AN ABSTRACT OF THE THESIS OF

Charles Hill for the degree of Master of Science in Computer Science presented on August 15, 2017.

Title: The Sum of its Parts: Investigating the Component Pieces of GenderMag

Abstract approved:

____

Margaret M. Burnett

Previous work introduced the GenderMag method, a software inspection method used to help software creators identify features within their software that are not gender-inclusive. 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. In this thesis, we investigate the two component parts of GenderMag: the customized cognitive walkthrough and the gendered personas. To investigate the cognitive walkthrough component, we analyze data collected from a field study that explores the experience of teams of software professionals using GenderMag. We pinpoint situations that we term "detours" during which teams were 6 times more likely to make errors on the cognitive walkthrough forms than they did outside of detours. To investigate the personas component, we present a lab study that investigates the tension between gendered personas and inappropriate stereotyping. We explore a new approach to personas: one that includes multiple photos (of males and females) for a single persona. Our results are encouraging about the use of personas with multiple pictures to expand participants' consideration of multiple genders without reducing their engagement with the persona. Between the two studies, we find answers to the question: How does each component part of GenderMag contribute to (or detract from) its overall goal of helping to identify gender biases in software?

The Sum of its Parts: Investigating the Component Pieces of GenderMag

by
Charles Hill

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 15, 2017
Commencement June 2018

Master of Science thesis of Charles Hill presented on August 15, 2017

APPROVED:

Major Professor, representing Electrical and Computer Engineering

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

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

Charles Hill, Author

ACKNOWLEDGEMENTS

“Don’t you dare switch majors again” – Mary Beth Hill

First and foremost, I thank my wife Mary Beth, without whom I wouldn’t have finished my first year of undergraduate, or my second, or my third, … or my seventh. She has constantly been there for me, never giving up on encouraging me to pursue what she knew I could accomplish. I’d also like to thank my son, Oliver, for being a constant reminder of what I’m working for, and for, on occasion, telling me to “focus, Papa!”

I also thank Dr. Margaret Burnett, my major advisor. I met Dr. Burnett during an undergraduate usability class and was fascinated with the material. Luckily, she had a position open in her lab, and so I applied. From then on, she’s encouraged me to be the best student, worker, and researcher I could be. Perhaps the most important lesson she’s taught me, both in professional and personal life, is to always ask “what if?”

I would like to thank Nicola Marsden and Maren Haag, who collaborated with us for one of the studies in this thesis. Their expertise on personas and the social issues surrounding them helped a great deal in forming and executing the study.

Finally, I thank all of the co-authors of the papers that are part of this thesis: Shannon Ernst, Alannah Oleson, Amber Horvath, Chris Mendez, and Anita Sarma. Their hard work and insights have helped shape this work through the years.

TABLE OF CONTENTS

Page
Introduction 1
Study 1: The Cognitive Walkthrough 2
Study 2: The personas 2

Literature Review 4
Background: The GenderMag Method 4
Related Work 7

Study 1 Method...11
Qualitative Analysis 11

Study 1 Results 14
GenderMag’s Persona(s): are they Working? 14
Reporting Inclusiveness Issues: The Good and the Bad 19

Study 1 Discussion...25
Team Sizes: More diversity, more buy-in, more errors?...............25
Genders of GenderMag Team Members 26
Stereotyping versus Promoting Diversity.....................................26

Study 2 Method...28
The Manipulation..28
Study 2.1(Eye tracking) Methodology 30
Study 2.2 Methodology 31

Study 2 Results 38
RQ1: How do people gender-stereotype personas.....................38
RQ2: Can we reduce stereotyping?...42

Study 2 Discussion...46
Stereotyping..46
Believability...47
Education..47
Engagement 48
Four pictures or one?...49

Conclusion..50
Study 1 Closing Thoughts 50
Study 2 Closing Thoughts 51
The Bigger Picture 52

Bibliography 55

TABLE OF CONTENTS (Continued)

Page

Appendix 1: Study 2 Materials 60

Questionnaire 1 60

Questionnaire 2: 61

Cognitive Walkthrough Forms 62

Appendix 2: Additional Statistical Results 66

LIST OF FIGURES

Figure Page

1. Issues Found in Previous GenderMag Work 1

2. The Tim Persona 6

3. Activity Theory Mapping 13

4. The Abby Persona 15

5. The Customizable Abby Persona 16

6. Abby Mentions by Role...18

7. Team Xs Form Errors..20

8. Recording Error Types 21

9. Missing Facet Example 21

10. Errors and Detours 23

11. The MultiAbby Persona..29

12. MultiAbby Footnote..29

13. Areas of Interest 32

14. BSRI Ratings 39

15. SCM Ratings...39

16. Reasons for Identifying Friends Like and Unlike Abby 41

17. SoloAbby Heatmap 45

18. MultiAbby Heatmap..45

LIST OF TABLES

Table Page

1. Organizations Using GenderMag....................12
2. Code Set Overview....................13
3. Abby Invocation Rate (Study 1)....................18
4. Error Rates by Team....................20
5. Detour Frequency....................23
6. Abby's BSRI Scores....................38
7. Abby Overall Invocation Rate (Study 2)....................41
8. Turbulence Rates....................43
9. Abby Invocation Rate by Treatment (Study 2)....................43
10. Duration of Gaze....................45
11. Statistical Results for Different Situations....................67

INTRODUCTION

Over the past decade, research spanning multiple diverse age groups and populations has shown that certain differences in the ways people use software tend to cluster by gender [5, 6, 9, 12, 13, 14, 15, 16, 18, 21, 26, 27, 41, 42, 46, 47, 49, 54, 61, 62, 65, 66, 68, 69, 72, 73, 79]. In addition, there is evidence to suggest that many software products are not designed to take these differences into account.

To help software creators address this issue, we have been working on a new method called the GenderMag method [10]. The GenderMag method (Gender Inclusiveness Magnifier) [10] is a software inspection method that aims to help software creators identify gender-inclusiveness issues in their technologies. We recently conducted a study of software teams in the field using GenderMag [11]. 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) — thus, 25% of features they evaluated had gender-inclusiveness issues [11].

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. In its first year alone, at a beta stage, GenderMag was used by more

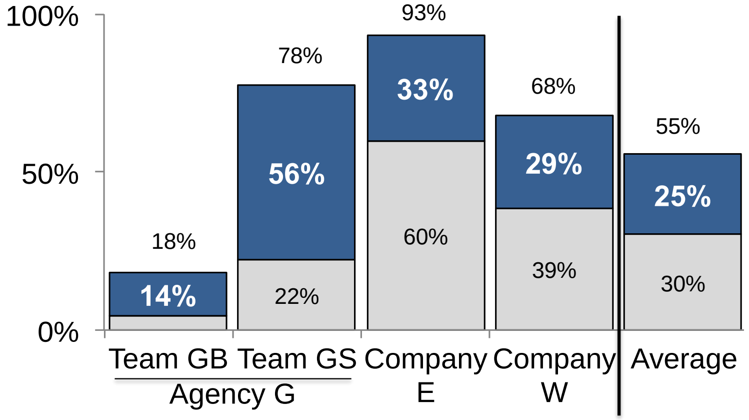


Figure 1: Issues from an early field study [6] that each team found as a percentage of the number of user actions and subgoals evaluated. Above bars: total issues. Dark blue: gender-inclusiveness issues. Light gray: other issues.

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.

The GenderMag method uses two component pieces to help software developers find gender-inclusiveness issues: cognitive walkthroughs and personas. In this thesis, we investigate both of these component pieces and how each contributes to GenderMag's overall goal of helping to identify (for eventual removal) gender biases in software. Study 1 investigates GenderMag's specialized cognitive walkthrough process, to understand the circumstances that cause GenderMag to be less (or more) effective at finding gender-inclusiveness issues. Study 2 investigates GenderMag's use of personas to evaluate whether it is possible to use GenderMag's gendered personas without promoting gender stereotyping.

STUDY 1: THE COGNITIVE WALKTHROUGH

In the field study mentioned above [11,] we observed that GenderMag's cognitive walkthrough can be cognitively taxing. Therefore, in Study 1 we consider where its 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?

STUDY 2: THE PERSONAS

GenderMag is based in part on gendered personas, which raises the possibility of unintended stereotyping. Stereotyping is an ingrained human characteristic [76], so we cannot hope to stamp it out entirely, but we can at least hope not to increase it. The type of stereotyping we focus on here is *techno*-stereotyping: if GenderMag's personas increased adverse techno-stereotyping of women, the method's components would be working against each other.

Therefore, Study 2 asks two questions: do GenderMag's personas promote adverse techno-stereotyping; and can we reduce stereotyping by introducing a diverse "cast" of personas all representing a single persona's traits–without reducing the persona's effectiveness?

Through these two studies, this thesis investigates the following overall research question: How does each component part of GenderMag contribute to (or detract from) its overall goal of helping to identify gender biases in software?

LITERATURE REVIEW

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.

GenderMag's foundations lie in research that shows people's problem-solving strategies tend to cluster by gender. GenderMag focuses on five facets of problem-solving that have been found in literature that cluster by gender. The method uses faceted personas to give life to these facets and embeds the personas' usage in a facet-focused specialization of the Cognitive Walkthrough (CW) [75, 80]. The five facets are:

Motivations: Over a decade's worth of research has found that females are more likely than males to be motivated to use technologies for the ends that they can accomplish with its help, whereas males are more often than females motivated by their interest in and enjoyment of technology itself [9, 12, 15, 41, 46, 49, 54, 72].

Information processing styles: Literature shows that females are statistically more likely to gather information comprehensively, forming a complete picture of the problem and its required background knowledge before trying to solve it. Males, on the other hand, are more likely to selectively process information, following the first piece of information that seems promising, then backtracking if the option doesn't pan out [13, 21, 61, 62, 68]. Both styles have advantages, but users of either are at a disadvantage if their style is not supported by the software they are using.

Computer self-efficacy: Empirical studies have found that females tend to have lower computer self-efficacy (area-specific confidence) than their male peers, which may affect the ways they interact with technology [5, 6, 9, 12, 27, 42, 47, 54, 65, 66, 73].

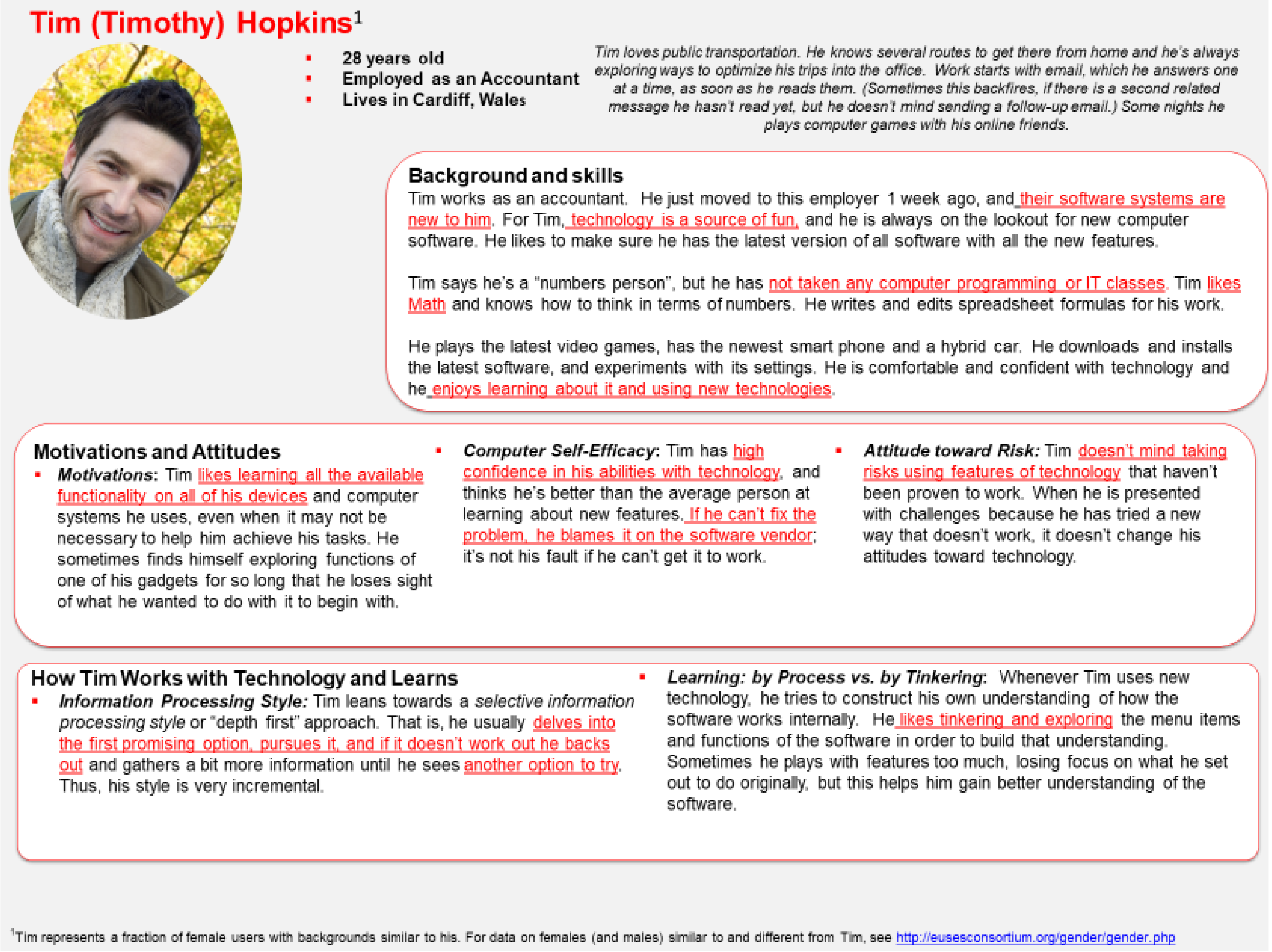
Risk: Prior research shows that females tend statistically to be more risk-averse than males when dealing with software [26], surveyed in [79], and meta-analyzed in [18].

Risk aversion may impact users' decisions regarding which features of software to use.

Tinkering: Research across many age groups and occupations reports females being statistically less likely to experiment ("tinker") with unfamiliar software features than males. If females do tinker, however, they are usually more likely to reflect on what they are doing, and thus may profit from the process more than males [6, 12, 14, 16, 46, 69].

GenderMag humanizes these facets with a set of four faceted personas—"Abby", "Pat(ricia)", "Pat(rick)" and "Tim" (see Figure 2) [10]. Each one represents a subset of a system's target users as they relate to these five problem-solving facets. To this end, Abby, Patricia, Patrick and Tim are identical in many aspects: all have the same job, live in the same town, and are equally comfortable with mathematics and with the technology that they use regularly. Their differences are strictly derived from existing gender research on the five facets. Tim's facet values represent those most frequently seen in males, while Abby's values are those seen in females that are the most different from Tim's. The two Pats' (identical) facet values represent a large portion of females' and males' problem-solving styles that are not covered by Abby's or Tim's facets. The Pats' identical facets highlight that differences relevant to inclusiveness lie not in a person's gender identity, but in the facet values themselves.

GenderMag combines the use of these personas with a specialized Cognitive Walkthrough (CW). The CW is a long-standing software inspection method used to uncover usability issues for users new to a program or feature [80]. During a GenderMag CW, evaluators step through a detailed use case (a goal and a list of actions) from the perspective of one of the personas and answer CW questions with respect to the five gendered problem-solving facets.



Tim (Timothy) Hopkins[1]

- 28 years old
- Employed as an Accountant
- Lives in Cardiff, Wales

Tim loves public transportation. He knows several routes to get there from home and he's always exploring ways to optimize his trips into the office. Work starts with email, which he answers one at a time, as soon as he reads them. (Sometimes this backfires, if there is a second related message he hasn't read yet, but he doesn't mind sending a follow-up email.) Some nights he plays computer games with his online friends.

Background and skills

Tim works as an accountant. He just moved to this employer 1 week ago, and their software systems are new to him. For Tim, technology is a source of fun, and he is always on the lookout for new computer software. He likes to make sure he has the latest version of all software with all the new features.

Tim says he's a "numbers person", but he has not taken any computer programming or IT classes. Tim likes Math and knows how to think in terms of numbers. He writes and edits spreadsheet formulas for his work.

He plays the latest video games, has the newest smart phone and a hybrid car. He downloads and installs the latest software, and experiments with its settings. He is comfortable and confident with technology and he enjoys learning about it and using new technologies.

Motivations and Attitudes

- ***Motivations***: Tim likes learning all the available functionality on all of his devices and computer systems he uses, even when it may not be necessary to help him achieve his tasks. He sometimes finds himself exploring functions of one of his gadgets for so long that he loses sight of what he wanted to do with it to begin with.
- ***Computer Self-Efficacy***: Tim has high confidence in his abilities with technology, and thinks he's better than the average person at learning about new features. If he can't fix the problem, he blames it on the software vendor; it's not his fault if he can't get it to work.
- ***Attitude toward Risk:*** Tim doesn't mind taking risks using features of technology that haven't been proven to work. When he is presented with challenges because he has tried a new way that doesn't work, it doesn't change his attitudes toward technology.

How Tim Works with Technology and Learns

- ***Information Processing Style:*** Tim leans towards a *selective information processing style* or "depth first" approach. That is, he usually delves into the first promising option, pursues it, and if it doesn't work out he backs out and gathers a bit more information until he sees another option to try. Thus, his style is very incremental.
- ***Learning: by Process vs. by Tinkering***: Whenever Tim uses new technology, he tries to construct his own understanding of how the software works internally. He likes tinkering and exploring the menu items and functions of the software in order to build that understanding. Sometimes he plays with features too much, losing focus on what he set out to do originally, but this helps him gain better understanding of the software.

[1] Tim represents a fraction of female users with backgrounds similar to his. For data on females (and males) similar to and different from Tim, see http://eusesconsortium.org/gender/gender.php

Figure 2: The Tim Persona.

RELATED WORK

Cognitive Walkthroughs

A specialized Cognitive Walkthrough (CW) forms the foundation of the GenderMag method. The most up-to-date, comprehensive study of CWs we could locate is the 2010 survey by Mahatody et al. [53]. Their survey describes many variations of the CW introduced by Lewis [52] and updated by Wharton et al. [80]. Later adaptations to the CW include such variations as having users in the CW during the process [36] or incorporating theories of cognition [28, 70]. Other modifications of the CW focus on solving problems identified with the classic CW process [71, 75].

One of the earliest responses to CW issues outside of revisions to the original method was that of Spencer's streamlined CWs [75]. Their work identified constraints of CWs that reduced the utility of the process in practice. After identifying these issues, Spencer changed the CW method in ways that attempted to fix these problems. Streamlined CWs reduced the number of questions in the CW to relieve the issues Spencer found.

More recently, Grigoreanu et al. [39] presented a CW variant called the Informal Cognitive Walkthrough. This method helps shorten the time necessary for the CW and boosts the reliability of the CW method by including representative users. However, this method relies heavily on a skilled researcher being present, limiting its usefulness in companies or groups lacking research staff.

Personas

Personas were created and developed by Cooper as a way to channel, clarify, and understand a user's goals and needs [20]. Today, personas are widely used in industry: sometimes simply to convey users' needs during software design, such as during informal role-playing tests or ideation [34, 59, 63, 67]. Recounted benefits of using personas include inducing empathy towards users [1] and facilitating communication about design choices [67]. Reasons cited for these benefits are: (a) that personas focus issues [45], (b) they provide uniform language to talk about the user and their needs

[58], (c) they reduce conflict over what the user's perceived goals are [1], and (d) they summarize data about users in a relatable and concise format [35].

However, researchers have also reported shortcomings and controversy surrounding personas. Creating an accurate, representative persona takes a significant amount of time and effort, and the persona is then too often ignored. For example, Friess reports that personas are referenced only 2% of the time in conversations regarding product decisions. [34] Friess also found that, even when evaluators use personas alongside CWs as focal points [34, 48], the personas themselves are only used 10% of the time [34].

Issues that have been reported with personas include the following: practitioners not believing personas are credible; finding personas to be abstract, misleading, or impersonal; and seeing the personas' personifying details as irrelevant [17, 59]. Furthermore, research suggests that personas are most often used by the people that created them, in part because they have firsthand knowledge of the persona's intent and formalized training on personas in general [59]. On the other hand, people who have not helped create the persona seem to prefer the raw data behind it, and are less likely to use the persona in design decisions [59]. We have also observed tensions between UX designers and software developers in which designers feel they must justify their personas' validities [55]. In addition, these findings suggest that software developers may have trouble empathizing with personas and that, for the persona to be accepted by developers, it must either be grounded in empirical work or a mainstream stereotype of a subset of users.

A persona photo or picture is part of most persona descriptions. Practitioners appreciate these images because they feel that they personalize the personas, although some worry that the photos might carry stereotypes [78]. To our knowledge, there have not been any studies tracking the actual use of persona pictures. Photos in person descriptions from other domains, such as resumes, receive a considerable amount of attention: for instance, eye tracking of LinkedIn profiles showed that recruiters spent almost one fifth of the time looking at the picture [29].

Grudin's analysis of the psychology of personas explains the importance of having a persona seem like a real person (and hence with a single appearance) [40]. As Grudin explains, personas promote engagement by leveraging a universal skill: humans' innate ability to build mental models of people by drawing from their experiences with others. The human skill of modeling people is very old, possibly dating back to humans' adoption of language, and fortunately, it transfers to an ability to build models of fictional people as well [40]. In essence, designers' ability to engage and empathize with personas comes in part from the fact that a persona seems like a person—not like a list of facts, a philosophical stance, or an educational document—but an actual person.

Perhaps not surprisingly, we have not been able to locate other research using multiple pictures on one persona. Nielsen [63] points to examples where several pictures are shown from one persona's everyday life, but analyses of personas showed they typically depict one person [64]. The only example of more than one person depicted on a persona appeared in a study of 170 personas, as part of persona descriptions that focused on a couple or on a family as the unit of reference [57]. Multiple pictures may run counter to the notion of a persona as a believable person that people want to engage with and target as representative of a user subgroup. As Adlin and Pruit put it, "Personas put a face on the user—a memorable, engaging, and actionable image" [1].

Related Work on Stereotypes

Personas rely on peoples' ability to create mental models about other people. Because of this, personas are subject to the stereotypes that the people who use the persona assign to the group of people the persona is part of. This type of stereotyping is called group stereotyping, which Grudin [40] defines as "a fixed set of characteristics assumed to be members of a shared group."

Because the GenderMag users engage with GenderMag personas in order to think about gender-inclusiveness, these personas' genders tend to be highlighted. But gender is a major source of bias in person perception, linked to prescribing certain roles and traits [8, 50]. Thus, these personas may be particularly subject to group stereotyping.

Gender is closely linked to the two basic dimensions that people rely on to judge other people: when we meet someone, we quickly make judgments of their warmth and their competence [22, 23, 32]. A prior content analysis of personas in use showed that male and female personas tended to be presented as equally competent, but tended to rely on stereotypes regarding the warmth dimension [57]. Additionally, group stereotypes typically assigned to females tend to be characteristics that are devalued [4, 31]. Another study that manipulated the gender of the GenderMag personas found that the presence of masculine problem-solving facets led people to attribute higher competence to the personas with those facets, even though each GenderMag persona was carefully designed to display equal competence to the others [56].

Because gender is so heavily stereotyped, we need to investigate GenderMag personas' relationship to gender stereotyping, especially as it relates to software and technology.

STUDY 1 METHOD

The data upon which we primarily base this investigation came from a previous field study [11]. 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.

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 (See Appendix 1 for sample Cognitive Walkthrough forms). For each user subgoal, the form asks:

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

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 [74]. 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 [34] method of measuring persona invocation, which was counting the percentage of conversational turns that invoked the personas.

	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.

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. 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.

Code Set	Method	Literature Source
Activities	Dual coding: interrater Jaccard agreement=80% over 22% of the data.	Activity Theory [74]
Detours	Dual coding: interrater Jaccard agreement=84% over 21% of the data	Activity Theory [74]
Invoking personas	Scan for persona names and pronouns referring to personas.	Persona research [34]
GenderMag CW recording errors	Dual coding: interrater Jaccard agreement=99.8% on 20% of the data.	--

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.)

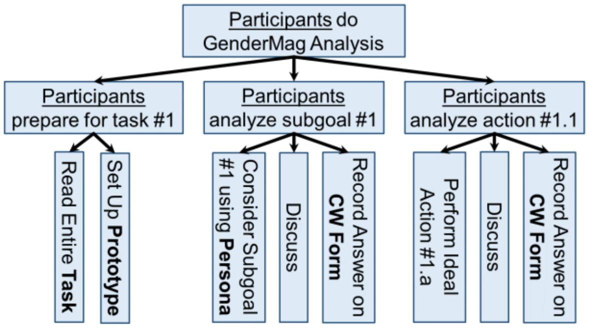

Participants do GenderMag Analysis
Participants prepare for task #1
Participants analyze subgoal #1
Participants analyze action #1.1
Read Entire Task
Set Up Prototype
Consider Subgoal #1 using Persona
Discuss
Record Answer on CW Form
Perform Ideal Action #1.a
Discuss
Record Answer on CW Form

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.

STUDY 1 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.

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 (Figure 4).

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 (Figure 5). 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].

Abby Jones

- **28 years old**
- **Employed as an Accountant**
- **Lives in Cardiff, Wales**

Abby has always liked music. When she is on her way to work in the mornings, she listens to music that spans a wide variety of styles. But when she arrives at work, she turns it off, and begins her day scanning all her emails first to get an overall picture before answering any of them. (This extra pass takes time but seems worth it.) Some nights she exercises or stretches, and sometimes she likes to play computer puzzle games like Sudoku.

Background and skills

Abby works as an accountant. She is comfortable with the technologies she uses regularly, but she just moved to this employer 1 week ago, and their software systems are new to her.

Abby says she's a "numbers person", but she has never taken any computer programming or IT systems classes. She likes Math and knows how to think with numbers. She writes and edits spreadsheet formulas in her work.

In her free time, she also enjoys working with numbers and logic. She especially likes working out puzzles and puzzle games, either on paper or on the computer.

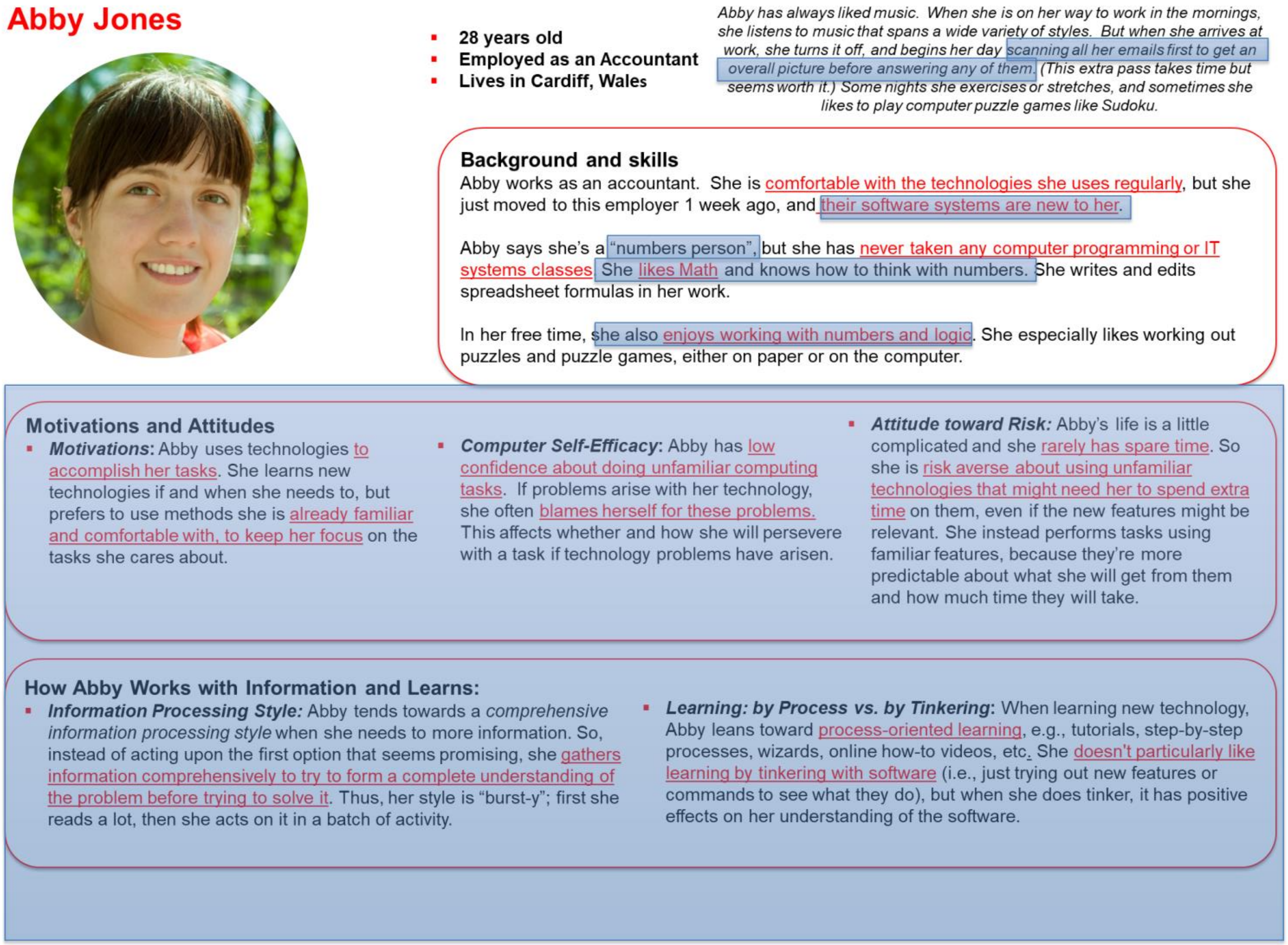
Motivations and Attitudes

- ***Motivations***: Abby uses technologies to accomplish her tasks. She learns new technologies if and when she needs to, but prefers to use methods she is already familiar and comfortable with, to keep her focus on the tasks she cares about.
- ***Computer Self-Efficacy***: Abby has low confidence about doing unfamiliar computing tasks. If problems arise with her technology, she often blames herself for these problems. This affects whether and how she will persevere with a task if technology problems have arisen.
- ***Attitude toward Risk:*** Abby's life is a little complicated and she rarely has spare time. So she is risk averse about using unfamiliar technologies that might need her to spend extra time on them, even if the new features might be relevant. She instead performs tasks using familiar features, because they're more predictable about what she will get from them and how much time they will take.

How Abby Works with Information and Learns:

- ***Information Processing Style:*** Abby tends towards a *comprehensive information processing style* when she needs to more information. So, instead of acting upon the first option that seems promising, she gathers information comprehensively to try to form a complete understanding of the problem before trying to solve it. Thus, her style is "burst-y"; first she reads a lot, then she acts on it in a batch of activity.
- ***Learning: by Process vs. by Tinkering***: 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 (i.e., just trying out new features or commands to see what they do), but when she does tinker, it has positive effects on her understanding of the software.

Figure 4: The Abby persona



Abby Jones

- 28 years old
- Employed as an Accountant
- Lives in Cardiff, Wales

Abby has always liked music. When she is on her way to work in the mornings, she listens to music that spans a wide variety of styles. But when she arrives at work, she turns it off, and begins her day scanning all her emails first to get an overall picture before answering any of them. (This extra pass takes time but seems worth it.) Some nights she exercises or stretches, and sometimes she likes to play computer puzzle games like Sudoku.

Background and skills

Abby works as an accountant. She is comfortable with the technologies she uses regularly, but she just moved to this employer 1 week ago, and their software systems are new to her.

Abby says she's a "numbers person", but she has never taken any computer programming or IT systems classes. She likes Math and knows how to think with numbers. She writes and edits spreadsheet formulas in her work.

In her free time, she also enjoys working with numbers and logic. She especially likes working out puzzles and puzzle games, either on paper or on the computer.

Motivations and Attitudes

- ***Motivations***: Abby uses technologies to accomplish her tasks. She learns new technologies if and when she needs to, but prefers to use methods she is already familiar and comfortable with, to keep her focus on the tasks she cares about.
- ***Computer Self-Efficacy***: Abby has low confidence about doing unfamiliar computing tasks. If problems arise with her technology, she often blames herself for these problems. This affects whether and how she will persevere with a task if technology problems have arisen.
- ***Attitude toward Risk:*** Abby's life is a little complicated and she rarely has spare time. So she is risk averse about using unfamiliar technologies that might need her to spend extra time on them, even if the new features might be relevant. She instead performs tasks using familiar features, because they're more predictable about what she will get from them and how much time they will take.

How Abby Works with Information and Learns:

- ***Information Processing Style:*** Abby tends towards a *comprehensive information processing style* when she needs to more information. So, instead of acting upon the first option that seems promising, she gathers information comprehensively to try to form a complete understanding of the problem before trying to solve it. Thus, her style is "burst-y"; first she reads a lot, then she acts on it in a batch of activity.
- ***Learning: by Process vs. by Tinkering***: 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 (i.e., just trying out new features or commands to see what they do), but when she does tinker, it has positive effects on her understanding of the software.

Figure 5: The text within the shaded boxes is not customizable.

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 [34]. Using the same method reported by Friess, we calculated the number of persona invocations per conversational turn. 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 at least 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.

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 6 (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 [34, 55].

	Turns that invoked personas	Turns that did not invoke personas	Total turns
Prior work [34]	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.

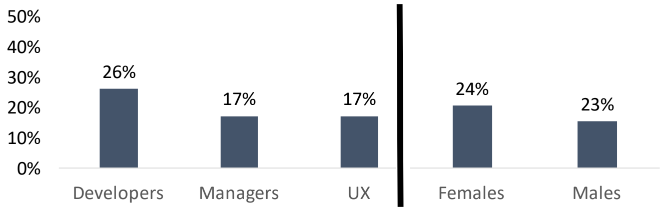


Figure 6: (Left) Managers and the UX intern referred to Abby in 17% of their conversational turns, but developers referred to her the most (26%). (Right): Females referred to Abby about the same amount as males did.

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.

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 very much 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., they 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 7 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 8, which breaks down the errors into five types. For example, neglecting to record facets (the most common error, illustrated in Figure 9) 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.

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.

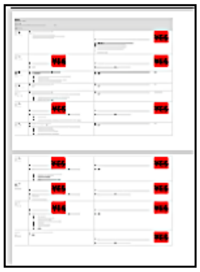


Figure 7: This form, from a Team Xs GenderMag session, had 12 recording errors (red blobs) out of 21 answers (i.e., 57%). The recorder originally pasted down through the form, then often forgot to update the Yes/No/Maybe's.

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.

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 [75] 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 10). Further, only 4% of segments without detours contained

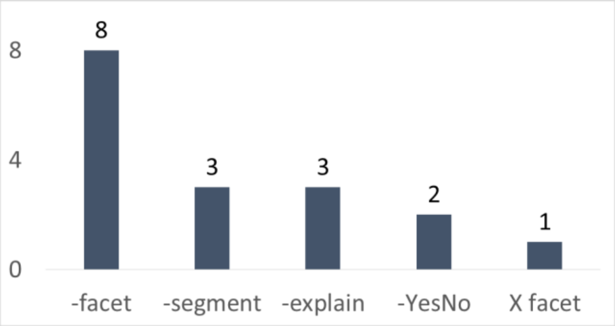
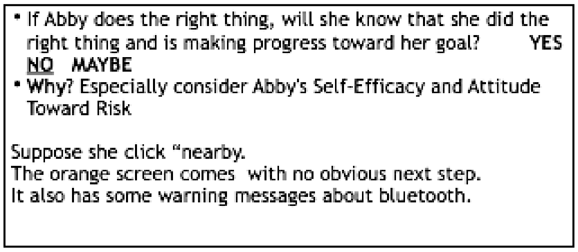


Figure 8: The five types of recording errors: "-" means "missing", and "X" means "wrong". The most common was missing facets ("-facets") with 8 instances.



- If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? YES [✓ NO] MAYBE
- Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk

Suppose she click "nearby.
The orange screen comes with no obvious next step.
It also has some warning messages about bluetooth.

Figure 9: 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.

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.

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.

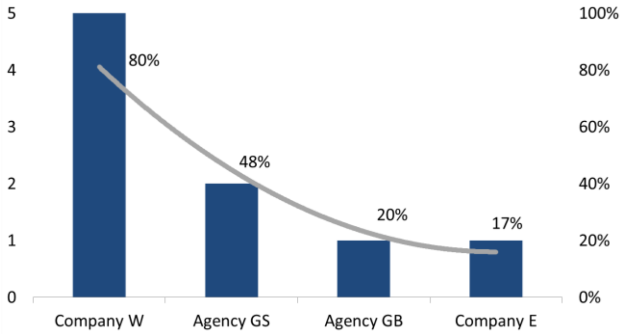


Figure 10: The bars are the number of CW form errors, and the line is the percentage of CW segments during which teams had detours. (Overall, 36% of segments contained at least one detour.)

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 the need to take a detour.

STUDY 1 DISCUSSION

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.

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.

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.)

STEREOTYPING VERSUS PROMOTING DIVERSITY

One issue that we continue to monitor is stereotyping. As the related work points out (e.g., [55]), 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.

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.

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.

With this in mind, we designed a new type of persona: the new Abby (as well as Tim and the Pats) includes photos of 3 more "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. However, we needed to make sure that our modifications to Abby didn't detract from the positive results we've discussed here. In Study #2, we investigate the effects of this manipulation.

STUDY 2 METHOD

In Study 2, we present a potential solution to the problem of possible gender stereotyping of personas: including multiple pictures of different people, males and females, on a single persona profile. Including multiple, diverse pictures on personas may reduce stereotyping of the persona's software usage habits. However, we recognize that changing an important aspect of the persona may have unintended consequences. Our goal is to investigate whether or not including multiple pictures on a persona reduces participants' stereotype activation without impacting their use of the GenderMag method. To evaluate our manipulation's effects, we conducted a controlled lab study, triangulating our results with eye tracking. We structure our investigation around two research questions:

- RQ1: How do people gender-stereotype personas in the context of gender-inclusiveness?
- RQ2: Can we reduce stereotyping by introducing a diverse "cast" of personas all representing a single persona's traits, and does that negatively affect engagement, learning, or turbulence?

THE MANIPULATION

Our manipulation consisted of presenting the GenderMag Abby persona with either a single picture or four different pictures. Figure 4 shows the full Abby persona with a single picture. Figure 11 shows the four-picture manipulation. All text is the same on both personas, with the exception of the footnote included on the single-picture persona (Figure 12). We specifically designed Abby for use with the GenderMag method, and she (with the single picture) has been employed by various companies that used GenderMag. For the four-picture treatment, we added three pictures to the persona to show that a persona with these problem-solving facets could possess socio-demographic attributes different from the young, white, female Abby on the original persona. Since Abby focuses on facets that have been shown to affect women more than men [10], the manipulation depicted more women than men. We added a footnote to the manipulated persona explaining that Abby represents users with motivations/attitudes and information/learning styles similar to hers and offered a link to find further information. For brevity, we refer to the manipulated version of Abby as multiAbby, and the non-manipulated version as soloAbby.

GenderMag has four different personas, but we used only one of them (Abby), for validity and feasibility. Specifically, GenderMag sessions always use only one persona, so validity required use of only one at a time. However, doing multiple sessions for different personas would have at least doubled the number of participants required, which was not feasible. Since stereotypes around technology usage are unfavorable to females [10], we prioritized our investigation on stereotyping of females.

To answer our research questions, we ran two studies. Study 2.1, conducted at Heilbronn University in Germany, used eye tracking to analyze participants' gaze on different parts of the persona description sheet. Study 2.2, based at Oregon State University, examined the effect of the manipulation in other situations - both with use in actual GenderMag sessions (referred to as the *GenderMag* situation) as well as in

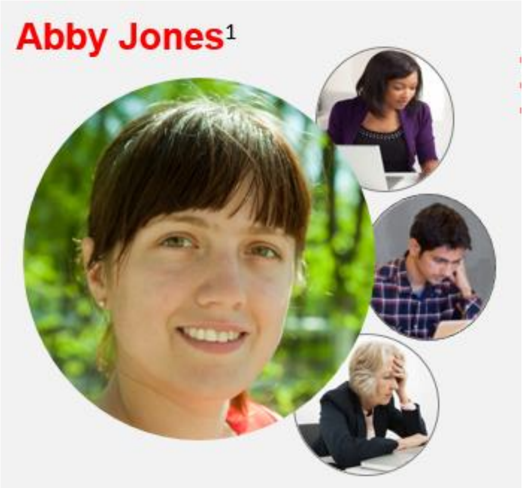


Figure 11: The pictures on the persona profiles. SoloAbby participants viewed personas with the large picture (left), and multiAbby participants viewed all four as shown here.

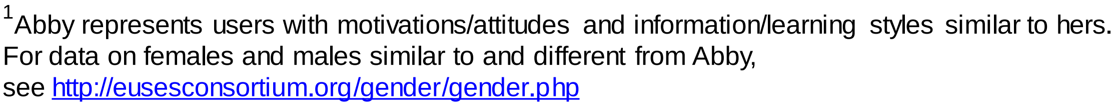

[1] Abby represents users with motivations/attitudes and information/learning styles similar to hers. For data on females and males similar to and different from Abby, see http://eusesconsortium.org/gender/gender.php

Figure 12: The multiAbby persona included this footnote to explain the multiple pictures.

sessions where participants only viewed the persona (referred to as the *PersonaOnly* situation). Including participants from both Heilbronn University and Oregon State University allowed us to collect data from two different cultures, as well as providing access to eye-tracking data. Both Study 2 groups gave their impressions on Abby and her problem-solving facets.

STUDY 2.1(EYE TRACKING) METHODOLOGY

Procedure

Participants of Study 2.1 were a convenience sample of professionals in the field of software development, research, and management. They were not compensated for their participation. 14 professionals (5 females, 9 males) participated in the eye-tracking study, filled out the questionnaire (Section Study 2.2 Methodology), and were debriefed. We instructed all participants in the same way at the beginning of the experiment about the usage of the devices and the procedure. We then presented Abby on a screen. In a between-participants design with two levels, we randomly assigned participants to view Abby with one or four pictures (soloAbby vs. multiAbby). Seven participants saw soloAbby, seven participants saw multiAbby. We collected eye-tracking data using a Tobii X60 Eye Tracker with preliminary data analysis in Tobii Studio, and performed further analysis in SPSS Statistics 22. We placed the eye tracker approximately 70 cm distance from the participant’s eyes, and the vertical angle that the screen made from the participant’s view was less than 35°.

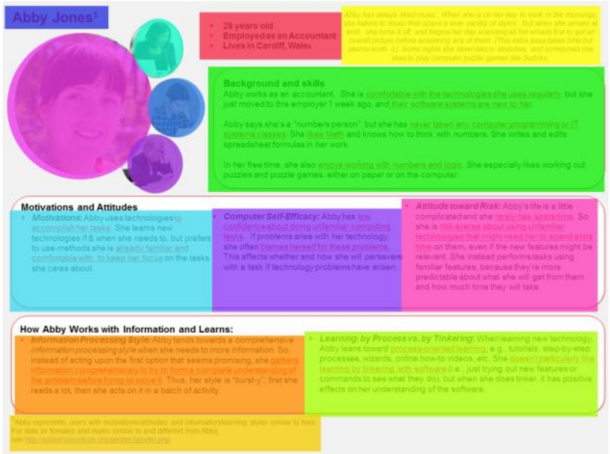
Data analysis

To analyze participants’ gaze on the different parts of the persona description and pictures, we defined areas of interest (AOI). Ten AOIs were defined for the soloAbby condition (name, picture Abby, age/employment etc., abstract, background and skills, and the five facets, i.e., motivations, computer self-efficacy, attitude towards risk, information processing style, tinkering). For the multiAbby condition, we defined 14 AOIs: three for the additional pictures and one for the footnote that was included to explain the usage of the four pictures (see Figure 13). The other AOIs were the same as in the soloAbby condition. As the dependent variable, we used the time spent inspecting the AOI, i.e., the duration of the visit in seconds. We measured the duration spent on each facet, accounting for length of facet description by measuring the duration of gaze per word in each facet. We also looked at the sequence in which the participants fixated the AOIs. Additionally, we used the number of fixations to create a visual overview ("heat map") of the gaze's dynamic.

STUDY 2.2 METHODOLOGY

Participants in Study 2.2 were mostly students at Oregon State University, though we did not limit participation to only students. We recruited participants by emailing announcements and distributing and posting flyers around campus and the surrounding areas. All participants were over 18 years old. The final participant count for Study 2.2 was 36 females and 36 males, spanning many age groups, academic majors and statuses.

We divided participants into two different situations: one where participants used the GenderMag method, and one where participants only looked at the persona. In each situation, the independent variable was the picture manipulation (one vs. four pictures). We randomly assigned participants to treatment sessions.



Abby Jones[1]

- 28 years old
- Employed as an Accountant
- Lives in Cardiff, Wales

Abby has always liked music. When she is on her way to work in the mornings, she listens to music that spans a wide variety of styles. But when she arrives at work, she turns it off, and begins her day scanning all her emails first to get an overall picture before answering any of them. (This extra pass takes time but seems worth it.) Some nights she exercises or stretches, and sometimes she likes to play computer puzzle games like Sudoku.

Background and skills

Abby works as an accountant. She is comfortable with the technologies she uses regularly, but she just moved to this employer 1 week ago, and their software systems are new to her.

Abby says she's a "numbers person", but she has never taken any computer programming or IT systems classes. She likes Math and knows how to think with numbers. She writes and edits spreadsheet formulas in her work.

In her free time, she also enjoys working with numbers and logic. She especially likes working out puzzles and puzzle games, either on paper or on the computer.

Motivations and Attitudes

- *Motivations:* Abby uses technologies to accomplish her tasks. She learns new technologies if & when she needs to, but prefers to use methods she is already familiar and comfortable with, to keep her focus on the tasks she cares about.
- *Computer Self-Efficacy:* Abby has low confidence about doing unfamiliar computing tasks. If problems arise with her technology, she often blames herself for these problems. This affects whether and how she will persevere with a task if technology problems have arisen.
- *Attitude toward Risk:* Abby's life is a little complicated and she rarely has spare time. So she is risk averse about using unfamiliar technologies that might need her to spend extra time on them, even if the new features might be relevant. She instead performs tasks using familiar features, because they're more predictable about what she will get from them and how much time they will take.

How Abby Works with Information and Learns:

- *Information Processing Style:* Abby tends towards a *comprehensive information processing style* when she needs to more information. So, instead of acting upon the first option that seems promising, she gathers information comprehensively to try to form a complete understanding of the problem before trying to solve it. Thus, her style is "burst-y"; first she reads a lot, then she acts on it in a batch of activity.
- *Learning: by Process vs. by Tinkering:* 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 (i.e., just trying out new features or commands to see what they do), but when she does tinker, it has positive effects on her understanding of the software.

[1] Abby represents users with motivations/attitudes and information/learning styles similar to hers. For data on females and males similar to and different from Abby, see http://eusesconsortium.org/gender/gender.php

Figure 13: Areas of interest (AOIs) on multiAbby. SoloAbby's AOIs were equivalently adapted to the gaze without the AOIs covering the three extra pictures and the footnote.

GenderMag Sessions

Participants in the GenderMag situation performed a GenderMag session among themselves. Each GenderMag session included 2 to 4 participants, all of whom were new to GenderMag. In Study 1 and previous GenderMag work [11, 44], team sizes ranged from 3 to 10 people. Larger group sizes do sometimes impact the use of the method, but group sizes were unlikely to impact the stereotyping of the persona in this study since each participant had their own copy of the persona and was asked to internalize the persona.

We gave participants a brief introduction to the method before they began, during which we asked them to “get to know” Abby by reading her persona. During the walkthrough, participants evaluated a feature of a popular word processing software using GenderMag from Abby’s perspective (See Appendix 1: Cognitive Walkthrough Forms).

We observed the roughly 1-hour sessions, and we also video-recorded (or audio-recorded, if the participants did not consent to video-recording) and later transcribed each session. After performing the GenderMag walkthrough, we temporarily removed Abby, and gave the participants the first questionnaire (described in Data Analysis). Following the questionnaire, we returned Abby to them and asked them to fill out a second questionnaire (also described in Data Analysis).

PersonaOnly Sessions

For the PersonaOnly situation, we presented participants with the Abby persona (either soloAbby or multiAbby) similarly to the GenderMag situation, but without introducing the rest of the GenderMag method. Participants in these groups read the persona silently and could ask the researcher for clarification if they had questions about Abby or her problem-solving traits. After participants read the persona, we gave them the same questionnaires as the GenderMag session groups.

Data Analysis

In both studies, participants filled out a questionnaire after they had viewed the persona or participated in the GenderMag session. We measured the effects of the manipulation regarding the following dependent variables:

- gender stereotyping
- facet perception
- gendering of search scope
- engagement with the persona
- confusion regarding the persona

The operationalization of the dependent variables is described in the following paragraphs. The study employed quantitative and qualitative measures, both in the questionnaire that was used and in the analysis of the GenderMag sessions that were recorded and transcribed.

Gender Stereotyping

To measure gender stereotyping, we measured to what extent participants applied traditional feminine attributes to Abby. This dependent variable was operationalized with two instruments: a short version [77] of the Bem Sex-Role Inventory BSRI [7], and warmth/competence questionnaire of the stereotype content model SCM [33], for a total of 17 items.

To elicit the extent to which the participants were stereotyping, we compared our participants' perceptions of Abby (measured by scores of the BSRI and the SCM) with established literature regarding peoples' stereotyping of other real people. found in representative samples: For the BSRI, the test values were based on Donnelly and Twenge [25], who found women's feminine role at M = 5.0 (on a 7 point likert scale), women's masculine role at M= 4.9, men's feminine role at M = 4.6, men's masculine role at M = 5.1. For the SCM we used Asbrock's [2] means of women's warmth at M = 5.6 and competence at M = 4.2, and men's warmth at M = 4.2 and competence at M = 5.6 (transformed from a five to a seven point Likert scale by proportional transformation [19]). This comparison allowed us to see whether or not Abby was subject to gender stereotyping at the same level as real people.

Gender stereotyping was submitted to analyses of variance (ANOVA) with the manipulation of the pictures (soloAbby vs. multiAbby) and the use in a GenderMag session (with GenderMag vs. PersonaOnly) as between-participant factors.

Facet Perception

We define facet perception as the participants' ability to correctly associate facet descriptions with the facets on the Abby Persona. We operationalized facet perception by applying the facet attributes of the GenderMag persona description to the persona [56]. We represented each facet was represented by two items on Questionnaire 1. For example, motivation (task orientation) was measured with "spends money on technology because new technology is fun or cool" (reverse code) and "spends time or money on technology mainly to accomplish some work or task goal". For the full questionnaire, see Appendix 1. The results for facet perception were submitted to ANOVA with the manipulation of the pictures (one vs. four pictures) and the use in a GenderMag session (with GenderMag vs. PersonaOnly) as between-participant factors.

For both the 17 gender-stereotyping items and the 10 facet-perception items, participants gave their impression of Abby by expressing to what extent the attributes applied to her. Agreement was measured on a seven-point Likert scale ranging from "not at all" to "extremely". The order of the items was randomized for each participant.

To determine whether a facet was recalled correctly, we determined acceptable Likert scale answers based on the descriptions of the facets in the persona: correct participant responses reflected the descriptions of the facets in the persona. For instance, the description of Abby's attitude towards risk contains the phrase *"Abby is risk averse when she uses computers to perform tasks."* Therefore, the *"tries to avoid risk"* item should be rated as greater than 4 (i.e., on the Likert scale: moderately, very, or extremely) to be considered correct.

Additionally, we measured facet perception through eye tracking, quantifying the duration of gaze on each facet.

Gendering of Search Scope

The second questionnaire consisted of qualitative questions meant to measure the extent to which the manipulation influenced the search scope of participants. We posed open questions about how they relate to the persona, which attributes of the persona the participant did or did not identify with, and we asked participants to name a few friends who were or weren't like Abby, and why (see Appendix 1). In line with previous research that linked automatic stereotyping with free recall [43], we used the answers to these questions to measure whether the persona description limited participants to thinking mainly of females. The dependent variable *"gendering of search scope"* was operationalized based on the assessment of Abby's likeness to male vs. female friends: For each friend mentioned by the participants, we asked the participants to identify the friend's gender. The ratio of female and male friends was calculated for the friends that were named and transformed into percentages, separately for the friends that were like Abby and the friends that were unlike Abby. We used T-tests to determine whether the manipulation of soloAbby vs. multiAbby had an effect of gendering or de-gendering the search scope.

Engagement with Persona

To identify indicators of participant engagement with Abby, we analyzed the transcripts of the GenderMag sessions. We operationalize *"engagement with the persona"* similarly to past literature [34]: we measured the invocation of Abby in relation to the number of conversational turns. We split the transcripts by conversational turn (as has been done in Study 1 and past literature[34]). We then counted the number of times the persona was invoked during each conversational turn. To be conservative with our measurements, we didn't count invocations if the participant was reading a question from the CW forms (e.g., "Would Abby know what to do at this step?").

We also measured engagement with the persona through eye tracking, considering visual engagement of the areas of interest (AOIs) by measuring gaze duration. We measured visual engagement by recording the fixation time on areas of interest (AOIs). To account for the different number of AOIs in the different treatments, we measured absolute fixation time per AOI rather than as a percentage of overall time.

Confusion regarding the Persona

To measure confusion, we identified instances in the cognitive walkthrough where participants may have faced confusion because of our manipulation. As a first conservative step, we identified instances of confusion when a participant said something that expressly stated they were confused by the persona. This includes confusion about Abby's gender (e.g., using pronouns), asking the researcher for clarification about the multiple pictures, and so forth. We call these "explicit turbulence". Since stereotype activation is typically measured with implicit measures [37, 43], we then expanded the code set to include instances of "implicit turbulence": if a participant stated something that implied confusion about the persona, rather than stating it outright. Examples of this include statements that signified participants were unsure about part of the persona ("she wouldn't do that...right?" in the manner of "I'm not quite sure about this"), making contradictory statements ("she's not a tinkery type but she's going to press everything"), and struggling to define or explain a concept ("she isn't tinkering she is just…she's just pressing stuff").

To come to an operationalization of "confusion regarding the persona", we analyzed transcripts through content analysis [60]. With the focus of identifying turbulences that might be caused by the manipulation, we searched the transcripts of the GenderMag session for statements indicating that the participants were confused about Abby. We looked for participants explicitly talking about being confused about the persona, but also for implicit cues of turbulences. In an iterative process, categories were formed and the data was structured.

Two researchers qualitatively coded each transcript of GenderMag session groups to get the turbulence count. We coded on conversational turns, i.e., each time the speaker in the transcript changed. After coding 12.75% of the data, we performed inter-rater reliability (IRR) analysis to assess the degree to which coders consistently assigned implicit turbulences to statements made by the participants. We used Cohen's Kappa to measure IRR, and obtained κ = .86, indicating substantial agreement [51]. The researchers then coded the remainder of the data set individually.

Results across Study 2.1 (Germany) and Study 2.2 (US) did not show any significant differences for any measured dependent variable after applying the Bonferroni correction (see Appendix 2 for details). The GenderMag and PersonaOnly situations' results were not significantly different either (see Appendix 2 for details). Therefore, in the results sections we report aggregated results with N = 86:

- soloAbby, PersonaOnly, n = 24 (7 Germany, 17 US);
- multiAbby, PersonaOnly, n = 23 (7 Germany, 16 US);
- soloAbby, with GenderMag, n = 18 (US);
- multiAbby, with GenderMag, n = 21 (US).

This multi-site study thus afforded generalization across two countries, and allowed triangulation of eye tracking results with questionnaire responses and session transcripts.

STUDY 2 RESULTS

RQ1: HOW DO PEOPLE GENDER-STEREOTYPE PERSONAS

Stereotyping

To measure stereotyping of the persona, we asked participants to rate the Abby persona on traditional feminine and masculine attributes in Bem’s Sex Role Inventory (BSRI), and asked them to evaluate her warmth and competence. Participants rated Abby’s BSRI-masculine score (M = 3.6, SD = 0.80) lower than her BSRI-feminine score (M = 4.5, SD = 0.61; $p < .001$). Participants rated Abby’s competence (M = 4.4, SD = 0.84) lower than her warmth (M = 4.9, SD = 0.77; p =.000). We compared these results with established literature regarding peoples’ stereotyping of other real people. We analyzed participants’ BSRI scores in a one sample t-test with the reference scores, and found that participants rated Abby’s BSRI masculine score lower than women’s or men’s masculine scores ($p < .01$). There was a tendential difference between our participants’ Abby BSRI-feminine scores and the reference score (p = .060) – i.e., participants rated Abby’s feminine traits and masculine traits lower than the BSRI women’s feminine and women’s masculine scores used as a reference (see Table 6 and Figure 14).

We performed a similar comparison for the SCM (warmth and competence). For the SCM, we used a reference score of women’s competence at M = 4.2 and warmth at M = 5.6, and men’s competence at M = 5.6 and warmth at M = 4.2 [2]. The one sample t-test yielded significant differences between our participants’ Abby scores for: men’s competence ($p < .01$), women’s warmth ($p < .01$), men’s warmth (p=.01), and

		Abby’s mean score	**Reference score [2, 25]**
Masculine		3.6 (SD 0.80)	Women’s: 4.9
			Men’s: 5.1
Feminine		4.5 (SD 0.61)	Women’s: 5.0
			Men’s: 4.6
Warmth		4.9 (SD 0.77)	Women’s: 5.6
			Men’s: 4.2
Competence		4.4 (SD 0.84)	Women’s: 4.2
			Men’s: 5.6

Table 6: Abby’s BSRI scores compared to reference scores (N=86).

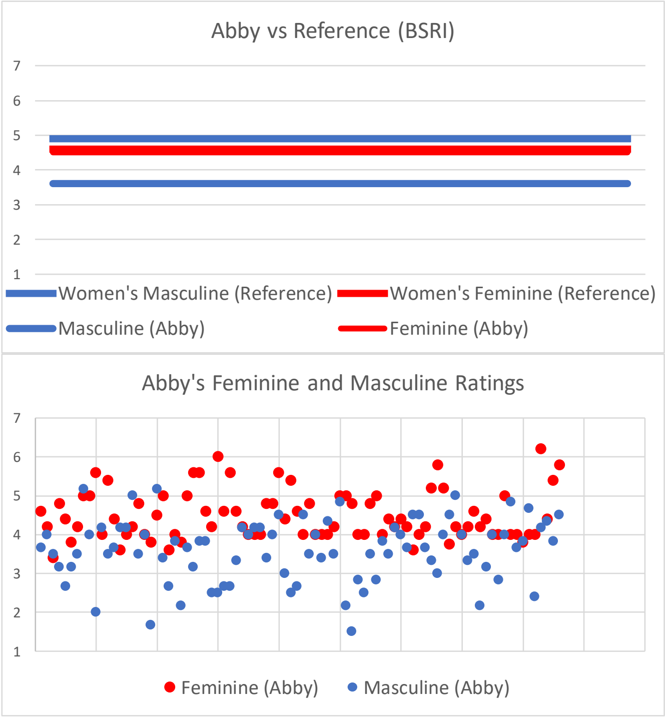


Figure 14: Participants rated Abby less masculine and less feminine than participants in the reference literature [2, 25] rated women (top) and as more feminine than masculine (bottom)

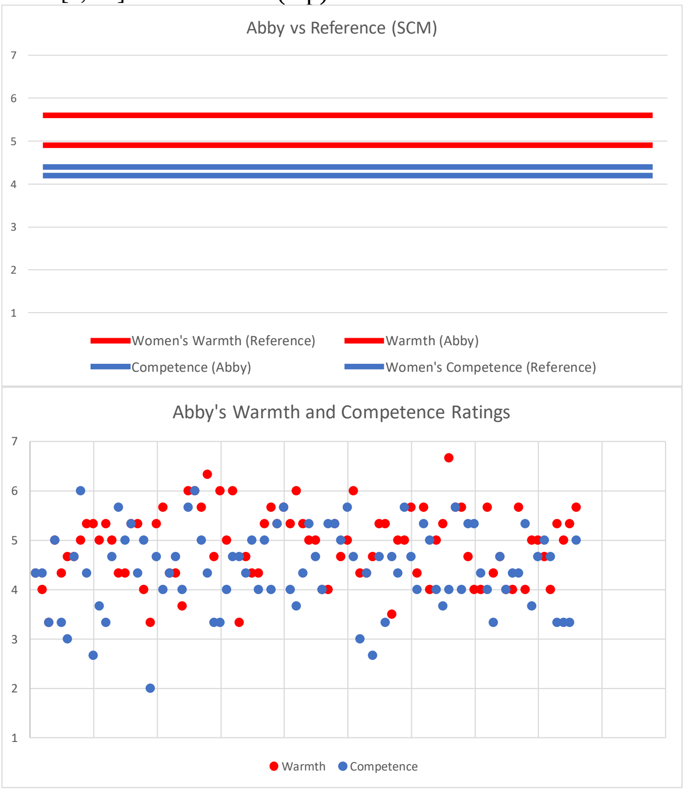


Figure 15: Participants rated Abby in between the reference scores [2, 25] for women's warmth and competence (top), and rated her equally warm and competent (bottom)

women’s competence (p = .018). That is, the warmth and competence our participants attributed to Abby were significantly different and *between* the average warmth and competence typically attributed to men and women (see Figure 15).

Thus, the results of the BSRI and SCM measurements suggest that our participants applied stereotypes to Abby less than people in the reference literature apply stereotypes to real people.

Gendered search scope

As a way of measuring whether participants felt that Abby only represented either men or women, we asked participants to identify friends like or unlike Abby, and their friends’ genders. We then compared the percentage of female and male friends that were named to be similar or unlike the GenderMag persona. Overall, participants named 38% male friends as like Abby, 62% female friends as like Abby (SD = 0.38); 31% female friends unlike Abby, and 69% male friends unlike Abby (SD = 0.40). This suggests that participants did not feel that Abby only represented men or women, but that she could represent either.

To delve more deeply into whether participants identified friends as like/unlike Abby for stereotypical reasons, we coded the reasons they identified friends as like or unlike. Abby. We classified their reasons into four types: background (e.g., “likes to go to the gym”), logic (e.g., “is a numbers person”), and facet (e.g., “doesn’t like tinkering”), or any other reasons that didn’t fit into the first three categories. Two researchers reached 92% interrater reliability over 20% of the data, and then one researcher finished the coding.

Figure 16 shows the distribution of reasons that participants identified friends as like or unlike Abby. If participants saw Abby as a stereotypical female, we would expect to see participants categorizing friends along stereotypical lines (e.g., female friends being like Abby due to facets but unlike Abby due to other reasons, and males being unlike Abby due to her facets but like Abby due to other reasons). However, note that most friends, regardless of gender, were like or unlike Abby based on her facets. This suggests that participants realized that Abby’s facets could represent either males or females, and also that not all women were like Abby.

Engagement

As mentioned in Literature Review, creating a persona takes a significant amount of time and effort, and the persona is often ignored. In prior field work on personas by Friess [34], personas were used between 2% and 10% of conversational turns. Thus, we measured engagement by how often participants invoked Abby during discussion in the GenderMag sessions. Study 1 showed invocation rates of up to 23%. We used the same metric to measure engagement with the persona: we counted the conversational turns between participants during which the participants invoked the Abby persona. Our participants invoked Abby during 34% of conversational turns (Table 7).

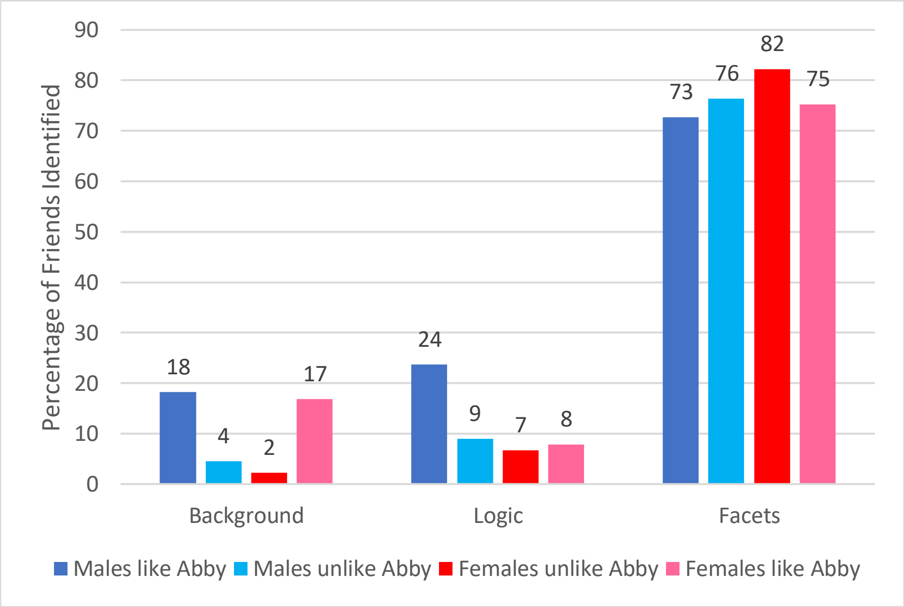


Figure 16: Participants’ reasons for identifying friends like or unlike Abby. The y-axis is the percentage of friends identified as like or unlike for the reasons on the x-axis. Note that multiple reasons could be applied to a single friend.

	Turns that invoked personas	**Turns that did not invoke personas**	**Total turns**
Friess [34]	94 (10%)	997 (90%)	1091
GenderMag field study [44]	601 (23%)	2006 (77%)	2607
Current work	736 (34%)	1429 (66%)	2165

Table 7: Invocations of Abby in this study vs. other work.

RQ2: CAN WE REDUCE STEREOTYPING?

Stereotyping

BSRI and SCM

We conducted two-way ANOVAs (N = 86) to examine the effect of our picture manipulation and the use of GenderMag sessions on gender stereotyping, using the results of participants' responses to the dependent variable gender stereotyping. Neither the statistical main effects nor the interaction effect yielded significant results: the groups did not show significant differences regarding their responses in the BSRI or the warmth/competence questionnaire of the stereotype content model (SCM).

Gendered search scope

The t-test comparing responses from multiAbby participants and soloAbby participants showed that multiAbby participants identified friends unlike Abby at M = 37% (SD = .45), whereas soloAbby participants named female friends unlike Abby at M = 26% (SD = .33, p = .001), i.e., multiAbby participants identified significantly more female friends unlike Abby. There was no significant difference between responses for friends like Abby.

Facet perception

To measure participants' facet perception, we asked the participants to express to what extent the facet attributes applied to the persona. We conducted two-way ANOVAs to examine the effect of our manipulation and GenderMag sessions on facet recollection by the participants. No significant differences could be found regarding multiAbby vs. soloAbby, GenderMag, or the statistical interaction between the two factors.

To determine whether all participants recalled a facet correctly, we compared the results to the Likert scale answers that would correctly reflect the facets in the persona. Both groups generally recalled facets correctly, except for one information processing style item ("selective in dealing with information" was M = 5, SD =1.55).

Turbulence

We identified implicit and explicit turbulences from the GenderMag transcripts. Recall that we define explicit turbulence as a participant expressly stating they were confused by the persona, and implicit turbulence as a participant implying they were confused by the persona without stating it outright. We identified a total of four instances of explicit turbulences in the transcripts; we found 216 instances of implicit turbulences across both groups (see Table 8 for more details). We performed Fisher's exact test to determine whether either treatment experienced more turbulence. No significant differences could be found for either implicit or explicit turbulence.

Engagement By Conversational Turn

We performed Fisher's exact test to determine whether either treatment referred to Abby more often. No significant differences could be found regarding multiAbby vs. soloAbby. MultiAbby participants invoked Abby during 35.95% of their turns, while soloAbby participants invoked Abby in 28.38% of their turns (Table 9).

Eye Tracking

The analysis of the eye-tracking data shows which parts of the persona the participants engaged with visually. All participants read through the text from top to bottom and looked at the AOI covering the footnote only at the end, and not immediately after viewing the pictures. With respect to the overall time that participants spent looking at the persona description, the manipulation did not show any significant differences: the average duration of the visits of all the areas of interest (AOIs) for N = 14 was M = 140.36 seconds and SD = 25.42 for multiAbby, and M = 145.35 seconds (SD = 35.63) for soloAbby. The aggregated gaze distribution over all multiAbby participants is shown in Figure 17, and for soloAbby in Figure 18.

The results show differences in the duration that participants look at background/skills and at the name of the persona: In the multiAbby condition participants looked at the

Treatment	**Implicit turbulence instances**	**Explicit turbulence instances**
soloAbby	6.75	0
multiAbby	4.26	0.21

Table 8: Mean instances of turbulence per participant.

Treatment	**# participants**	**Total turns**	**Turns w/ >=1 invocation**
soloAbby	18	1231	392 (28%)
multiAbby	21	934	344 (36%)

Table 9: How often participants in each group invoked Abby.

background/skills longer (M = 32.72 seconds, SD = 3.15) than in the soloAbby condition (M = 30.10 seconds, SD = 9.27, p = .010). Also, multiAbby participants tended to look at the name longer (M = 0.80 seconds, SD = 0.50) than soloAbby participants (M = 0.25 seconds, SD = 0.22, p = .056).

Regarding the percentage of time spent on the facets, there was a difference between the conditions: Taken the AOIs for the five facets together, the analysis showed that participants spent a higher percentage of the time looking at the facets with soloAbby (M = 39%, SD = 0.01) than with multiAbby (M = 36%, SD = 0.02, p = .029).

To account for different lengths of facet descriptions, we measured gaze duration per word as a measure of the time spent on each facet. A significant difference between soloAbby and multiAbby could not be found. The average durations that the participants spent on each facet can be seen in Table 10. The average time per word spent on the facets was higher for computer self-efficacy compared to all other facets (p < .05). The comparison of the other facets did not yield significant results.

The average time soloAbby participants spent on the picture (M = 2.18 seconds, SD = 3.35) did not differ significantly from the time multiAbby participants looked at the four pictures altogether (M = 1.72 seconds, SD = 2.10). In fact, participants in both groups looked at the picture for less than 2% of their total time. Male and female participants did not differ significantly (women: M = 2.92 seconds, SD = 3.82; men: M = 1.41 seconds, SD = 1.90). MultiAbby participants spent a similar amount of time reading the footnote (M = 2.17 seconds, SD = 2.77). Of these participants, one person did not look at the footnote at all. The sequence of the eye fixations showed that the other participants (n = 6) looked at the footnote after looking at all the other AOIs.

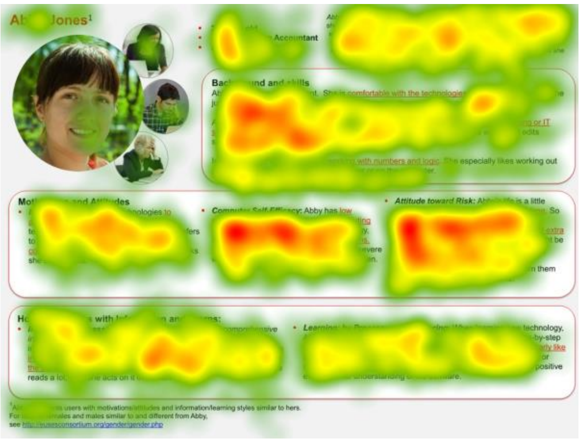


Figure 17: Eye fixations, aggregated over all participant in multiAbby eye-tracking condition (n=7; darkest red = 30.16 counts)

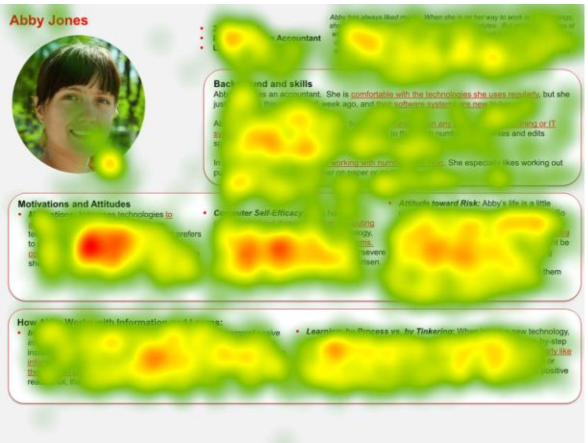


Figure 18: The soloAbby heat map

Areas of Interest (AOI)	**Total seconds spent on AOI (SD)**	**Seconds per word (SD)**
Motivations	13.33 (3.77)	.32 (.09)
Computer self-efficacy	15.22 (3.82)	.37 (.09)*
Risk	21.46 (5.48)	.32 (.08)
Info. processing style	20.00 (6.55)	.29 (.10)
Learning style	16.88 (5.72)	.27 (.09)

Table 10: Mean durations of gaze (seconds). *=different from all other facets (p < .05)

STUDY 2 DISCUSSION

For GenderMag or any persona-based software inspection method to succeed, participants must *engage* with the persona; that is, they must utilize the persona and refer to the persona in their discussion. To help participants engage with the persona, they are typically designed as a person: background, picture, and facets. To this end, personas must be *believable*. GenderMag carries an additional directive: to *educate* participants about the problem-solving strategies of various genders so those strategies can be accounted for in software design. GenderMag strives to encourage engagement, believability, and education without promoting gender-based *stereotypes*. We now consider each goal.

STEREOTYPING

Forming person perceptions and stereotypes is an automatic process, and this problem extends to personas. Through the use of two questionnaires, we measured stereotyping by the BSRI and SCM metrics, as well as the genders of "friends like or unlike Abby".

Our analysis of the application of gender stereotypes to Abby (both soloAbby and multiAbby) revealed an interesting insight: the perception of the persona – regardless of whether it had one or four pictures – was not subject to gender stereotypes as strongly as real people are. This held true for participants in both countries (US and Germany), and was seen in both the BSRI and SCM that we used as measures of gender stereotyping.

In particular, the BSRI has four categories: masculine, feminine, androgynous, and undifferentiated. Our result shows that Abby would be classified as undifferentiated in the BSRI. This may explain why soloAbby vs. multiAbby did not yield different results: the attribution of feminine or masculine traits was rather low altogether. The results of the SCM show a similar picture: participants neither perceived Abby as a "typical woman" nor a "typical man" with regard to warmth and competence. This result suggests that our participants' attitude towards Abby was not gender-biased – but the participants also did not admire her or see her as part of their in-group [33].

In fact, there may not have been enough gender stereotyping going on with soloAbby to further reduce it – at least not by changing the persona's picture. When asked to identify friends like or unlike Abby, participants identified both male and female friends as like Abby, based primarily on her facets. Participants seemed to grasp the idea that Abby could represent a range of people, regardless of whether Abby had one picture or four – and this generalized across countries (US vs. Germany), gender (male vs. female), and experience (professionals vs. students).

BELIEVABILITY

We performed content analysis to determine whether participants were confused by having multiple pictures on the persona. We coded the transcripts as described in the Methodology section. No significant difference was found between groups for either implicit or explicit turbulence.

However, only multiAbby participants experienced instances of explicit turbulence. Explicit turbulence instances were rare though – only 4 instances appeared in the transcripts (as opposed to 216 instances of implicit turbulence), and these might have occurred because of confusion in the use of the "correct" pronouns in the persona description.

EDUCATION

To help software teams identify features in their software that are not gender-inclusive, GenderMag has to educate the teams on the facets as part of their use of the method. So far, GenderMag has been effective at teaching teams about the facets of the personas [55], but we needed to make sure that adding extra pictures to Abby didn't negatively affect participants' learning of facets.

To measure this, we compared groups' responses on the Questionnaire 1. There were no significant differences between participant groups; participants in all groups tended to recall facets correctly, with the exception of an Information Processing Style item. The item that the participants did not recall in line with our expectations was *"selective in dealing with information"*. The information processing style facet is designed to refer to a cognitive approach of information processing. Because the opposite item, *"processes information comprehensively"*, was answered correctly, we suspect the first item may have been misleading: for instance, it could be interpreted as this excerpt from the Abby persona: *"focused on the tasks she cares about"*.

ENGAGEMENT

There was a danger that adding pictures to Abby could negatively affect engagement: multiple pictures might not allow evaluators to empathize as easily and cause them to avoid invoking the persona. However, our results showed that the manipulation did not harm engagement with the persona; participants in both soloAbby GenderMag and multiAbby GenderMag groups referred to Abby equally as often. In related work, persona engagement has often been a problem [17, 34, 48, 59].

To ward off such problems, we took several measures in designing the GenderMag personas. For example, we explained that there was extensive data behind the personas, and to make them quickly digestible we made them fit on one page and used bullets, boldface, and red, underlined text. These measures appear to have paid off, because according to established metrics [34], our participants over all GenderMag sessions engaged with the personas much more than in previous non-GenderMag studies of personas (34% vs. 10% of conversational turns). This result also outperforms Study 1 (with a prior version of soloAbby) in which GenderMag users engaged with the personas in 23% of conversational turns [44].

However, adding the extra pictures to form multiAbby may have slightly altered *how* participants engaged with Abby: it changed how they distributed their attention within the persona. SoloAbby participants spent more time reading the facets than multiAbby participants. On the other hand, multiAbby participants spent more time reading the name and the background/skills. However, the effect size was very small; the difference amounts were only about 2 seconds out of 140-145 seconds total.

FOUR PICTURES OR ONE?

We discovered something interesting about the attention paid to the persona's pictures: participants didn't look at the persona pictures for long, regardless of treatment. Participants spent 2.18 seconds in soloAbby and 1.72 seconds in multiAbby treatments looking at the pictures – a small fraction of the two minutes they spent on the entire persona. Two seconds might be considered average for looking at a picture of a face on a web site [24] – but it is a very short time to spend looking at pictures of four different people, or trying to get to know a person from a picture. We speculate that instead of taking in or thinking about picture(s) on the persona, multiAbby participants simply glanced at the pictures, or only looked at them long enough to register the pictures' presence, but not long enough to examine them.

There are no published eye-tracking studies of personas that allow for a comparison between the duration of visual attention given to the persona picture(s) vs. the textual description. Therefore, we do not have a direct basis to which we can compare our findings. However, some literature addresses the use of non-fictitious person profiles where a person's photo complements a description of the person. For instance, job recruiters spend roughly 20% of their time on a profile studying the profile picture [29]. This underscores the brevity of our participants' gazes on the persona's picture(s), which accounted for less than 2% of the time spent looking at the persona. Future work should investigate differences in attention given to pictures in fictitious vs non-fictitious person profiles, as well as the influence that the viewer's task (e.g., recruiting vs. GenderMag) has on the attention given to pictures.

Why did participants spend so little time looking at the picture? Perhaps they were aware that the picture on the persona is just an illustration not conveying any "real" information about a person – an arbitrary picture of a person, meant to illustrate and underline the persona description. Thus, illustrating a persona description with four instead of one picture accentuates the message that Abby could be any age, any ethnicity, and/or any gender.

CONCLUSION

The goal of this thesis was to investigate how each component part of GenderMag contributes to (or detracts from) its overall goal of helping to identify gender biases in software.

STUDY 1 CLOSING THOUGHTS

Study 1 investigated the *process* component of GenderMag, namely GenderMag's specialized cognitive walkthrough. The experiences of four software teams who used GenderMag to evaluate the inclusiveness of their own software revealed insights into situations in which GenderMag was less (or more) effective in finding gender-inclusiveness issues. Among our results were:

- *Persona(s) engagement during the process*: 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 at the same time 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 being 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 results of Study 1 suggest that the primary source of errors in GenderMag walkthroughs are detours, which we believe are caused by a high cognitive load on participants. On the other hand, participants were much more engaged with GenderMag's personas than prior literature suggests.

STUDY 2 CLOSING THOUGHTS

In Study 2, we investigated GenderMag's use of personas (the second component of GenderMag) to evaluate whether it is possible to use GenderMag's gendered personas without promoting inappropriate gender stereotyping. We presented groups of participants with one of two personas: one with a single portrait, and the other with multiple portraits of people with different ages, ethnicities, and genders. We then measured participants' perceptions of this modified persona, focusing especially on gender stereotyping.

This study provided evidence to suggest that people's perceptions of personas are perhaps not as straightforward as they seem:

- *Stereotyping:* Although Abby, a gendered persona, represents a range of problem-solving facets that disproportionately affect women, participants in all conditions and in both countries viewed Abby as neither stereotypically feminine nor masculine. This suggests that neither soloAbby nor multiAbby triggered adverse techno-stereotyping of women.

- *Engagement:* We had feared that multiAbby might reduce engagement, but we found no differences between solo- and multiAbby groups in the amounts participants engaged with the persona, either verbally or visually. Further, the addition of multiple pictures did not seem to harm engagement with the persona, the learning of facets, or overall believability.

- *Pictures:* Participants looked at Abby's picture - whether solo or multi - for less than 2% of the time they spent looking at the persona description. We

> expected the picture to receive a much larger portion of participants' attention. This suggests that the participants realized that the persona's appearance was not an important aspect of the persona.

The key takeaway is that, although participants did not stereotype Abby as either traditionally masculine or feminine, they engaged with her more than most other personas in the literature, and they understood her problem-solving strategies, even when Abby was represented by four pictures. Thus, it appears that we can have it both ways—avoiding the promotion of inappropriate stereotypes while illustrating a persona's gender-inclusiveness using a diverse group of people.

THE BIGGER PICTURE

What has been done

Study 1 uncovered some unexpected problems and strengths with the GenderMag CW. When thinking about the scope and impact of these problems, we must consider that the CW forms are the only record a team typically has of their decisions and thoughts about the GenderMag process, and in turn, whether or not a feature had a Gender-Inclusiveness issue. Thus, issues that affect the CW forms affect the outcome and effectiveness of the GenderMag method.

Even though participants had a high (93%) success rate at capturing team discussion on the CW forms, they missed up to a third of gender-inclusiveness issues in their software. We looked deeper in to why those issues were being missed, and found that a few different types of detours were dominating the list of mistakes that participants made: troubleshooting, procedure, where are we, and researcher clarification detours made up 77% of the detours present.

Of these detours, "where are we" was perhaps the most indicative of underlying issues in GenderMag's cognitive walkthrough forms. The GenderMag forms may have been confusing to participants, and may not have been specific enough to guide the participants through the walkthrough on their own. We have begun to attempt to address these problems through the use of simplified CW forms.

Study 2 investigates the other component of the GenderMag walkthroughs: the persona. Study 1 found that participants were surprisingly engaged with the persona they were using. This was promising, but we wondered why participants were so engaged with GenderMag's personas. Ideally, it would be because the participants perceived the persona as real and representative of some subset of their users.

However, we were concerned that this feeling of representativeness might encourage to inappropriate stereotyping (e.g., "all women are like Abby, and only women are like Abby"). If participants were relying on stereotypes to find gender-inclusiveness issues, it would undermine the message of GenderMag and no matter how much we improve the CWs, the GenderMag method wouldn't be a positive influence. Thus, we needed to make sure that we measured whether participants adversely stereotyped Abby.

Because we felt having one gender represented on the personas could encourage adverse stereotyping, we measured whether a persona that included people from different genders and socio-demographics would increase or decrease stereotyping. However, we also wanted to make sure that including more than one picture didn't reduce the believability or engagement/empathy with the persona – without these, the persona wouldn't be useful in the method.

Study 2 suggests that Abby would be classified as undifferentiated in the BSRI, regardless of whether there were one or four pictures on her persona. Furthermore, participants seemed to grasp the idea that Abby didn't represent all women, and that she didn't only represent women.

Implications Beyond GenderMag

The studies presented in this thesis enumerate some of the strengths and weaknesses of the component parts of the GenderMag method. As we continue to gain a better understanding of what helps GenderMag work and what hinders its aim of helping developers find gender-inclusiveness issues in their software, we update the method.

These updates aim to make GenderMag more useful to and successful for its participants. For instance, GenderMag's CW forms were sometimes confusing to teams. Thus, we are currently exploring ways to reduce the forms' complexities without compromising their effectiveness. Finding and recording the difficulties that GenderMag teams have experienced will help us help future teams avoid those difficulties. Additionally, the work presented in this thesis provides information to guide developers of inclusiveness-based methodologies to preempt some of the issues we've found.

The work done on GenderMag through the years has also provided insight into how future inclusiveness methodologies can be constructed using GenderMag as a framework. GenderMag's workflow consists of cognitive walkthroughs and research-based personas. These component parts were adapted from existing methodologies and combined to help identify gender-inclusiveness issues in software. We expect that these methodologies can be combined in a similar fashion for other inclusiveness issues, e.g., customizing a cognitive walkthrough and personas people of low socio-economic status using software to gain access to social services.

Thus, the findings presented in this thesis have an impact broader than gender-inclusiveness: we can use the lessons learned here as a guide to creating new inclusiveness methodologies.

BIBLIOGRAPHY

[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] Frank Asbrock. 2010. Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology* 41, 2: 76–81.

[3] John A. Bargh. 2013. *Social psychology and the unconscious: The automaticity of higher mental processes.* Psychology Press.

[4] Manuela Barreto, Naomi Ellemers. 2015. Detecting and Experiencing Prejudice: New Answers to Old Questions. *Advances in Experimental Social Psychology* 52: 139-219.

[5] Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. 2005. Effectiveness of end-user debugging software features: Are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '05), 869-878. http://doi.acm.org/10.1145/1054972.1055094

[6] 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* (CHI '06), 231-240. http://doi.acm.org/10.1145/1124772.1124808

[7] Sandra L. Bem. 1981. *Bem Sex-Role Inventory: Professional Manual*. Consulting Psychologists Press, Palo Alto, CA.

[8] Marilynn B. Brewer. 1988. A dual process model of impression formation. In *Advances in Social Cognition* (Vol. 1), Thomas K. Srull and Robert S. Wyer (eds.). Psychology Press, New York, 1-36.

[9] Margaret 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: 450–460.

[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*, online January 2016. doi:10.1093/iwc/iwv046.

[11] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender inclusiveness software issues with GenderMag: A field investigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '16), 2586-2598. http://doi.acm.org/10.1145/2858036.2858036.2858274

[12] Margaret Burnett, Scott D. Fleming, Shamsi Iqbal, Gina Venolia, Viyda 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* (ESEM '10), 28. http://doi.acm.org/10.1145/1852786.1852824

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

[14] Jill Cao, Kyle Rector, Thomas H. Park, Scott D. Fleming, Margaret Burnett, and Susan Wiedenbeck. 2010. A debugging perspective on end-user mashup programming. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing* (IEEE '10), 149-156.

[15] J. Cassell. 2002. Genderizing HCI. In *The Handbook of Human-Computer Interaction,* M.G. Helander, T.K. Landauer, and P.V. Prabhu (eds.). L. Erlbaum Associates Inc., Hillsdale, NJ, 402-411.

[16] Shou Chang, Vikas Kumar, Eric Gilbert, and Loren 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 (CSCW '14), 674-686. http://doi.acm.org/10.1145/2531602.2531660

[17] Christopher N. Chapman and Russell Milham. 2006. The personas' new clothes: methodological and practical arguments against a popular method. *Human Factors and Ergonomics Society Annual Meeting* 50: 634-637.

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

[19] Andrew Colman, Claire Norris, and Carolyn Preston. 1997. Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports* 80: 355-362.

[20] Alan Cooper. 2004. *The Inmates Are Running the Asylum*. Sams Publishing.

[21] Constantinos K. Coursaris, Sarah J. Swierenga, and Ethan Watrall. 2008. An empirical investigation of color temperature and gender effects on web aesthetics. *Journal of Usability Studies* 3, 3: 103-117.

[22] Amy Cuddy, Susan Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology* 40: 61-149.

[23] Amy Cuddy, Susan Fiske, Virginia Kwan, Peter Glick, Stephanie Demoulin, Jacques-Philippe Leyens, et al. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology* 48, 1: 1-33.

[24] Soussan Djamasbi, Marisa Siegel, and Tom S. Tullis. 2012. Faces and viewing behavior: An exploratory investigation. *AIS Transactions on Human-Computer Interaction* 4, 3: 190-211.

[25] Kristin Donnelly and Jean M. Twenge. 2016. Masculine and feminine traits on the Bem sex-role inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*: 1-10. http://dx.doi.org/10.1007/s11199-016-0625-y.

[26] 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: 522–550.

[27] Alan Durndell and Zsolt Haag. 2002. 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* 18, 5: 521–535.

[28] Joel U. Eden. 2007. Distributed cognitive walkthrough (DCW): a walkthrough-style usability evaluation method based on theories of distributed cognition. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition* (C&C '07), 283-283 http://doi.acm.org/10.1145/1254960.1255019

[29] Will Evans. 2012. Eye-tracking online metacognition: cognitive complexity and recruiter decisionmaking. Retrieved September 9, 2015 from http://cdn.theladders.net/static/images/basicSite/pdfs/TheLadders-EyeTracking-StudyC2.pdf

[30] Susan Fiske. 2000. Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology* 30, 3: 299-322.

[31] Susan Fiske. 2015. Intergroup biases: a focus on stereotype content. *Current Opinion in Behavioral Sciences* 3: 45-50.

[32] Susan Fiske, Amy Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences,* 11, 2: 77-83.

[33] Susan Fiske, Amy Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82, 6: 878-902.

[34] Erin Friess. 2012. Personas and decision making in the design process: an ethnographic case study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1209-1218.

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

[36] Toni Granollers and J. Lorés. 2004. Incorporation of users in the evaluation of usability by cognitive walkthrough. In *HCI Related Papers of Interacción*. 243-255.

[37] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1: 4.

[38] Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97, 1: 17-41.

[39] Valentina Grigoreanu and Manal Mohanna. 2013. Informal cognitive walkthroughs (ICW): paring down and pairing up for an agile world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 3093-3096. http://doi.acm.org/10.1145/2470654.2466421

[40] Jonathan Grudin. 2006. Why personas work: The psychological evidence. In *The Persona LifeCycle: Keeping People in Mind Throughout Product Design*, John Pruitt and Tamara Adlin (aut). Morgan Kaufmann Publishers.

[41] 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: 137-149.

[42] Kathleen Hartzel. 2003. How self-efficacy and gender issues affect software adoption and use. *Communications of the ACM – Why CS students need math* 46, 9: 167–171.

[43] E. Tory Higgins. 1996. Knowledge activation: Accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles*, Guilford Press, New York, 133-168.

[44] 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 the IEEE Symposium on Visual Languages and Human-centric Computing* (IEEE '16).

[45] Karen Holtzblatt, Jessamyn B. Wendell, and Shelley Wood. 2004. *Rapid Contextual design: A How-to Guide to Key Techniques for User-Centered Design*. Morgan Kaufmann, San Francisco, CA.

[46] Weimin Hou, Manpreet Kaur, Anita Komlodi, Wayne G. Lutters, Lee Boot, Shelia R. Cotten, Claudia Morrell, A. Ant Ozok, and Zeynep Tufekci. 2006. Girls don't waste time: Pre-adolescent attitudes toward ICT. In *Proceedings of the CHI EA Conference on Human Factors in Computing Systems* (CHI '06), 875-880. http://doi.acm/org/10.1145/1125451.1125622

[47] Ann H. Huffman, Jason Whetten, William H. Huffman. 2013. Using technology in higher education: The influence of gender roles on technology self-efficacy. *Computers in Human Behavior* 29, 4: 1779–1786.

[48] Tejinder K. Judge, Tara Matthews, and Steve Whittaker. 2012. Comparing collaboration and individual personas for the design and evaluation of collaboration software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1997-2000. http://doi.acm.org/10.1145/2207676.2208344

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

[50] Mary E. Kite, Kay Deaux, and Elizabeth L. Haines. 2008. Gender stereotypes. In *Psychology of women: A handbook of issues and theories* (2nd ed.), Florence L. Denmark and Michele A. Paludi

(eds.). Greenwood Publishing Group, 205-236.

[51] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1: 159-174.

[52] Clayton Lewis, Peter G. Polson, Cathleen Wharton, and John Rieman. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90), 235-242. http://doi.acm.org/10.1145/97243.97279

[53] Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. 2010. State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human-Computer Interaction* 26, 8: 741-85.

[54] Jane Margolis and Allan Fisher. 2003. *Unlocking the Clubhouse: Women in Computing*. MIT Press.

[55] Nicola Marsden and Maren Haag. 2016. Stereotypes and politics: reflections on personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 4017-4031. http://doi.acm.org/10.1145/2858036.2858151

[56] Nicola Marsden and Maren Haag. 2016. Evaluation of gendermag personas based on persona attributes and persona gender. In *HCI International 2016 – Posters' Extended Abstracts: 18th International Conference, HCI International 2016, Toronto, Canada, July 17-22, 2016, Proceedings, Part I*, Constantine Stephanidis (Ed.). Cham: Springer International Publishing, 122-127.

[57] Nicola Marsden, Jasmin Link, and Elisabeth Büllesfeld. 2015. Geschlechterstereotype in Persona-Beschreibungen. In *Mensch und Computer 2015 Tagungsband*, Sarah Diefenbach, Niels Henze and Martin Pielot (eds.), Stuttgart: Oldenbourg Wissenschaftsverlag, 113-122.

[58] Adrienne L. Massanari. 2010. Designing for imaginary friends: information architecture, personas, and the politics of user-centered design. *New Media & Society* 12, 3: 401-416.

[59] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1219-1228. http://doi.acm.org/10.1145/2207676.2208573

[60] Philipp Mayring. 2014. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Beltz, Klagenfurg.

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

[62] Joan Meyers-Levy and Durairaj Maheswaran. 1991. Exploring differences in males' and females' processing strategies. *Journal of Consumer Research* 18, 1: 63–70.

[63] Lene Nielsen and Kira S. Hansen. 2014. Personas is applicable: a study on the use of personas in Denmark. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 1665-1674. http:/doi.acm.org/10.1145/2556288.2557080

[64] Lene Nielsen, Kira S. Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *International Journal of Sociotechnology and Knowledge Development* 7, 1: 45-61.

[65] Anne O'Leary-Kelly, Bill Hardgrave, Vicki McKinney, and Darryl Wilson. 2004. The influence of professional identification on the retention of women and racial minorities in the IT workforce. In *NSF Info. Tech. Workforce & Info. Tech. Res. PI Conf.*: 65-69.

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

[67] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003*

conference on Designing for user experiences (DUX ‘03), 1-15. http:/doi.acm.org/10.1145/997078.997089

[68] René Riedl, Marco Hubert, and Peter Kenning. 2010. Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of EBay offers. *MIS Quarterly* 34, 2: 397-428.

[69] Daniela Rosner and Jonathan Bean. 2009. Learning from IKEA hacking: I’m not one to decoupage a tabletop and call it a day. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ‘09), 419-422. http://doi.acm.org/10.1145/1518701.1518768

[70] Hokyoung Ryu and Andrew F. Monk. 2004. Analysing interaction problems with cyclic interaction theory: Low-level interaction walkthrough. *Psychology Journal* 2, 3: 304-330.

[71] Andrew Sears. 1997. Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction* 9, 3: 213-234.

[72] Steven J. Simon. 2001. The impact of culture and gender on web sites: an empirical study. *The Data Base for Advances in Information Systems* 32, 1: 18-37.

[73] 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: 739–746.

[74] Mads Soegaard and Rikke Friis Dam, (eds.). 2014. *The Encyclopedia of Human-Computer Interaction*, 2nd ed, The Interaction Design Foundation.

[75] 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* (CHI ‘00), 353-359. http://doi.acm/org/10.1145/332040.332456

[76] Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? *Design Studies* 32, 1: 30-44.

[77] Afshin Vafaei, Beatriz Alvarado, Concepcion Tomás, Carmen Muro, Beatriz Martinez, and Maria Victoria Zunzunegui. 2014. The validity of the 12-item Bem Sex Role Inventory in older Spanish population: An examination of the androgyny model. *Archives of gerontology and geriatrics* 59, 2: 257-263.

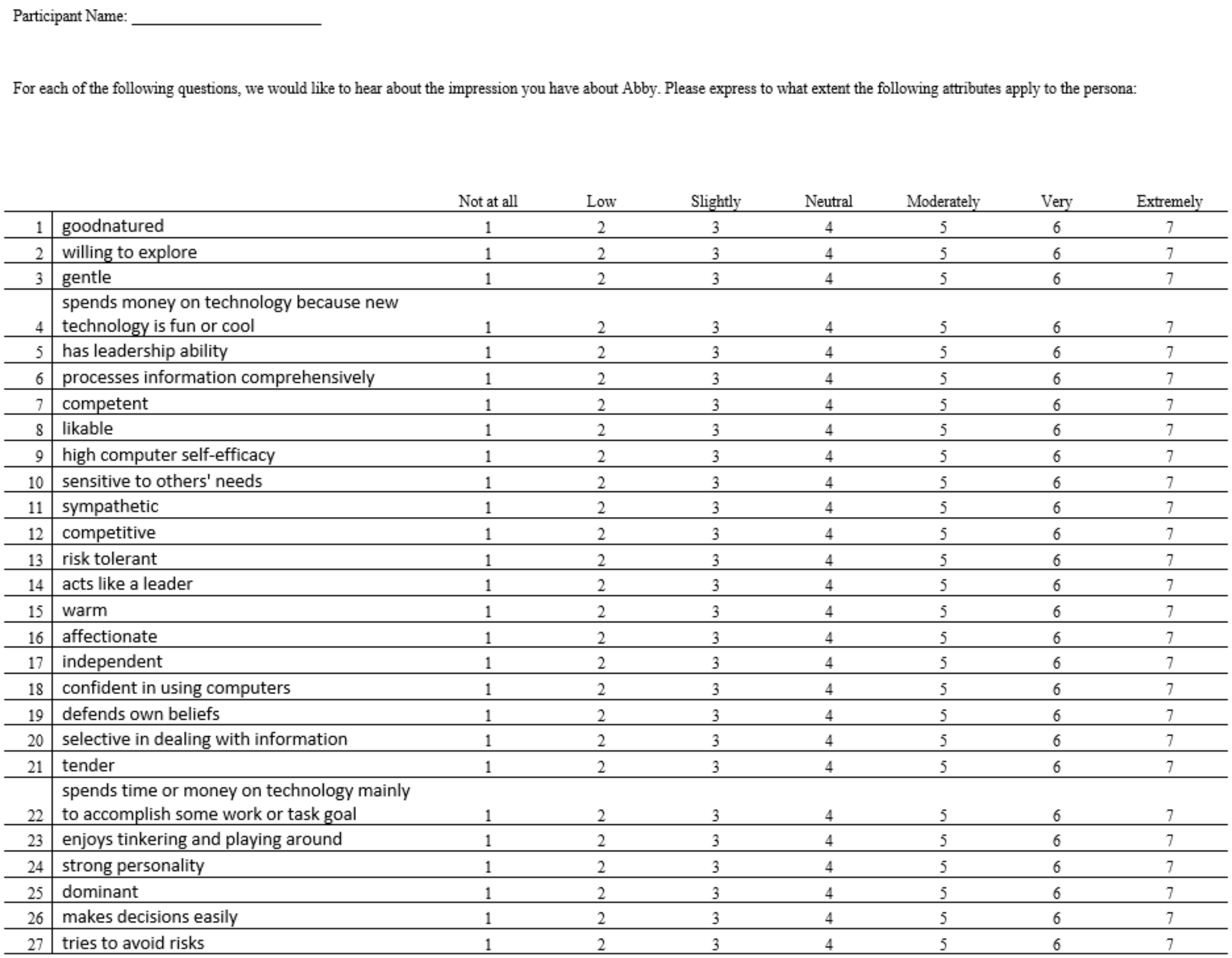
[78] Gabriela Viana and Jean-Marc Robert. 2016. The practitioners’ points of view on the creation and use of personas for user interface design. In *Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I*, Masaaki Kurosu (Ed.). Cham: Springer International Publishing, 233-244.

[79] Elke U. Weber, Ann-Renée Blais, and Nancy E. Betz. 2002. A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making* 15, 4: 263-290.

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

APPENDIX 1: STUDY 2 MATERIALS

QUESTIONNAIRE 1



Participant Name: ______________________

For each of the following questions, we would like to hear about the impression you have about Abby. Please express to what extent the following attributes apply to the persona:

		Not at all	Low	Slightly	Neutral	Moderately	Very	Extremely
1	goodnatured	1	2	3	4	5	6	7
2	willing to explore	1	2	3	4	5	6	7
3	gentle	1	2	3	4	5	6	7
4	spends money on technology because new technology is fun or cool	1	2	3	4	5	6	7
5	has leadership ability	1	2	3	4	5	6	7
6	processes information comprehensively	1	2	3	4	5	6	7
7	competent	1	2	3	4	5	6	7
8	likable	1	2	3	4	5	6	7
9	high computer self-efficacy	1	2	3	4	5	6	7
10	sensitive to others' needs	1	2	3	4	5	6	7
11	sympathetic	1	2	3	4	5	6	7
12	competitive	1	2	3	4	5	6	7
13	risk tolerant	1	2	3	4	5	6	7
14	acts like a leader	1	2	3	4	5	6	7
15	warm	1	2	3	4	5	6	7
16	affectionate	1	2	3	4	5	6	7
17	independent	1	2	3	4	5	6	7
18	confident in using computers	1	2	3	4	5	6	7
19	defends own beliefs	1	2	3	4	5	6	7
20	selective in dealing with information	1	2	3	4	5	6	7
21	tender	1	2	3	4	5	6	7
22	spends time or money on technology mainly to accomplish some work or task goal	1	2	3	4	5	6	7
23	enjoys tinkering and playing around	1	2	3	4	5	6	7
24	strong personality	1	2	3	4	5	6	7
25	dominant	1	2	3	4	5	6	7
26	makes decisions easily	1	2	3	4	5	6	7
27	tries to avoid risks	1	2	3	4	5	6	7

QUESTIONNAIRE 2:

Participant Name:_______________________

The persona you looked at had certain attributes that affect their way of dealing with technology. From your experience, are there other aspects influencing the use of technology that we might have missed? Which ones are they? How might they affect a person's use of technology?

With which attributes of the persona do you most associate yourself? If so, which ones?

- ____________________ Explanation:
- ____________________
- ____________________
- ____________________
- ____________________

With which attributes of the persona do you least associate yourself. If so, which ones?

- ____________________ Explanation:
- ____________________
- ____________________
- ____________________
- ____________________

Name a few friends who remind you of the persona that you just viewed/ worked with. For each friend you chose, please tell us why you chose this person.

Name a few friends who seem very different from the persona you just viewed/worked with. For each friend you chose, please tell us why you chose this person.

Cognitive Walkthrough Forms

Scenario name: Abby needs to insert a line of 3pt thickness into her document

- **Subgoal #: 1**
- **Subgoal name:** Insert a line

Analysis questions:

- Will Abby have formed this sub-goal as a step to her overall goal? **YES NO MAYBE** (Choose one)
- **Why**? Especially consider Abby's Motivations & Strategies

• Action #: 1a • Name: Click the “insert” menu button	• Will Abby know what to do at this step? **YES NO MAYBE** • Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)	• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? **YES NO MAYBE** • Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk
• Action #: 1b • Name: Click “Shapes”	• Will Abby know what to do at this step? **YES NO MAYBE** • Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)	• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? **YES NO MAYBE** • Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk

• Action #: 1c • Name: Click the picture of the line	• Will Abby know what to do at this step? **YES NO MAYBE** • Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)	• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? **YES NO MAYBE** • Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk
• Action #: 1d • Name: Click and drag the line on the page to draw it.	• Will Abby know what to do at this step? **YES NO MAYBE** • Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)	• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? **YES NO MAYBE** • Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk

<table>
<tr><td colspan="3">• <b>Subgoal #: 2</b>
• <b>Subgoal name:</b> Change the thickness of the line

<b>Analysis questions</b>:
• Will Abby have formed this sub-goal as a step to her overall goal? <b>YES NO MAYBE</b> (Choose one)
• <b>Why</b>? Especially consider Abby's Motivations & Strategies</td></tr>
<tr><td>• Action #: 2a
• Name: Click shape outline menu option</td><td>• Will Abby know what to do at this step?
<b>YES NO MAYBE</b>
• Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)</td><td>• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal?
<b>YES NO MAYBE</b>
• Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk</td></tr>
<tr><td>• Action #: 2b
• Name: click weight option”</td><td>• Will Abby know what to do at this step?
<b>YES NO MAYBE</b>
• Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)</td><td>• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal?
<b>YES NO MAYBE</b>
• Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk</td></tr>
</table>

• Action #: 2c • Name: click 3pt.	• Will Abby know what to do at this step? **YES NO MAYBE** • Why? Especially consider Abby's Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)	• If Abby does the right thing, will she know that she did the right thing and is making progress toward her goal? **YES NO MAYBE** • Why? Especially consider Abby's Self-Efficacy and Attitude Toward Risk

APPENDIX 2: ADDITIONAL STATISTICAL RESULTS

We compared each dependent variable between each of the situations in Study 2. Each variable was subject to a one-way ANOVA. Table 11 reports the p-value for each test run, comparing the mean values that participants entered on questionnaire 1.

Without the Bonferroni correction, there was a significant difference between Heilbronn University and Oregon State University participants' reporting of the tinkering facet and the items from the BSRI masculinity scale.

Participants from Heilbronn University reported Abby's tinkering at 3.8 (of 7 on a Likert scale), while Oregon State University participants reported Abby's tinkering at 2.7 (of 7 on a Likert scale). These are both correct answers (when measuring participant understanding of the facet), and as such do not change our interpretation of the results.

The BSRI Masculine rating was only one of six ways we measured stereotyping: masculinity, femininity, warmth, competence, friends like or unlike Abby, and reasons friends were identified as like or unlike Abby. Other than the BSRI masculine rating, no other stereotyping results were significant when comparing Heilbronn University to Oregon State University. Thus, we are not inclined to believe that Heilbronn University participants stereotyped more than Oregon State University participants.

Variable	p-Value (Heilbronn v OSU)	Bonferroni Correction (α=.0055)	p-Value (GenderMag v PersonaOnly)	Bonferroni Correction (α=.0055)
Self-Efficacy	.16	not significant	.89	not significant
Risk	.70	not significant	.17	not significant
Motivation	.41	not significant	.71	not significant
Information Processing Style	.29	not significant	.85	not significant
Tinkering	.006**	not significant	.09	not significant
BSRI Masculine	.016**	not significant	.68	not significant
BSRI Feminine	.10	not significant	.69	not significant
SCM Warmth	.23	not significant	.81	not significant
SCM Competence	.65	not significant	.08	not significant

Table 11: Results of one-way ANOVAs for the different situations in study 2.
**=significant difference (p < .05)