meetings. When a team decided to use the method, we asked if we could observe. The context of each case was that the teams had already done the set-up necessary to run GenderMag and knew the basics of using the method, with set-up help offered when needed. Because we used the results of each session to iteratively inform and refine the method, the GenderMag method improved between some of the sessions. We observed the sessions, which usually lasted about 2 hours, and attempted to reduce effects of our presence by positioning ourselves outside the participant group (e.g., at the other end of a conference room). We also video-recorded and later transcribed each session, and collected the forms each team filled out during each session. Sessions spanned multiple software types and platforms, software maturity levels, gender make-up of the teams, and personas the teams chose to use (Table 1). Because we did not obtain videos or transcripts for the Company W's third and fourth sessions, we omit them from this paper. We also combine Company W's first two sessions here because the second was simply a continuation of the first.

B. *Qualitative Analysis*

To analyze the transcripts, we began by aligning them with answers on the GenderMag forms the teams had filled out as they talked. For each user subgoal, the form asks:

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

	Govt. Agency G	Company E	Company W
Teams & Sessions	2 mixed-gender teams (GB & GS), each team in own session.	1 session (all-male team).	4 sessions (overlapping set of mixed-gender team members).
Personas	Abby	Abby	Session 1-3: Abby, Session 4: Tim.
Software	Travel situation problem-solving.	Machine learning algorithm analyzer.	Mobile app for document delivery.
Software maturity	Very mature (10 years old).	Pre-release (initial development).	Post-release, active evolution restarting.
Software is for...	Operators capturing travel information to inform travelers.	Software developer wanting to use an ML algorithm.	Any smart phone user.

Table 1: The organizations using GenderMag on their own products covered a range of situations.

Thus, transcript dialog during the team's time working on the above question was one segment (i.e., from the time they first started this question until they moved on to the next question). The form then asked these questions about each user action to carry out that subgoal, both of which also became segments:

> Will <persona> know what to do at this step? (Yes/no/maybe, why)
>
> If <persona> does the right thing, will s/he know s/he did the right thing & is making progress toward their goal? (Yes/no/maybe, why)

We then categorized the segmented transcripts and forms using the code sets overviewed in Table 2. (Details of each code set will be presented in the relevant results sections.) Three of these four code sets were inspired by prior literature as follows.

The first and second code sets were informed by Activity Theory [48]. Activity theory defines activities as a three-level hierarchy of "has-a" relationships between subjects and objects. To understand issues arising from teams ("subjects", in Activity theory terminology) trying to conduct the GenderMag activity with its collection of objects (the prototype, the persona, the task, and the GenderMag CW forms), we coded as per Table 2 to the nodes in the activity chart in Figure 3.

Our third analysis used the Friess [19] method of measuring persona invocation, which was counting the percentage of conversational turns that invoked the personas.

Code Set	Method	Literature Source	Detailed in Section…
Activities	Dual coding: interrater Jaccard agreement=80% over 22% of the data.	Activity Theory [48]	Section IV.C
Detours	Dual coding: interrater Jaccard agreement=84% over 21% of the data	Activity Theory [48]	Section IV.C
Invoking personas	Scan for persona names and pronouns referring to personas.	Persona research [19]	Section IV.A
GenderMag CW recording errors	Dual coding: interrater Jaccard agreement=99.8% on 20% of the data.	--	Section IV.B

Table 2: Overview of our four code sets. Detailed code sets are enumerated in the sections that use them. (The bottom row's agreement was particularly high because our coding rules settled on fixed keywords and phrases to identify errors.)

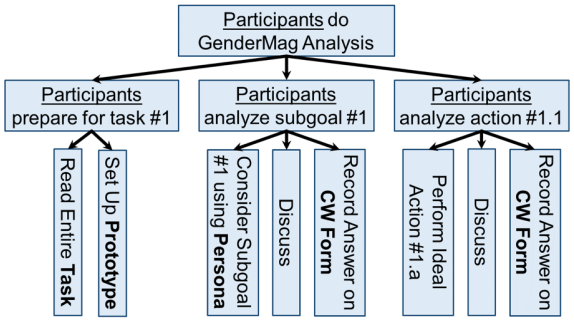


Figure 3: The GenderMag method, broken down by the stages of Activity Theory. This chart shows the first steps of the analysis phase (task #1, subgoal #1, and ideal action #1.1); subsequent tasks, subgoals, and actions repeat the same process.

Finally, we coded recording errors teams made in filling out the GenderMag forms, such as erroneously omitting issues or facets teams had identified during the discussion (detailed in Section IV.B). By matching such errors against particular activities or objects in the activities chart, we hoped to identify problematic aspects of GenderMag'ing as an activity.

IV. Results

We have mentioned that our early field study of software teams suggested that GenderMag was very effective: teams found gender-inclusiveness issues in 25% of the software features they evaluated. Here, we consider what about the process worked well in producing these results, and what about the process undermined them.

A. GenderMag's Persona(s): Are They Working?

The GenderMag personas are critical to the GenderMag method, because they are GenderMag's only means of educating those using GenderMag (i.e., the teams) about the facets and their ranges of values. Thus, if team members declined to engage with the personas as in some prior reports, they would rely more on their own opinions than on the GenderMag personas in deciding which features were problematic, which could undermine the results.

To ward off problems like those reported above, we took several measures in the design of the personas. To make them quickly digestible, we made them fit on one page and used bullets, boldface, and red, underlined text to enable readers of the persona to find the important parts. For flexibility, we also made small portions of the personas tailorable to the teams' target audience, allowing such features as profession, education, background, hobbies, age, and location to be tailored. Finally, we linked the personas to the research and data behind each persona to build credibility (http://eusesconsortium.org/gender), as per Adlin and Pruitt's notion of public "foundation documents" [1].

1) How much did GenderMag participants use Abby?

We compare the team members' invocation of personas with Friess's best-case result, in which 10% of conversational turns during a cognitive walkthrough referred to the personas [19]. Using the same method reported by Friess, we calculated the number of persona invocations per conversational turn, in which a turn began when one speaker started to speak, and ended when s/he stopped speaking. (Since all sessions in this analysis used the Abby persona, in this section we will concretely refer to Abby.) As with Friess, if a team member referred to the Abby persona by name or by pronoun, we counted it as a reference to Abby, except if team members were merely reading a CW question aloud (which contained Abby's name). To be conservative, we still included these readings in the total count of conversational turns.

We were surprised at how much GenderMag teams used the Abby persona. GenderMag teams referred to Abby in

20%–31% of the conversational turns in their teams, for an average of 23%—all of which are more than twice as often as Friess's 10%. As Table 3 shows, the GenderMag team members' rates of referring to the persona were significantly higher than the Friess counts of referring to personas (Fisher's exact test, $p<.0001$).

This raises another question: how much persona engagement is "enough" in GenderMag? One way of measuring "enough"ness is measuring the extent teams referred to Abby per step in their CW analysis. Thus, we measured the rate of persona invocation per question on the CW form (i.e., per *segment*, as defined in Section III).

The results were that Teams GB, GS, E, and W explicitly referred to Abby in 42%, 88%, 79%, and 93% of the CW segments, respectively, or an average of 79% of their CW segments. (Team GB's markedly lower rate than the other teams' may have been due to the fact that they were using the first version of the personas and the CW forms, which we improved before the other teams used them.) We view this high rate of explicitly considering Abby in 4/5 of the questions to be very encouraging.

2) Who referred to Abby the most?

In some prior work, developers who participated in persona-based sessions were often less empathetic or less involved in using personas. To consider whether this was true among the team members in this study, we counted Abby references by each of the 21 team members (across all four teams): 15 developers, 5 managers, and 1 UX (user experience) intern. Interestingly, as Figure 4 (left) shows, the developers referred to Abby *more* than the other team members did.

In fact, of the 21 team members, only one failed to refer to Abby. (This team member rarely talked at all, with only 10 turns compared to the average of 57 turns per session.) All 20 of the other team members referred to Abby multiple times, ranging from 7%–42% of their utterances, and (including the 21st team member) averaging 23% overall. This too is in marked contrast to related work pointing to disengagement of a sizeable fraction of discussants [19, 33].

	Turns that invoked personas	Turns that did not invoke personas	Total turns
Prior work [19]	94 (10%)	997 (90%)	1091
GenderMag	601 (23%)	2006 (77%)	2607

Table 3: Rate of invoking personas per conversational turn during cognitive walkthroughs. GenderMag team members' rates were significantly higher than prior results.

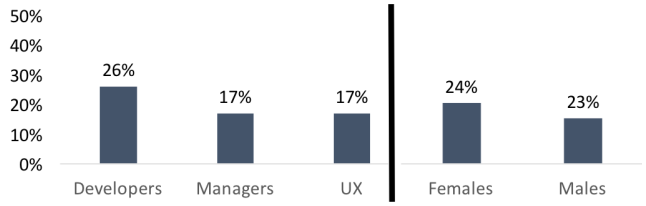


Figure 4: (Left) Managers and the UX intern referred to Abby in 17% of their conversational turns, but developers referred to her the most (26%). (Right): Females referred to Abby about the same amount as males did.

In sum, the results in this section suggest that the GenderMag persona(s) worked quite well in encouraging team members to engage in the needs of the Abby persona. We thus turn our attention to the reporting aspect of GenderMag.

B. Reporting Inclusiveness Issues: The Good and the Bad

With GenderMag, teams report the gender-inclusiveness issues they find via specialized CW forms. These forms not only record which features raised issues, but also why they are issues (free-form explanations) and what makes them gender-inclusiveness issues (listing the facets involved). The forms are key, because they are the team's only record of their decisions as to where gender-inclusiveness issues lay and why.

The success of GenderMag recording rests on the team member who has accepted the role of "recorder." The recorder has a lot to do: they must accurately record what step in the action sequence is being discussed; they must capture whether and why the team thought the action might be problematic for the persona at hand (Abby, in these examples); and they must capture which of Abby's facets caused the team to believe the action was problematic.

Fortunately, even though working with GenderMag is intense during the sessions, a session does not take a long length of time; perhaps this is why every organization in the field study has done long-term follow-up. Further, the teams' recorders succeeded in capturing a surprisingly high number of inclusiveness issues. Together they identified 22 gender-inclusiveness issues (i.e., found issues in 25% of the features they evaluated) as mentioned in the introduction. The large majority of most teams' records reflected those teams' deliberations with good accuracy (Table 4, "good" column).

But the bad news is that not all teams shared in the high rate of accuracy. Of the teams' shared total of 17 errors (second column of Table 4), Team W made 10 of them, which affected fully one-third of their segments. We have observed even higher error rates among other teams. For example, Figure 5 shows a Team Xs GenderMag form with a 57% error rate: 12 erroneous segments out of the 21 total segments.

The 9 erroneous segments in our primary data source had a potentially disproportionate impact on the inclusiveness issues GenderMag can reveal. Any segment error might mean a gender-inclusiveness issue overlooked. To see how, consider Figure 6, which breaks down the errors into five types. For example, neglecting to record facets (the most common error, illustrated in Figure 7) or explanations could cause issues to incorrectly not be counted as gender-inclusiveness issues; omitting the yes/no/maybe could prevent an issue from being counted as an issue at all.

Thus, in the worst case, if all 9 erroneous segments caused an inclusiveness issue to go unrecorded, then the correct number of gender-inclusiveness issues to record should have been

31 (22+9), rather than 22—meaning that one-third (9/31) of the gender-inclusiveness errors were missed in the worst case, due to recording errors.

C. When and How Did Errors Happen?

Team W, with its relatively high error rate, gave us our first clue into the pattern behind these errors: Team W took more detours than the other teams. For the purposes of this paper, we define a *detour* as any time team members left the GenderMag activity chart (Figure 3) in one of the nine ways described in Table 5. An example detour is a team deviating into a design discussion about how to fix a problem they just identified ("Proposing Fixes").

Spencer [49] has specifically advised CW users to avoid such detours, and his advice is on-point here. The number of errors our teams made closely aligned with how often they detoured (Figure 8). Further, only 4% of segments without detours contained errors, but 27% of segments *with* detours contained errors—an error rate over 6 times as high when detours were involved.

Of the nine kinds of detours that occurred in segments in which teams made errors, 4 types dominated (Table 5). Those four types were Troubleshooting, GenderMag Procedure, "Where are we?", and Researcher Clarifications. These four together accounted for 77 (82%) of the 94 detour instances.

For example, here the team got so disoriented ("Where are we?") that they started talking about a different part of the prototype than the one they were recording.

GS1m: ...She didn't like tinkering ... going to the balloon
GS5m: We're not on that step yet.

In this example ("Troubleshooting"), the team got so confused by odd prototype behaviors that the recorder was unable to sort out all the relevant bits he needed to record, and at least one facet was never recorded:

W4m: This is mine. And yours looks like that. So those are the two options. I don't know why.
<Team compares prototypes for two and a half minutes, talking about Abby interspersed>

Two of these four types, GenderMag Procedure and Researcher Clarification, may simply be a matter of inexperience. After a team uses GenderMag a few times, they will be less likely to ask about proper procedure or seek a researcher for clarification.

However, "Troubleshooting" and "Where are we?" seem indicative of the high cognitive load that can arise when teams attempt to attend to a prototype, the form questions, the persona, the facets, and each other all at the same time. Given this, it makes sense that a distraction (detour) in the face of this load would generate a significant rise in error rate, exactly as happened here. This suggests an opportunity for a tool that helps remind GenderMag teams where they are, so that they can in some cases avoid detours (e.g., by the tool telling them where they are), and can get back on track when they do find

Team	# good (error-free) segments	# recording errors
GB	49/50 (98%)	2
GS	43/45 (96%)	2
E	27/28 (96%)	3
W	10/15 (67%)	10
Total	129/138 (93%)	17 (in 9 segments)

Table 4: The good (1st column), and bad (2nd column) by team. Note Team W's error rate: they made over three times as many errors as the other teams.

Figure 5: This form, from a Team Xs GenderMag session, had 12 recording errors (red blobs) out of 21 answers (i.e., 57%). The recorder originally pasted down through the form, then often forgot to update the Yes/No/Maybe's.

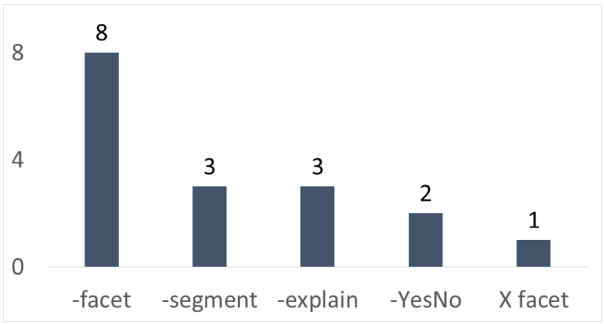


Figure 6: The five types of recording errors: "-" means "missing", and "X" means "wrong". The most common was missing facets ("-facets") with 8 instances.

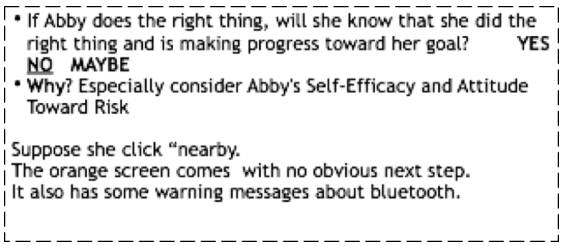


Figure 7: Example of missing facets: This Team Xs recorder captured the team's Yes/No/Maybe decisions and explanations, but he failed to capture *even one* of the facets his team discussed. For example, in this segment the team referred to Abby's risk aversion, but the recorder did not write it down.

the need to take a detour.

V. Discussion: Unique Challenges

We still have work to do on GenderMag. The current version is in beta status, and we iteratively improve it as we learn more about its strengths and weaknesses from teams like those in this paper. Further, GenderMag has unique challenges that arise from the blend of its diversity mission, its particular use of personas, and its specialized CW, a few of which we discuss in this section.

A. Stereotyping versus Promoting Diversity

One issue that we continue to monitor is stereotyping. As the related work points out (e.g., [33]), personas and stereotyping are closely related—every persona inherently represents a "representative member" of some group of target users. This close relationship is particularly discomfiting for GenderMag, because its very mission is to promote inclusiveness by recognizing *diverse* sets of users, including some often overlooked by software development groups.

Code	Example	# CW form segments
	Prototype	
Troubleshooting	GS2m: ...it's hidden …I haven't figured out … [how to] overcome this.	14
Proposing Fixes	E2m: I think I would like it better button-less.	1
Prototype Misnavigation	W1m: Whoops and I just touched the screen and lost the message.	6
Prototype Error	GS6f: Why did that come up? GS3f: I don't know.	4
	GenderMag Walkthrough	
GenderMag Procedural Confusion	GS1m: Are we making the assumption though, that this is a new piece of existing design than has been out there that Abby should be expecting to use?	21
Where are we?	W1m: ...which page should I be on?	23
Researcher Clarification	E1m: Have we just sort of abducted [Abby] and made them use [product]...? Res.: [Abby] started [the job] a week ago.	19
	Persona	
Misunderstanding Persona	GS1m: She … would feel comfortable [doing step that requires tinkering]	1
Persona Appropriation	W7m: I don't think she would have either, because [W3m] had to tell me [where] to go… and [this] before.	5
Total		94

Table 5: The frequency of appearance of each of the detours across the four teams.

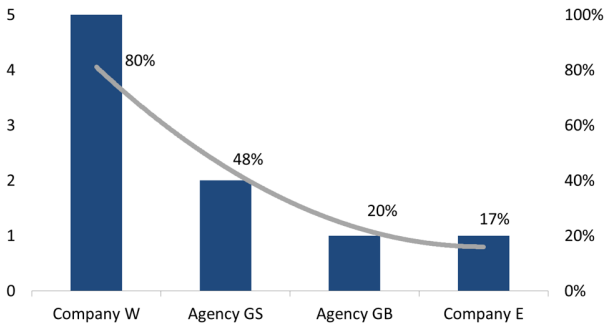


Figure 8: The bars are the number of CW form errors, and the line is the percentage of CW segments during which teams had detours. (Overall, 36% of segments contained at least one detour.)

To guard against inappropriate stereotyping, we have taken several measures. For example, we have four personas, two males and two females, to communicate that none of the personas are "the typical" male or female. We have also made two of the personas (the "Pats") twins, to emphasize that many males and females have problem-solving traits in common. We also bestowed upon all four personas identical educational backgrounds, job titles, and skills at mathematics and logic. Finally, we show the research and distribution data behind the four personas in an on-line personas foundation document, and include bar graphs like the one shown in Background and Related Work (Figure 2) in the downloadable GenderMag kit and in public presentations about GenderMag. However, we are still working on how best to navigate the challenge of promoting GenderMag's mission of inclusiveness without encouraging GenderMag's users to overgeneralize and inappropriately stereotype.

In general, gender-inclusiveness methods share a dual burden—they must both *find* issues that disproportionately affect one gender, and *educate* those in the room about inclusiveness. Sometimes, these two goals can be at odds with one another, especially when it comes to inappropriate stereotyping. We continue to work on this problem. For example, we are experimenting with a new version of the personas: the new Abby (as well as Tim and the Pats) includes photos of "other people with Abby's facet values". The aim is to prevent GenderMag users from concluding inappropriate take-aways like "all females... [do something one particular way]", by pictorially depicting gender distributions in the data behind the personas.

B. Genders of GenderMag Team Members

The genders of the members of a GenderMag team could have an impact on the way teams experience GenderMag as well as the outcomes of their sessions. For example, females or males on a same-gender team might feel "safer" than they otherwise would, to talk about controversial topics that can arise in a GenderMag session, such as inappropriate stereotyping, or gender politics in moving forward with GenderMag.

We have also seen differences in how female and male team members feel individually. For example, one Team Xs male views himself as an advocate, but is uncomfortable as a man conversing about why the gender-inclusiveness issues that come out of GenderMag affect more women than men. Some females on our teams have strongly identified with one or another of the female personas, and because of this see GenderMag as finally giving them a voice in how their team's software is being designed. On the other hand, the female personas are very different from some females on the teams, and

those females sometimes feel uncomfortably pigeon-holed into an image that is not at all representative of the way they work. We suspect that males might also encounter this feeling if they overly identify with a persona that they suspect they should not identify with (e.g., I'm secretly like Abby females, but I don't want anyone to know it.)

C. Team Sizes: More diversity, more buy-in, more errors?

Interest in GenderMag sessions at software organizations has sometimes arisen from a desire to *educate* developers on diversity. In cases like this, the teams still use GenderMag to find problems in their own software, but they also invite large fractions of their team members to attend, so that everyone can gain some insights into the diversity of individual problem-solving approaches. Some GenderMag sessions have had as many as 11 team members.

A large team size seems to have impacts on the success of GenderMag, in ways both good and bad. Because GenderMag is about supporting diversity, GenderMag CWs harvest the views of everyone in the room on each question, and record the union (not the consensus) of these views. Ideally then, the more people in the room, the greater the chance that an inclusiveness issue will be spotted. Another advantage of large GenderMag teams is more buy-in: it avoids the need to later convince a team member who was not present to fix an issue that they did not help discover.

However, disadvantages of large teams are that they slow down the process, and also seem to increase recording errors. Recall that one of the Team Xs recorders made errors in more than 50% of her form segments: her session included 7 team members. and Team W (with 9 team members in the room at its maximum) had an error rate twice as high as Team GS (the next-smaller team in our study with 6 team members), and five times as high as Team GB and Team E, who had 3 and 2 team members, respectively. This high error rate may simply be because capturing so many views makes the recorder's job much more difficult.

Given the advantages of sometimes having large GenderMag sessions—both from a diversity education perspective and from the quality that comes from collecting a diversity of viewpoints—the challenge then arises of how to best support these teams in the process. We are currently thinking about how a GenderMag "recorder's assistant" tool might better enable large teams to move through the process efficiently without losing the educational and diversity benefits that come from large teams.

VI. Conclusion

In this paper, we have investigated the experiences of four software teams who have used GenderMag to evaluate the inclusiveness of their own software. Among our results were:

- *Persona(s)*: The software teams were surprisingly engaged with the Abby persona: 20/21 team members referred to her during their GenderMag sessions—the developers even more than the UX and managers. On average, teams explicitly referenced Abby in 79% of their CW discussions. This is a significantly higher persona engagement rate than has been reported elsewhere in the literature.
- *Workload*: Teams have a heavy workload in a GenderMag session. The recorders have a particularly heavy workload: they must stay oriented to the team discussion in relation to the correct segment of the form, while accurately capturing all team members' views, explanations, and relevant facet values. Most teams handled this remarkably well, with 93% of their segments adequately captured. Still, the errors they did make had far-reaching consequences, resulting in up to a third of the gender-inclusiveness issues to be omitted.
- *Where Errors Happened*: We identified the most likely circumstances for errors to occur—two-thirds of errors occurred during detours. In fact, teams were over 6 times as likely to make an error during detours as they were when they were not detouring. Among the four most common detour types, two are likely to resolve naturally as teams become more experienced with GenderMag, and the other two seem to have good potential for tool support.

The interplay among GenderMag's diversity mission, its use of personas, and its specialization of the CW, raise unique challenges, and we still have significant work ahead in investigating how to support software teams who are encountering these challenges. We hope that GenderMag's successes and difficulties will help to inform other gender-inclusiveness methods as they emerge. GenderMag's forays into navigating the stereotype minefield and issues with cognitive load can also help inform other inclusiveness methods. We hope other researchers will join in working to address such challenges on the journey to increasing software's ability to support and nurture diverse ways of thinking and engaging with software.

VII. Acknowledgments

The GenderMag kit is freely downloadable at http://eusesconsortium.org/gender/. This work was supported by NSF #1240957, 1314384, and 1528061.

References

[1] T. Adlin and J. Pruitt, The Essential Persona Lifecycle: Your Guide to Building and Using Personas, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2010.

[2] L. Beckwith, M. Burnett, S. Wiedenbeck, C. Cook, S. Sorte, and M. Hastings, Effectiveness of end-user debugging software features: Are there gender issues? ACM CHI, 2005, pp. 869-878.

[3] L. Beckwith, C. Kissinger, M. Burnett, S. Wiedenbeck, J. Lawrance, A. Blackwell, and C. Cook, Tinkering and gender in end-user programmers' debugging, ACM CHI, 2006, pp. 231-240.

[4] M. Burnett, L. Beckwith, S. Wiedenbeck, S. D. Fleming, J. Cao, T. H. Park, V. Grigoreanu, and K. Rector, Gender pluralism in problem-solving software, Interacting with Computers, 23, 5, 450–460, 2011.

[5] M. Burnett, S. D. Fleming, S. Iqbal, G. Venolia, V. Rajaram, U. Farooq, V.Grigoreanu, and M. Czerwinski, Gender differences and programming environments: Across programming populations, IEEE Empirical Soft. Eng. and Measurement, 2010.

[6] M. Burnett, A. Peters, C. Hill, and N. Elarief, Finding gender-inclusiveness software issues with GenderMag: A field investigation, ACM CHI, May 2016 (to appear).

[7] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, GenderMag: A method for evaluating software's gender inclusiveness, Interacting with Computers, in press.

[8] P. Cafferata and A. M. Tybout, Gender Differences in Information Processing: A Selectivity Interpretation, Cognitive and Affective Responses to Advertising, Lexington Books, 1989

[9] J. Cao, K. Rector, T. H. Park, S. D. Fleming, M. Burnett, and S. Wiedenbeck, A debugging perspective on end-user mashup programming, IEEE VL/HCC, 2010, pp. 149-156.

[10] J. Cassell, Genderizing HCI, The Handbook of Human-Computer Interaction, L. Erlbaum Associates Inc., Hillsdale, NJ, 2002, 402-411.

[11] S. Chang, V. Kumar, E. Gilbert, L. Terveen, Specialization, homophily, and gender in a social curation site: findings from Pinterest, ACM Conf. Comp. Supported Coop. Work & Social Computing, 2014, pp. 674-686.

[12] C. N. Chapman and R. Milham, The personas' new clothes: methodological and practical arguments against a popular method, Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 634-637, 2006.

[13] G. Charness and U. Gneezy, Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization, vo1. 83(1), pp. 50–58, 2012.

[14] A. Cooper, The Inmates Are Running the Asylum, Sams Pub., 2004

[15] C. Coursaris, S. Swierenga, and E. Watrall, An empirical investigation of color temperature and gender effects on web aesthetics, Journal of Usability Studies vol. 3(3), pp. 103-117, May 2008.

[16] T. Dohmen, A. Falk, D. Huffman, U. Sunde, J. Schupp, G. Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences, J. European Econ. Assoc. vol 9(3), pp. 522–550, 2011.

[17] A. Durndell and Z. Haag, Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample, Computers in Human Behavior, vol. 18, pp. 521–535, 2002.

[18] J. U. Eden, Distributed cognitive walkthrough (DCW): a walkthrough-style usability evaluation method based on theories of distributed cognition, ACM Conf. Creativity & Cognition, 2007.

[19] E. Friess, 2012. Personas and decision making in the design process: an ethnographic case study, ACM CHI, 2012, pp. 1209-1218.

[20] K. Goodwin, Designing for the Digital Age: How to Create Human-Centered Products and Services, Wiley, Indianapolis, IN, 2009.

[21] T. Granollers and J. Lorés, Incorporation of users in the evaluation of usability by cognitive walkthrough, HCI Related Papers of Interacción, 2004, pp. 243-55.

[22] V. Grigoreanu and M. Mohanna, Informal cognitive walkthroughs (ICW): paring down and pairing up for an agile world, ACM CHI, 2013, pp. 3093-3096.

[23] J. Hallstrom, H. Elvstrand, and K. Hellberg, Gender and technology in free play in Swedish early childhood education, Int J. Technology and Design Education, vol. 25, pp. 137-149, 2015.

[24] K. Hartzel, 2003. How self-efficacy and gender issues affect software adoption and use, Commun. ACM vol. 46(9), pp. 167–171, 2003.

[25] K. Holtzblatt, J. B. Wendell, and S. Wood, Rapid Contextual design: A How-to Guide to Key Techniques for User-Centered Design, Morgan Kaufmann, San Francisco, CA, USA 2004.

[26] W. Hou, M. Kaur, A. Komlodi, W. G. Lutters, L. Boot, S. R. Cotten, C. Morrell, A. A. Ozok, Z. Tufekci, Girls don't waste time: Pre-adolescent attitudes toward ICT, ACM CHI 2006 Extended Abstracts, 2006.

[27] A. H. Huffman, J. Whetten, W. H. Huffman, Using technology in higher education: The influence of gender roles on technology self-efficacy, Computers in Human Behavior, vol. 29(4), pp. 1779–1786, 2013.

[28] T. K. Judge, T. Matthews, and S. Whittaker, Comparing collaboration and individual personas for the design and evaluation of collaboration software, ACM CHI, pp. 1997-2000, 2012.

[29] C. Kelleher, Barriers to programming engagement, Advances in Gender and Education, vol. 1, pp. 5-10, 2009.

[30] C. Lewis, P. G. Polson, C. Wharton, and J. Rieman, Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces, ACM CHI 1990 (CHI '90), 1990.

[31] T. Mahatody, M. Sagar, and C. Kolski, State of the art on the cognitive walkthrough method, its variants and evolutions, Int. Journal of Human-Computer Interaction, vol. 26(8), pp. 741-85, 2010.

[32] J. Margolis and A. Fisher, Unlocking the Clubhouse: Women in Computing, MIT Press, 2003.

[33] N. Marsden and M. Haag, Stereotypes and politics: reflections on personas, ACM CHI 2016, to appear.

[34] A. L. Massanari, Designing for imaginary friends: information architecture, personas, and the politics of user-centered design, New Media & Society, vol. 12(3), pp. 401-416, 2010.

[35] T. Matthews, T. Judge, and S. Whittaker, How do designers and user experience professionals actually perceive and use personas? ACM CHI, 2012, pp. 1219-1228.

[36] J. Meyers-Levy, B. Loken, Revisiting gender differences: What we know and what lies ahead, J. Consumer Psychology, vol. 25(1),129-149, 2015.

[37] J. Meyers-Levy, D. Maheswaran, Exploring differences in males' and females' processing strategies, J. Consumer Research, vol. 18, pp. 63–70, 1991.

[38] L. Nielsen and K. S. Hansen, Personas is applicable: a study on the use of personas in Denmark. ACM CHI, 2014, pp. 1665-1674.

[39] A. O'Leary-Kelly, B. Hardgrave, V. McKinney, and D. Wilson, The influence of professional identification on the retention of women and racial minorities in the IT workforce, NSF Info. Tech. Workforce & Info. Tech. Res. PI Conf., 2004, pp. 65-69.

[40] Piazza Blog, STEM confidence gap. Retrieved September 24th, 2015 from http://blog.piazza.com/stem-confidence-gap/

[41] J .Pruitt and J. Grudin, Personas: practice and theory. ACM Conf. Designing for UX, 2003, pp. 1-15.

[42] R. Riedl, M. Hubert, and P. Kenning, Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of E-Bay offers, MIS Quarterly, vol. 34(2), pp. 397-428, June 2010.

[43] D. Rosner and J. Bean, Learning from IKEA hacking: I'm not one to decoupage a tabletop and call it a day, ACM CHI, 2009, pp. 419-422.

[44] H. Ryu and A. F. Monk, Analysing interaction problems with cyclic interaction theory: low-level interaction walkthrough, PsychNology Journal, vol. 2(3), pp. 304-330, 2004.

[45] A. Sears, Heuristic walkthroughs: finding the problems without the noise, Int, Journal of Human-Computer Interaction, vol. 9(3), pp. 213-234, 1997.

[46] S. J. Simon, The impact of culture and gender on web sites: an empirical study, The Data Base for Advances in Information Systems, vol. 32(1), pp. 18-37, Winter 2001.

[47] A. Singh, V. Bhadauria, A. Jain, and A. Gurung, Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets, Computers in Human Behavior, vol. 29(3), pp. 739–746, May 2013.

[48] M. Soegaard and R. F. Dam, (eds.), The Encyclopedia of Human-Computer Interaction, 2nd ed, The Interaction Design Foundation, 2014.

[49] R. Spencer, The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company, ACM CHI, 2000, pp. 353-359.

[50] E. U. Weber, A. Blais, and N. E. Betz, A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors, J. Behavioral and Decision Making, vol. 15, pp. 263-290, 2002.

[51] C. Wharton, J. Rieman, C. Lewis, and P. Polson. The cognitive walkthrough method: A practitioner's guide, Usability Inspection Methods, John Wiley, NY, 1994, pp. 105-140.