Finding Gender-Inclusiveness Software Issues with GenderMag: A Field Investigation

Margaret Burnett[1], **Anicia Peters**[1], **Charles Hill**[1], **and Noha Elarief**[2]
[1]Oregon State University
Corvallis, Oregon, USA
{burnett, peterani, hillc}@eecs.oregonstate.edu

[2]Hewlett Packard
Corvallis, Oregon USA
noha.elarief@hp.com

ABSTRACT
Gender inclusiveness in computing settings is receiving a lot of attention, but one potentially critical factor has mostly been overlooked—software itself. To help close this gap, we recently created GenderMag, a systematic inspection method to enable software practitioners to evaluate their software for issues of gender-inclusiveness. In this paper, we present the first real-world investigation of software practitioners' ability to identify gender-inclusiveness issues in software they create/maintain using this method. Our investigation was a multiple-case field study of software teams at three major U.S. technology organizations. The results were that, using GenderMag to evaluate software, these software practitioners identified a surprisingly high number of gender-inclusiveness issues: 25% of the software features they evaluated had gender-inclusiveness issues.

Author Keywords
GenderMag; gender; usability; field study.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces; K.4.m. Computers and society: Miscellaneous

INTRODUCTION
Significant research has come to light in recent years revealing the lack of gender-inclusiveness in numerous computing situations—spanning education, the computing workforce, and more generally across STEM populations [25, 47]. Numerous efforts have arisen to try to solve these problems, such as changes in curricular and pedagogical practices, and changes in workforce or education climate. (See http://ncwit.org for summaries of much of this work). However, none of these efforts consider the *software itself* that people use on a daily basis to perform computing tasks.

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858274

Evidence has emerged over the past decade that software is subtly undermining females' problem-solving abilities. Recent research has shown that the ways people use software features often cluster by gender, and further, that many software features are inadvertently designed around the way males tend to work with software, as we detail in the next sections. In fact, research shows (at least) five factors that can directly impact the ways males and females use software: their motivations for using the software, their style of processing information, their computer self-efficacy, their attitudes toward risk, and their willingness to tinker.

To put these findings into the hands of software practitioners in an actionable form, we have created GenderMag (Gender-Inclusiveness Magnifier), a new software inspection method to enable ordinary software practitioners to find gender-inclusiveness issues in their software. Its goal is to enable past gender research to make a difference in the design of today's software.

To evaluate GenderMag's effectiveness in the hands of software practitioners developing real software projects, we conducted a multiple-case field investigation on three large, U.S.-based organizations: an agency of state government, and teams at two multi-national technology companies located on opposite sides of the U.S. Our goal was to investigate GenderMag's effectiveness; its value; and its strengths, weaknesses, and problems in real-world situations. We structured our investigation around the following research questions:

- RQ1 (*Gender-inclusiveness issues*): Does the GenderMag method reveal gender-inclusiveness issues in real-world software?
- RQ2 (*Personas and facets*): Which of the GenderMag personas and persona facets are the most useful to real-world software practitioners at revealing gender-inclusiveness issues?
- RQ3 (*Utility*): Do GenderMag's results have practical utility? If so, how do real-world practitioners take follow-up action?
- RQ4 (*Gender*): How do practitioners who use the method ultimately view its interactions with gender as a concept and with their own gender identities?

BACKGROUND AND RELATED WORK

Prior research has already established that the individual differences in how people use software features aimed at supporting problem solving tend to cluster by gender. Such research has spanned numerous software domains; a partial list is spreadsheets [5, 6, 7, 28, 29, 33, 60], visualization systems [8, 61], online classwork platforms [50], web and home automation [18, 53, 55], working with intelligent agents [39], and programming tools [13].

Research to address the need to improve the gender-inclusiveness of software falls generally into two categories. The first category is to develop demonstration software products, and the second category is to develop general methods that can be used across a class of software to either improve or to evaluate the gender inclusiveness of that class of software.

In the "demonstration software" category, some software projects aim to appeal specifically to females. Kafai and Burke term this kind of approach "building new clubhouses" [36], as a counterpoint to the well-known work by Margolis and Fisher about "unlocking" the (male-only) computing clubhouse [43]. Examples of tech products designed specifically for females include Goldiblox and Storytelling Alice [38]. Goldiblox is an interactive book series plus accompanying construction set starring Goldie, the girl inventor who loves to build. The products are marketed as "construction toys for girls" (http://www.goldieblox.com/) and intertwine the engineering play with appearances and themes common in toys for girls. Storytelling Alice takes into account the difference in males' and females' motivations toward using technologies. It then extends the Alice programming language and environment by supporting storytelling through programming, thereby increasing middle-school girls' learning of computer programming [38].

Other demonstration projects aim to appeal to both males and females, often by removing barriers or enhancing features that tend to particularly affect one gender. This kind of approach has a pluralism theme such as advocated by Bardzell [4], i.e., the idea that most individuals do not fit statically into a single gender bin [16], and that removing barriers to entry can help everyone regardless of the gender with which they identify.

An example of the pluralism approach is Gidget, a debugging game for novice programmers. Its gender inclusiveness comes from innovating certain programming environment characteristics, such as portraying the computer as fallible, personifying error messages, and presenting explanatory help in forms compatible with both females' tendency toward comprehensive information processing and males' tendencies toward selective information processing [35, 40, 41]. Another example is LilyPad [10, 11]. LilyPad is a "maker" product with the same functionality as Arduino, but for wearable computing projects. Thus, it combines the "build it" tradition of boys' play worlds with craft traditions like sewing and textiles of girls' play worlds [36].

Still another example is StratCel [27], an add-on for Excel, which supports problem-solving strategies statistically associated with females in addition to those statistically associated with males [60].

Demonstration projects like these are important for not only demonstrating that greater inclusiveness is possible in problem-solving technologies, but also for providing exemplars of how to go about it. However, a disadvantage of these kinds of projects is that they tend not to scale up. That is, each such project tends to be expensive, requiring extensive research, prototype building, and empirical work for each separate project.

At the opposite end of the spectrum from developing demonstration projects is developing new methods and practices for avoiding or identifying gender-inclusiveness issues in software. The advantage of methods is scalability: if the methods are shown to be effective, they can help large numbers of projects detect and/or avoid gender-inclusiveness issues. Processes for design and decision-making are examples of such methods. For example, Williams captures a number of design process recommendations that are about including females in the decision-making processes that shape software [64]. The GenderMag method used in this paper's field investigation falls into the methods category. This paper is the first to empirically investigate the use of GenderMag in the field.

THE GENDERMAG METHOD

GenderMag (Gender-Inclusiveness Magnifier) is a usability inspection method for evaluating problem-solving software. We have detailed elsewhere [14] the formation of GenderMag, its formative empirical work, and a controlled lab study. Thus, here we present only enough of the method needed for interpreting the in-the-field results presented in upcoming sections.

GenderMag focuses on five facets of gender differences that have been extensively investigated in the literature pertaining to problem solving. It encapsulates them into a set of faceted personas to bring them to life, and embeds their use into a systematic process based on a gender specialization of the Cognitive Walkthrough (CW) [59, 63]. The five facets are:

Motivation: Research spanning over a decade has found that females tend (statistically) to be motivated to use technology for what they can accomplish with it, whereas males are often motivated by their enjoyment of technology per se [12, 13, 19, 31, 33, 37, 43, 57]. This difference can affect which software features users choose to use.

Information processing styles: To solve problems, people often need to process new information. Females are more likely (statistically) to process new information comprehensively—gathering fairly complete information before proceeding—but males are more likely to use selective styles—following the first promising information, then backtracking if needed [17, 22, 45, 46, 52]. Each style

has advantages, but either is at a disadvantage when not supported by the software.

Computer self-efficacy: Self-efficacy is a person's confidence about succeeding at a specific task, and influences their use of cognitive strategies, persistence, and strategies for coping with obstacles [3]. Empirical data have shown that females often have lower computer self-efficacy than males, and this can affect their behavior with technology [5, 6, 12, 13, 24, 32, 34, 43, 49, 50, 58].

Risk aversion: Research shows that females tend statistically to be more risk-averse than males [23], surveyed in [62], and meta-analyzed in [21]. These results span numerous decision-making domains, such as in ethical decisions, investment decisions, gambling decisions, health/safety decisions, career decisions, and others. Risk aversion with software usage can impact users' decisions as to which feature sets to use.

Tinkering: Research across age groups and professions reports females being statistically less likely to playfully experiment ("tinker") with software features new to them, compared to males. However, when females do tinker, they tend to be more likely to reflect during the process and thereby sometimes profit from it more than males do [6, 13, 18, 20, 33, 54].

GenderMag brings these facets to life with a set of four faceted personas—"Abby", "Pat(ricia)", "Pat(rick)" and "Tim". Each persona represents a subset of a system's target users. Personas are widely used in industry, sometimes simply to communicate about user needs during design phases of software development, such as via ideation and role-playing during informal tests, and sometimes for much more [26, 44, 48, 51].

Abby, Patricia, Patrick and Tim are identical in several ways: all have the same job, live in the same place, and all are equally comfortable with mathematics and with the technology they regularly use. Their differences are strictly derived from the gender research on the five facets. Abby and Tim have existed for over a year; the Pats were added about the time of Company E's session. Tim's facet values are those most frequently seen in males (e.g., Figure 1), Abby's facet values are those frequently seen in females that are the most different from Tim's, and the two Pats' (identical) facet values add coverage of a large fraction of females and males different from both Abby and Tim. The

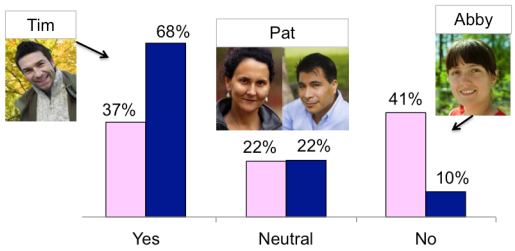

Figure 1: A portion of the empirical foundations behind the personas' Motivation facet values. (See text.)

two Pats' identical facet values are to raise awareness that differences relevant to inclusiveness lie not in a person's gender identity, but in the facet values themselves.

To illustrate how the foundational data maps to the personas, Figure 1 demonstrates a portion of the data behind the Motivation facet values of each persona. As with all the facets, Motivation is backed by multiple studies, but for simplicity of presentation, only one of them is illustrated in the figure, namely a study in [13]. In that study, about 2/3 of males and 1/3 of females were motivated by exploring next-generation technology, and this value for the Motivation facet is covered by Tim; about 1/5 of both males and females felt neutral about it (covered by the two Pats). The largest percentage of females and smallest percentage of males did not enjoy exploring next-generation technology (covered by Abby). Figure 2 shows the persona side of such mappings, with snippets of how another facet, Self-efficacy, maps to each persona.

GenderMag intertwines these personas with a specialized Cognitive Walkthrough (CW). The CW is a long-standing inspection method for identifying usability issues for users new to a system or feature [63]. In a GenderMag CW, evaluators answer the following questions for each step of a detailed use case (goal and list of actions). Red font shows the GenderMag specializations:

(Subgoals question): Will <persona> have formed this sub-goal as a step to their overall goal? Why? (refers evaluator to the pertinent facets)
(Actions question #1) Will <persona> know what to do at this step? Why? (refers evaluator to the pertinent facets)
(Actions question #2) If <persona> does the right thing, will s/he know that s/he did the right thing, and is making progress towards their goal? Why? (refers evaluator to the pertinent facets)

EMPIRICAL METHODOLOGY

To investigate GenderMag's usage under real-world conditions, we used a multi-case study design. Each "case" was one organization's software teams' usage of the GenderMag method to find gender-inclusiveness issues in their own products.

Organizations learned about GenderMag from our website or from talks at conferences and meetings. When an organization decided to use the method, we asked if we could observe. Three organizations agreed: a state **g**overnment agency (abbreviated G), one **e**ast-coast-based team at a

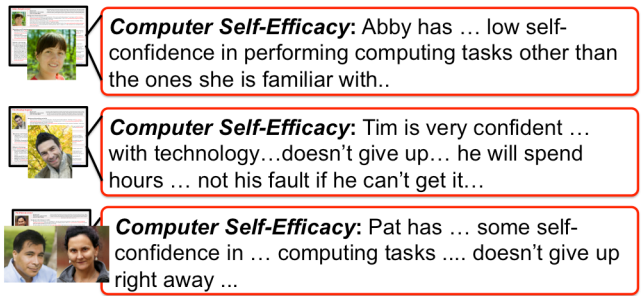

Figure 2: The self-efficacy portions of Abby, Tim, and 2 Pats, drawn from self-efficacy theory and empirical data (see text).

multi-national hardware/software company (E), and a **w**est-coast-based team at another multi-national hardware/software company (W). As Table 1 summarizes, sessions spanned multiple software types and platforms, software maturity levels, gender make-up of the teams, and personas the teams chose to use.

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. When needed, we helped them with the set-up. We used the results of each session to iteratively inform and refine the method, so the GenderMag method itself changed between some of the sessions.

The "case" officially began when the team-appointed facilitator started the evaluation session, which usually lasted about 2 hours. A team-appointed recorder captured the issues on GenderMag forms. Everyone served as evaluators. We observed, video-recorded and later transcribed each session. At the end of each session, we observed the team debriefing, and performed a short, semi-structured group interview about the issues they had found and any follow-up actions they planned. Two weeks later, we conducted follow-up interviews with any individual team members willing and available to be interviewed.

To analyze the data, we qualitatively coded four data sets: transcriptions of the main evaluation, group debriefings, follow-up interviews, and the handwritten, gender-specialized CW forms the teams had filled out during their evaluation sessions. Table 2 details the code set. We coded the transcribed data in 30-second segments, and coded the forms at each feature evaluated. Multiple codes were allowed. Two independent coders achieved inter-rater reliability rates of 80% agreement on 20% of the data, so one researcher then finished coding the rest of the data.

We guarded the rigor in our case study through triangulation—independent sources showing the same phenomenon. In case studies, each "case" is the counterpart to an entire experiment [56, 65], so a counterpart to increasing the number of chances to refute a conclusion in a case study is triangulation. We triangulated in three ways. First, we used methodological triangulation (using observation, group debriefing interviews, and follow-up interviews to collect data). Second, we used investigator triangulation (multiple observers and IRR in coding). Finally, we used data source triangulation: data whose sources were multiple, independent situations. We did this at both a high level (different organization types, geographical locations, genders, etc.) and a low level (multiple data sources reporting the same issue in the same software), as will be seen in the sections that follow.

Code	Description
Issues:	Any issue/problem in the software, and/or the fix.
Method:	Comments on ease or difficulties in using GenderMag.
Persona:	Comments about a persona or personas in general.
Gender:	Comments on the concept of gender.
Facets (General):	About facets but not specific about which one.
Facets (Specific):	Referred explicitly to one of the 5 facets (GenderMag section): M,I,SE,R,T, plus F(amiliarity) as a convenience later allocated across the 4 facets that refer to it.

Table 2: Code set used in data analysis.

RESULTS: GENDER-INCLUSIVENESS ISSUES (RQ1)

The ultimate purpose of GenderMag, and indeed of any usability inspection method, is to find issues, so we begin with the "bottom line."

All four teams found issues from the perspectives of the persona they used. In fact, 3 of the 4 teams found large numbers of them. As Figure 3 summarizes, Agency G's Team GB found issues in 18% of the actions and subgoals they evaluated, and the other three teams found issues in at least 68% of the actions and subgoals they evaluated. In total, the four teams found issues in 55 of the 99 actions/subgoals considered (55%).

A surprisingly large fraction of these issues were gender-inclusiveness issues. To calculate this fraction, we defined an issue as being a gender-inclusiveness issue if the team used one or more of the facet values in their persona to identify it, because the facets (attitude toward tinkering, risk, etc.) represent the empirical findings of individual differences by gender discussed in the previous sections. The result was that 25 of the 55 issues were gender-inclusiveness issues (45%). Stated another way, these software practitioners found gender-inclusiveness issues in one out of every four actions/subgoals they evaluated in their own software products (25/99), a total of 25%.

	Govt. Agency G	Company E	Company W
Teams & Sessions	2 mixed-gender teams, 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	Traffic situation problem-solving.	Machine learning algorithm analyzer.	Mobile app for document delivery.
Software maturity	Very mature (about 10 years old).	Pre-release (still in initial development).	Post-release, active evolution restarting.
Software is for...	Operators capturing travel information to inform travellers.	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.

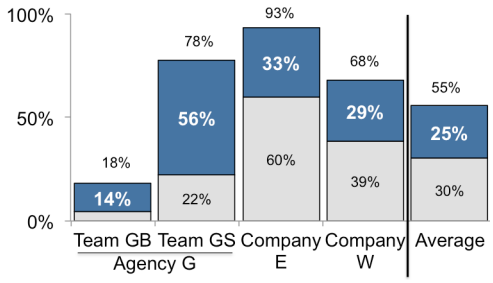

Figure 3: Issues each team found as a percentage of the number of user actions and subgoals evaluated. Above bars: total issues. Dark blue: gender-inclusiveness issues. Light gray: other issues.

GOVERNMENT AGENCY G & RQ2-RQ3

To gain insight into the differences between Agency G's Team GB versus the other teams requires a closer look, starting with Agency G.

The impetus behind Agency G's interest in trying GenderMag was that almost all of their software's end users are female. Their end users are operators—mostly female—who use the software to capture and summarize travel-related information as quickly and precisely as possible and then to inform travelers via a website. The software has been in use for about 10 years.

Two Agency G teams (GB and GS) elected to try GenderMag, via separate sessions. The teams used GenderMag to evaluate the same software, and both chose Abby (the only female persona who was ready at that time) because of their software's female user base. However, they evaluated different tasks (use cases) of that software.

Initially, Team GB's interest was fairly low. They conducted the session because they were told to do so: their supervisor had seen GenderMag in pilots and required his team to try it, but the supervisor was not able to attend himself. The team's evaluators were the lead developer (male) and two intern developers (1 male, 1 female). The software undergoes regular maintenance, and this team does most of that maintenance when the "client organization" (Team GS) requires it. Team GB is not co-located with an operations center, and they rarely see the operators.

The client organization, Team GS, had their session four months after Team GB's. Unlike Team GB, Team GS was eager to try GenderMag, because by then they had learned that Team GB had identified gender-inclusiveness issues in their shared software. Six of Team GS attended (4 male, 2 female): three were IT managers, one was software project manager, one was a developer, and one a systems analyst.

Agency G meets Abby & the facets

Team GB's experience was uneven, starting with the Abby persona itself. They used the earliest version of the method, and at that time, we had left implicit the fact that the software was new to Abby. Newness is important, because CWs are about early experiences, not usability by expert users. Since that aspect was not specified, Team GB's customization of Abby put her on the job for six months. Given this mismatch to the CW component, they found few issues, instead answering most of the CW questions using their expectations of the operators' training program:

> GB1f (debriefing session): she was proficient with the technology ... we referred to <that> a lot.

Team GB did not seem to have much empathy for Abby, and occasionally found her to be frustrating. In contrast, Team GS's found Abby to be such a good match to much of their user population, they at first mistook Abby to be one of their real operators:

> GS3f (in response to the description of Abby): Oh, you mean <operator name>, right?
> GS1m: There is at least one person in every <unit> that can fit this <description>.
> GS3f: <Our users are> majority female, most risk-averse, ... few are tinkerers ... introducing new features is difficult, you have to show them the value or it's an uphill battle.

Note how GS3f's "majority female" quote refers to at least three of Abby's facet values: risk, tinkering, and motivation. The facets have an important function in the GenderMag method: they are the method's primary technology transfer devices for making the research foundations actionable. Figure 4 summarizes how much each facet played into the teams' deliberations. (The figure shows maxima instead of totals, because spoken evaluations could duplicate written evaluations if teams read aloud what they wrote.) Both teams used tinkering the most.

Perhaps the most telling measure of the facets' effectiveness is that both teams realized by the end of the sessions that the gender inclusiveness issues they were finding were not simplistic, gender "binary" values. Instead, as we had hoped, they realized that *inclusiveness lies in supporting a range of facet values*, not in gender stereotyping:

> GB2m: It wasn't necessarily about the gender of the persona though, it's just ... her <facet values>.

GenderMag's utility to Agency G

Table 3 enumerates the gender-inclusiveness issues the two teams identified and how they identified them. To be conservative, we report only the issues explicitly identified by facets on the teams' written forms. As the table shows, the results are triangulated across two and sometimes even three data sources (with "form" used as the gold standard).

Empirical research has previously established that a high percentage of issues CWs reveal are indeed valid issues (i.e., that CWs have a low false positive rate). For example, Mahatody's survey reports false positive rates ranging from about 5% to about 10% [42]. Thus, it seems reasonable to assume that almost all of the issues the teams found using our variant of the CW were indeed real usability issues.

However, prior research showing that CW issues are mostly valid does not answer the following: did the teams believe that these issues, presented by GenderMag as mattering the most to Abby-like users, actually were issues for Abby-like users, and did the teams deem them to be important?

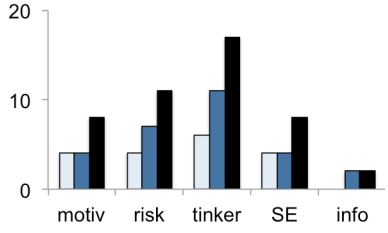

Figure 4: Each facet's maximum number of written and spoken mentions. (Light=by Team GB, Medium=by Team GS, Black=in total.) Both teams referred to tinkering the most, but both also used 3-4 other facets.

Team	User actions with gender-inclusiveness issues	Why issue for Abby	Data sources: form	video	eval	Facet that found
GB	B-T1.s3: See what changed	GB3m: ...do not know if <Abby> would notice... she likes to avoid more troublesome features	√	√		T
GB	B-T2.1c: Open <screen>	GB3m: ...depends on <familiarity>	√	√		F
GB	B-T2.2a: (same as above)		√	√		F
GS	S-T1.s2: Enter <tag>	GS3f: How would <Abby> know...? ... She prefers <to> follow a step-by-step...	√	√	√	T
GS	S-T1.2b: Enter <value> in <field>	GS1m: She avoids troublesome features... it's not intuitive... she tends to blame herself...	√	√	√	T,S
GS	S-T1.2c: Press enter	GS6f: If you had moved your mouse ... you're in trouble. I don't think it's intuitive and.there's not a step-by-step...	√	√	√	T
GS	S-T1.s4: Update map	GS2m: There's no clear indication of where you need to ... the next action which will take them to a map.	√	√	√	T
GS	S-T1.4a: Click <button>	GS3f: I don't even know why you would hit that ...she doesn't tinker so she's going to be hesitant to just push buttons and see what they do.	√	√		T
GS	S-T1.4c: On the map, click <place>	GS1m: You have to understand the system entirely...she didn't like tinkering GS4m: She doesn't like to tinker, ... she's risk-<averse>	√	√	√	R,T,F
GS	S-T2.s1: Provide details.	GS3f: No, it's gotta be in her checklist or training... how does she know where to go next, 'cause <object> is way over there?	√	√		F
GS	S-T2.1a: Select <location>	GS6f: ...averse to tinkering... don't know what the rules ...	√	√		T
GS	S-T2.1b: Select <value>	GS6f: There's no um... GS3f <interjects>: ...tutorial or ... guide... GS3f: We would have been totally lost.	√	√	√	F
GS	S-T2.1c: In <field>, enter <details>	GS2m: ...what you cannot see here is this field... GS3f: If she's risk-averse, she's going to...	√	√	√	R,T

Table 3: The 13 gender-inclusiveness issues Agency G teams found using Abby. The teams also found 5 other usability issues (not shown). Data sources columns: the CW forms the teams filled out, their videotaped discussions of Abby, or an evaluator experiencing it him/herself. The facet column uses the codes in Table 2.

At first, Team GB was not sure what they thought about the value of using GenderMag. In fact, Team GB initially (in the debriefing session) said that they were unlikely to use the method again. However, when they described the issues they had found to their manager, he confirmed that he had actually witnessed one of their users having exactly that issue. This confirmation was very convincing to Team GB, and they ultimately decided that all the issues they had found warranted a discussion with the client. Further, their belief in the validity of the issues changed Team GB's mind, and they decided to conduct more GenderMag sessions on a new software product they are developing.

Team GS regarded the issues to be very credible right away, in part because they themselves experienced many of them. In fact, Team GS's transcripts show 27 different instances in which the team themselves expressed confusion about the interface, sometimes for long stretches of time:

> GS1m: Well, I am just trying to figure out ... even know to click the <label> button.
> GS6f: So the <tag> changes without an explanation why...

Follow-through: What came next at Agency G?

The ultimate test of utility is follow-through, and the maturity of the software paired with the distributed nature of decision-making relating to that software made issue follow-through difficult. Inspection methods work best when the software's owners have motivations to change the software—and budgets to do so. As GS4m put it:

> GS4m: We needed this ... to happen 9 years ago.

Despite that, a total of four of the issues were deemed by the teams to be important enough to investigate fixing. Also, both teams decided to keep using GenderMag in different ways. We already mentioned that Team GB plans to use it on a different product they are working on. Team GS also abstractly discussed using it when designing new software, but their immediate plans are to incorporate aspects of GenderMag in their emerging GUI standard revision:

> GS1m: So ... we're revising our GUI standards and we will include some high-level GenderMag concepts in that.

The Next iteration: What came next to GenderMag?

As the first two teams ever to try GenderMag in real-world situations, Teams GB and GS revealed a number of ways we could improve the method, so we made several changes.

First, we needed to address a "groupthink" problem. Some team members allowed themselves to be talked into or out of their opinions. To solve this, we added an explicit "maybe" choice to all the CW questions, and changed the method's instructions to emphasize that teams needn't agree, but only to record *all* the team members' perspectives. Second, we made explicit that the software systems were new to the personas. Without that, we noticed that Team GB members convinced each other that many issues would be taken care of as a result of training, which may have masked several issues. Finally, we changed to a variant of Spencer's streamlined CW's evaluation questions [59] for clarity and efficiency. We then took the updated GenderMag method to Company E.

COMPANY E'S RESULTS FOR RQ2-RQ3

Company E's team had a high degree of buy-in, perhaps because one of its members advocated heavily for its use. The team evaluated a system with a fully developed back-

end, but a user interface still in its early stages. The system aims to help developers who are *not* machine learning experts to select a machine learning algorithm to insert as a "black box" into an application. To use it, a developer inputs data, and the system provides statistics on several algorithms to help the developer choose which to incorporate.

Initially, three team members planned to participate in the session, but due to scheduling problems, the original advocate (female) was unable to attend, leaving only two of the team (both males) to participate. The two participating team members were the software's UI developer and a machine learning expert who helped to inform the project.

Unfortunately, Company E was particularly strapped for time, and they could not manage to do much of the setup in advance. Thus, Company E went into the evaluation with only a high-level task sequence. More critically, they arrived with only the base Abby persona, without customizing her, so Abby lacked the necessary background to use their software.

Despite the abbreviated setup, the team compensated. They worked out task details on the fly when needed, and used "a bit of imagination" to quickly transition into evaluating the task as though Abby had the necessary background:

E2m: We would need a bit of imagination because we don't think an Accountant would use this...
E1m: ...she may not understand the whys. But if she does, if she makes the connection between the list and the chart, I guess she understands what is going on.

Company E: "Channeling Abby"

Once they had mentally adjusted Abby's background, the team stepped fully into her character, and remained in-character throughout. Key to their results, although the team adjusted Abby's background in an ad hoc fashion, they were careful not to damage the essence of Abby, namely her facet values. Indeed, they embraced emulating Abby's cautious nature, her low self-efficacy with the tool, and her non-tinkering ways of going about her work:

E1m: (imitating nervous person) if you are nervous and you are not sure what you are doing and you want to click and suddenly it moves to a new place...
E2m: Yeah, you are completely in Abby now.
E1m: Yeah, I am channeling Abby ... and I am not having fun with this program.

Ultimately, both team members decided that the Abby persona was realistic and nuanced in ways that were both useful and challenging:

E1m: I definitely know... people who are... nervous about just randomly smacking keys around & pointing at things & clicking...
E2m: She is ... a bit curious but then she is not, and then you somehow have to draw a line yourself ... She was not too curious, that was good...She is more down-to-earth ... I liked her.

Interestingly, as both Agency G teams had independently come to realize, the Company E team also eventually realized that the method's power was not in gender per se, but in the facet values:

User actions with gender-inclusiveness issues	Data sources: form	Data sources: video	Data sources: eval	Facets that found
T1.1a: Click 'choose file'	√	√	√	R,T
T1.1d: Click submit	√	√	√	R,T
T2.s3: Monitor performance	√	√		R
T1.2b: View statistics	√	√	√	T
T1.3a: Click 'like'	√	√	√	T

Table 4: The 5 gender-inclusiveness issues Company E found. The team also found 9 other usability issues (not shown).

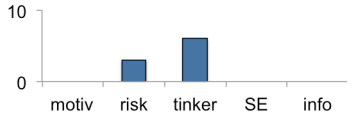

Figure 5: Company E, like the Agency G teams, emphasized Tinkering the most, followed by the Risk facet. Company E did not use the other facets.

E1m: Is it really about gender differences, or is it just about people differences?
E1m: ... using the core personality characteristics ... will be helping women as well.

Utility to Company E: "Easy to detect"

Despite what E2m had originally described as a "simple interface", the team soon realized there was no shortage of issues to be found. Almost every feature had an issue. In total, the team found 14 issues in the software. Of these, 5 were gender-inclusiveness issues, which they found using the risk and tinkering facets (Table 4 and Figure 5).

E2m: As Abby, I found this <system> to be an unmitigated disaster.

Within two weeks of the session, the team had fixed three of the issues that they found. As E2m summarized in his interview, the fresh perspective on their software that GenderMag had brought made it possible for the two of them, neither of whom were usability professionals, to see issues that would be important to users with Abby's facets:

E2m: Once you have a different viewpoint, it makes it easy to detect those things.

COMPANY W AND RQ2-RQ3

At first, the Company W team was ambivalent about the idea of using GenderMag. Then one member of the group persistently championed the idea, and the team became enthusiastic about trying it. They decided to conduct two sessions—one with Abby and the other with Tim. So many team members turned up for the first session that they had to split up who would attend the first session and who would attend the second. Four males and one female participated in the first session, using Abby as their persona (Session #1). Since Session #1 did not complete the entire task sequence they had planned, they decided that Session #2 should continue with the Abby persona. At that session, two females and six males participated. There was about a 50% overlap between the evaluators in the two sessions, with 3 people (two males and one female) attending both sessions.

A Company W researcher observed the two sessions along with two external researchers.

The team deemed both sessions very useful, so they decided to conduct two more sessions. Although these latter two sessions did not have external researchers present, we have data from their forms and a debriefing interview, which we include here. In Session #3, the team completed the evaluation of the task list from Sessions #1 and #2 with the Abby persona. In Session #4, they re-evaluated the entire task list—this time using the Tim persona.

The software they evaluated was a mobile app that had been released earlier, and was now entering a new stage of active improvement. Each team member had a mobile phone with a freshly installed app for evaluation. The phones had different versions of the Android operating system and different levels of cellular service (some devices had no cell service at all) to cover a variety of real-world situations.

How Company W reacted to Abby's world

When working with the Abby persona, the team was surprised at some of her facets. For example, W4m wondered whether anyone really blamed themselves for software's bad behavior, so he decided to ask his wife about that facet. Later, another member of the team mentioned that W4m said his wife's response to the conversation was:

> "Welcome to my world!"

When we asked him about the conversation in follow-up interviews, he explained:

> W4m: She had the same characteristic as Abby in that if she tried something new, it was only because she had to, and if it didn't work, she would blame herself for the failure rather than the software...
> I asked her if that is *really* how she thinks and she <said>: "yes, that is exactly how I think."

W4m's story was one of several examples in which Team W's use of Abby opened their eyes to users they had not thought about before:

> W1m: Her attitude to technology was the most interesting ... 'Cause that's something that's really different from me.
> W3m: …made us think about how a real person would approach this. As opposed to a person we wish would be there.

However, by providing only two personas at extremes from one another, one male and one female, the method was in danger of stereotyping men and women; some users of the method might see the personas as being representative of all males or all females. Indeed, W9f told us that she had overheard male participants from Session #1 talking in the hallways, saying things she paraphrased as:

> "Today I learned that women are this and women are like that."

To help guard against such stereotyping, we had already begun developing two additional personas, Patrick and Patricia, described in the GenderMag section. We then showed the two Pats to Team W members during follow-up interviews, with mildly favorable response, but have not yet had an opportunity to investigate their actual use in the field.

> W4m: If you have only a limited amount of time for this kind of process then 2 <personas> is as much as you do. <But> if you have a lot of time and you are really focused on this,... then definitely 3 or 4 would be to make sure you cover your bases.

Company W: "GenderMag has infected us"

Company W identified 7 gender-inclusiveness issues in total (Table 5): 6 gender-inclusiveness issues for Abby, and 1 gender-inclusiveness issue for Tim. As Figure 6 shows, all five facets were discussed in finding these issues. Some of these issues had remained unnoticed for months:

> W6m: I've <done the sequence> many times now and I never have taken the time to methodically go through ... each screen.

We have already pointed out that one measure of the method's utility is follow-through in terms of fixing issues found during the session. This turned out to be a little more complicated than expected because, as with Agency G, Company W's case featured distributed responsibility for the software. For some issues, the original software designers were the ones who needed to make the gender-inclusiveness issue fixes that Company W's team envisioned. Still, Company W's team felt so strongly about the importance of three of those issues, they eventually convinced the original designers to fix them. The method also led the team to start developing a new usability (automatically logged) metric

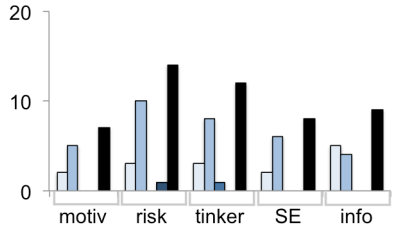

Figure 6: The Company W team used all five of the facets, with Session 2 especially emphasizing the Risk facet, and Session 1 especially emphasizing the Information Processing facet. (Light blue to dark blue=Sessions 1 to 4. Black=Total.)

User actions with gender-inclusiveness issues	Session#: Persona	Notes	Data sources: Form	video	debrief	eval	Facets that found
T1.s1: Use <app>	#1:Abby		√	√			T
T1.1b: Select <option>	#4:Tim	Reported *Abby* issue			√		T
T1.1b: Select <option>	#4:Tim	*Tim* issue (≠ Abby's)			√		T
T1.s2: Pick the <device>	#1:Abby		√	√		√	R
T1.2a: Skip Intro	#1:Abby		√	√		√	R
T1.2b: Opt in	#2:Abby, #4:Tim	Both reported *Abby* issue	√	√	√	√	R,M
T1.4d: Settings	#3:Abby				√		T

Table 5: The Company W team found 7 gender inclusiveness issues, and 13 other usability issues (not shown). The debrief data source was an interview about sessions #3 and #4.

for their deployed software that would more easily attract the original designer's interest, namely users actually completing certain (tool-detectable) use cases they started.

Company W also showed their view of the method's utility another way: by taking the initiative to conduct GenderMag evaluations on other applications.

> W3m: Well, I did some ... other <GenderMag evaluations>...

In fact, W3m decided to give his entire testing team experience with the method, and to start using it more broadly:

> W3m: We ... want to do it across more <software, platforms and devices>... really powerful the way they are written up.... That was really helpful to me in testing.... Epiphanies... of "wow, this is a huge problem".

Ultimately, W6m summed up GenderMag's effects on the Company W team as permanent and impactful:

> W6m: <GenderMag> has already infected us.

GENDER (RQ4): RESULTS AND DISCUSSION

Should the personas have a gender?

One point that was raised by every team was whether gender should be explicit in the method. Table 6 shows their differing points of view.

On one hand, as already pointed out, every team realized that the essence of the approach was not gender categorizations, but rather the facets. Further, some worried that having the personas be assigned genders that align with the statistical foundations—or having any gender at all—might be inappropriate. On the other hand, some team members thought the personas' gender assignments imparted subtle wake-up signals that may have helped them realize the importance of taking that persona's point of view. As Table 6 shows (➔🡐), some team members even took both sides of the question.

As to the idea of genderless personas, some team members found the "person-ness" of the persona to be immensely valuable and necessary, consistent with the ideas of advocates of personas; e.g., "Personas put a face on the user—a memorable, engaging, and actionable image" [1].

Grudin's analysis of the psychology of personas is consistent with this view, i.e., that having a persona seem like a real person (hence having a gender) matters [30]. As Grudin explains, the point of personas is to promote engagement: the more a designer engages with a persona, the better able the designer is to predict and evaluate how such people will behave in new situations such as with the designer's software features.

Personas promote engagement by leveraging a universal skill: humans' ability to build models of people by drawing from their experiences with those people and other people. 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 [30]. In essence, designers' ability to engage and empathize with personas comes in part from the fact that personas seem like people instead of like lists of facts. This explains why the teams felt able to predict what Abby or Tim would do with their software.

Even if it were possible to present a convincingly person-like yet genderless persona, Bradley et al. found that when given the genderless word "user" and asked to draw and classify the user as a gender, males classified their users as males 80% of the time, and even females classified their users as males 60% of the time [9]. Thus, if we did not provide a gender, most personas would have genders anyway—and the genders would mostly be male.

The teams' genders

The ways in which the team members reacted to the experience of using GenderMag seemed to cluster according to their own genders. The males tended almost unanimously toward three reactions: (1) strongly identifying with the Tim persona; (2) an emerging set of real insights into the Abbys of the world, and (3) an appreciation of the value of their newfound insights:

> W1m: I think we're missing out on an awful lot of Abbys.

The female team members who, like the males were also software developers and technical managers, experienced many of the above insights—but they also experienced more complex and nuanced reactions. First, as the males had done with Tim, these technical females strongly identified with Abby—except for GB1f. In fact, GB1f was uncomfortable and frustrated with Abby. She wanted Abby to succeed the way she (the evaluator) would, consistent with

Con: Having gendered personas is irrelevant or worse		**Pro:** Personas should be gendered & person-like; gender is pertinent
GB1f: I don't know if I was necessarily considering gender as much as just ... her personality.		GS2m: ... you are so focused on your project team members... <but> you have this broad audience, gender is always going to fall into that.
E2m: I did not see it as anything related to gender … this could also be a guy... It is kind of a form of discrimination		W3m: <person-like characteristics> help to personify the facets.
GB3m: I don't know how it would help to identify and find the different behavior between genders.	➔🡐	GB3m: Maybe I *should* think she's a woman. ... Because I never think like a woman!
E1m: Is it really about gender differences, or is it just about...characteristics of *people*?	➔🡐	E1m: ... <think more about> women, who are not traditionally taken in consideration.
W9f: I'm not convinced at this point that it has anything to do with gender.	➔🡐	W9f: We as society are trained to help take care of ... females, <so> a female persona can pull out the "I really need to pay attention"... <but for> a male persona ... "he ought to have figured out a way around this".

Table 6: Examples of the teams' deliberating pro vs. con as to whether gendering the personas matters.

her own facet values, not the facet values Abby had:

> W9f: I'm Abby! … I was risk averse and I thought this was perfect.
> GB1f: I am not like Abby... It was hard for me to step into those shoes... <I wanted to tell her> "Come on! It's okay to click something!"

On the other hand, in every team that included females (i.e., all teams except Company E), the females championed Abby's characteristics and working style more frequently than their male peers did:

> GB1f: She knows what she is doing.
> GS3f: She is proficient in what she uses.

> W2f: Abby really enjoys using technology that she's familiar and comfortable with.
> W9f: She's proficient with technology she uses and knows a lot about mobile phones.

This identifying with and championing of Abby by technical females has also occurred in earlier presentations and meetings about GenderMag. At times, technical females seem to experience a sense of validation—that the ways in which they problem-solve are not, after all, deficient simply because they are different from many technical males' ways, and that voices like theirs are finally able to be heard.

		Agency G		Company E	Company W			
		Team GB	Team GS		#1 Abby	#2 Abby	#3 Abby	#4 Tim
RQ1	Found inclusiveness issues	√	√	√	√	√	√	√
	% of actions with gender-inclusiveness issues	14%	56%	33%	50%	50%	7%	10%
RQ2	Most useful persona	Abby>Tim	Abby>Tim	Abby>Tim	Abby>Tim			
	Most useful facets	Tinker> Risk=Motiv.=SE	Tinker>Risk> Motiv.=SE> Info.	Tinker>Risk	Risk>Tinker>Info> SE>Motiv.			
	Understood that inclusiveness lies in supporting diverse facet values	√	√	√	√			
RQ3	Utility	√	√	√	√			
	Follow-up type	Took to client, use again	Standards	Fixed 3 issues	More sessions, fixed 3 issues, wider adoption			

Table 7: Triangulation of results across sessions and teams.

RESULTS TRIANGULATION

Table 7 shows the triangulation of results across the cases. Each research question occupies a major row of the table. Subrows add additional details to the major rows.

In the table, each "√" denotes a session that provided evidence relevant to a particular research question (row) and a particular case (column). If multiple sessions contributed different results, the notations in parentheses clarify the session that produced that evidence.

As the table shows, evidence from multiple real-world cases and software platforms pointed to the same results for three of the four research questions. The fourth, gender, gave more nuanced similarities and differences, as we discussed in the previous section.

CONCLUSION

This paper presents the first in-the-field investigation of professional software practitioners' use of GenderMag—or of *any* systematic method—to concretely identify gender-inclusiveness issues with their software's features.

All three organizations, one of which was a state government agency and two of which were multinational hardware/software companies, were able to put GenderMag to good use. Their software spanned a range of types and maturity levels, their evaluation teams spanned a range of job titles in the software industry, and the organizations had varying decision-making situations and cultures. Despite these differences, the results consistently showed that:

- *Gender-inclusiveness issues*: All teams found gender-inclusiveness issues in their software: in total, 25% of the software features they evaluated showed gender-inclusiveness issues. In many cases, these issues had gone unnoticed for months or even years.
- *Personas*: Despite some hiccups along the way, the two personas that teams used (Abby and Tim) both enabled the teams to identify gender-inclusiveness issues. Most issues were Abby issues, but the team using Tim also uncovered a Tim issue.
- *Utility*: All the teams found value in using the method. Agency G's teams found four issues that they deemed important enough to pursue fixing, even though their software had been in maintenance status for years. Team E fixed 3 issues right away, and Team W convinced the software's designers to fix 3 issues. Teams GB, GS, and W also made longer-term follow-up plans involving GenderMag.
- *Facets and Gender*: Perhaps most important, all of the teams ultimately realized that for software to be gender-inclusive, it needs to support a range of facet values, not just facet values matching their own personal styles. In essence, they realized that gender inclusiveness is not about sorting people into gender bins, it is about pluralism.

We still have work to do on improving GenderMag and understanding the roles it can play in real-world software organizations, but a beta is available [15]. Ultimately, we hope that GenderMag will enable problem-solvers of any gender finally to have a virtual, yet concretely actionable way to express Ashcraft and DuBow's point [2]:

"*Women in tech do not generally need extra help, but the current environment in which they work does need help.*"

ACKNOWLEDGMENTS

This work was supported in part by NSF #1240957, 1314384, and 1528061.

REFERENCES

1. Tamara Adlin and John Pruitt. 2010. The Essential Persona Lifecycle: Your Guide to Building and Using Personas. Morgan Kaufmann/Elsevier.
2. Catherine Ashcraft and Wendy Dubow The Tricky (And Necessary) Business of Being A Male Advocate For Gender Equality. 2015. Retrieved September 24th, 2015 from http://www.fastcompany.com/3046555/strong-female-lead/the-tricky-and-necessary-business-of-being-a-male-advocate-for-gender-equ
3. Albert Bandura. 1986. Social Foundations of Thought and Action. Prentice Hall.
4. Shaowen Bardzell. 2010. Feminist HCI: Taking stock and outlining an agenda for design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), 1301-1310. http://doi.acm.org/10.1145/1753326.1753521
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. Laura Beckwith, Derek Inman, Kyle Rector, and Margaret Burnett. 2007. On to the real world: Gender and self-efficacy in Excel. In Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '07). 119-126. http://dx.doi.org/10.1109/VLHCC.2007.42
8. Michelle A. Borkin, Chelsea S. Yeh, Madelaine Boyd, Peter Macko, Krzysztof Z. Gajos, Margo Seltzer, and Hanspeter Pfister. 2013. Evaluation of filesystem provenance visualization tools. IEEE Transactions on Visualization and Computer Graphics 19, 12 (December 2013), 2476-2485. http://dx.doi.org/10.1109/TVCG.2013.155
9. Adam Bradley, Cayley MacArthur, Mark Hancock, and Sheelagh Carpendale. 2015. Gendered or neutral? Considering the language of HCI, In Proceedings of the 41st Graphics Interface Conference (GI '15), 163-170.
10. Leah Buechley and Michael Eisenberg. 2008. The LilyPad Arduino: Toward wearable engineering for everyone. IEEE Pervasive Computing 7, 2 (April 2008), 12-15. http://dx.doi.org/10.1109/MPRV.2008.38
11. Leah Buechley and Benjamin Mako Hill. 2010. LilyPad in the Wild: How hardware's long tail is supporting new engineering and design communities. In Proceedings of the 8th ACM Conference on Designing Interactive systems (DIS '10), 199-207. http://doi.acm.org/10.1145/1858171.1858206
12. Margaret Burnett, Laura Beckwith, Susan Wiedenbeck, Scott D. Fleming, Jill Cao, Thomas H. Park, Valentia Grigoreanu, and Kyle Rector. 2011. Gender pluralism in problem-solving software. Interacting with Computers, 23, 5 (September 2011), 450–460. http://dx.doi.org/10.1016/j.intcom.2011.06.004
13. Margaret Burnett, Scott D. Fleming, Shamsi Iqbal, Gina Venolia, Vidya Rajaram, Umer Farooq, Valentina Grigoreanu, and Mary Czerwinski. 2010. Gender differences and programming environments: across programming populations. In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '10). 10 pages. http://doi.acm.org/10.1145/1852786.1852824
14. Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, William Jernigan, GenderMag: A method for evaluating software's gender inclusiveness, Interacting with Computers (to appear). DOI 10.1093/iwc/iwv046
15. Margaret Burnett, Simone Stumpf, Laura Beckwith, and Anicia Peters, The GenderMag Kit: How to Use the GenderMag Method to Find Inclusiveness Issues through a Gender Lens, Retrieved Oct. 10, 2015, from http://eusesconsortium.org/gender.
16. Judith Butler. 1999. Gender Trouble: Feminism and the Subversion of Identity. Routledge.
17. Patricia Cafferata and Alice M. Tybout. 1989. Gender Differences in Information Processing: A Selectivity Interpretation, Cognitive and Affective Responses to Advertising. Lexington Books.
18. 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 the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '10), 149-156. http://dx.doi.org/10.1109/VLHCC.2010.29
19. Justine Cassell. 2002. Genderizing HCI. In The Handbook of Human-Computer Interaction, Julie A. Jacko and Andrew Sears (Eds.). L. Erlbaum Associates Inc., Hillsadale, NJ, USA 402-411.
20. Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G. Terveen. 2014. Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW '14), 674-686. http://doi.acm.org/10.1145/2531602.2531660

21. Gary Charness and Uri Gneezy. 2012. Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization 83, 1 (June 2012), 50–58.
22. Constantinos Coursaris, Sarah Swierenga, and Ethan Watrall. 2008. An empirical investigation of color temperature and gender effects on web aesthetics. Journal of Usability Studies 3, 3 (May 2008), 103-117.
23. 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 (April 2009), 522–550.
24. 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, 521–535.
25. Executive Office of the President. 2013. Women and Girls in Science, Technology, Engineering, and Math (STEM). Retrieved September 24th, 2015 from www.whitehouse.gov/ostp/women
26. 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. http://doi.acm.org/10.1145/2207676.2208572
27. Valentina Grigoreanu, Margaret Burnett, and George Robertson. 2010. A strategy-centric approach to the design of end-user debugging tools. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), 713-722. http://doi.acm.org/10.1145/1753326.1753431
28. Valentina Grigoreanu, Margaret Burnett, Susan Wiedenbeck, Jill Cao, Kyle Rector, and Irwin Kwan. 2012. End-user debugging strategies: A sensemaking perspective. Transactions on Computer-Human Interaction 19, 1, 5 (May 2012).
29. Valentina Grigoreanu, Jill Cao, Todd Kulesza, Christopher Bogart, Kyle Rector, Margaret Burnett, and Susan Wiedenbeck. 2008. Can feature design reduce the gender gap in end-user software development environments? In Proceedings of the 2008 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '08), 149-156. http://dx.doi.org/10.1109/VLHCC.2008.4639077
30. Jonathan Grudin. 2006. Why personas work: The psychological evidence. In John Pruitt and Tamara Adlin, The Persona LifeCycle: Keeping People in Mind Throughout Product Design, Morgan Kaufmann Publishers.
31. Jonas Hallstrom, Helene Elvstrand, and Kristina Hellberg. 2015. Gender and technology in free play in Swedish early childhood education, Int J. Technology and Design Education, 25, 137-149. DOI 10.1007/s10798-014-9274-z.
32. Kathleen Hartzel. 2003. How self-efficacy and gender issues affect software adoption and use. Commun. ACM 46, 9 (September 2003), 167–171. http://doi.acm.org/10.1145/903893.903933
33. 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 CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06), 875-880. http://doi.acm.org/10.1145/1125451.1125622
34. Ann H. Huffman, Jason Whetten, and 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.
35. William Jernigan, Amber Horvath, Michael Lee, Margaret Burnett, Taylor Cuilty, Sandeep Kuttal, Anicia Peters, Irwin Kwan, Faezeh Bahmani, and Andrew Ko. 2015. A principled evaluation for a principled Idea Garden. In Proceedings of the 2015 IEEE Symposium on Visual Languages and Human-Centric Computing, October 2015. 8 pages.
36. Yasmin B. Kafai and Quinn Burke. 2014. Beyond game design for broadening participation: Building new clubhouses of computing for girls. In Proceedings of Gender and IT Appropriation. Science and Practice on Dialogue – Forum for Interdisciplinary Exchange (Gender IT '14). 8 pages.
37. Caitlin Kelleher. 2009. Barriers to programming engagement. Advances in Gender and Education 1, 5-10.
38. Caitlin Kelleher, Randy Pausch, and Sara Kiesler. 2007. Storytelling Alice motivates middle school girls to learn computer programming. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 1455-1464.
39. Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naïve Bayes text classification. ACM Trans. Interact. Intell. Syst. 1, 1, Article 2 (October 2011), 31 pages. http://doi.acm.org/10.1145/2030365.2030367
40. Michael Lee, Faezeh Bahmani, Irwin Kwan, Jilian Laferte, Polina Charters, Amber Horvath, Fanny Luor, Jill Cao, Catherine Law, Mihcael Bethwetherick, Sheridan Long, Margaret Burnett, and Andrew Ko. 2014. Principles of a debugging-first puzzle game for computing education. In Proceedings of the 2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '14), 57-64.

41. Michael Lee and Andrew Ko. 2011. Personifying programming tool feedback improves novice programmers' learning. In Proceedings of the 7th International Workshop on Computing Education Research (ICER '11), 109-116. http://doi.acm.org/10.1145/2016911.2016934
42. Thomas Mahatody, Mouldi Sagar, and Christopher Kolski. 2010. State of the art on the Cognitive Walkthrough method, its variants and evolutions, International Journal HCI 26, 8 (July 2010), 741-785.
43. Jane Margolis and Allan Fisher. 2003. Unlocking the Clubhouse: Women in Computing. MIT Press.
44. 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
45. 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.
46. Joan Meyers-Levy and Durairaj Maheswaran. 1991. Exploring differences in males' and females' processing strategies. Journal Consumer Research 18, 63–70.
47. National Center for Women & IT. 2014. By the numbers, Version 02282014. Retrieved September 24th, 2015 from http://www.ncwit.org/sites/default/files/resources/btn_02282014web.pdf
48. Lene Nielsen and Kira Storgaard Hansen. 2014. Personas is applicable: A study on the use of personas in Denmark. In Proceedings of the 32nd Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '14), 1665-1674. http://doi.acm.org/10.1145/2556288.2557080
49. 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. NSF ITWF & ITR/EWF Principal Investigator Conference, 65-69.
50. Piazza Blog. 2015. STEM Confidence Gap. Retrieved September 24th, 2015 from http://blog.piazza.com/stem-confidence-gap/
51. 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
52. 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 (June 2010), 397-428.
53. Jennifer Ann Rode. 2008. An ethnographic examination of the relationship of gender & end-user programming. Ph.D Thesis. University of California Irvine, Irvine, CA.
54. 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
55. Mary Beth Rosson, Hansa Sinha, and Tisha Edor. 2010. Design planning in end-user web development: gender, feature exploration, and feelings of success. In Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '10), 141-148. http://dx.doi.org/10.1109/VLHCC.2010.28
56. Per Runeson, Martin Host, Austen Rainer, and Bjorn Regnell. 2012. Case Study Research in Software Engineering: Guidelines and Examples. Wiley Publishing.
57. Steven John Simon. 2001. The impact of culture and gender on web sites: An empirical study. The Data Base for Advances in Information Systems 32, 1 (Winter 2001), 18-37.
58. Anil Singh, Vikram Bhadauria, Anurag Jain, and Anil Gurung. 2013. Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets. Computers in Human Behavior 29, 3 (May 2013), 739–746.
59. 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
60. Neeraja Subrahmaniyan, Laura Beckwith, Valentina Grigoreanu, Margaret Burnett, Susan Wiedenbeck, Vaishnavi Narayanan, Karin Bucht, Russell Drummond, and Xiaoli Fern. 2008. Testing vs. code inspection vs. ... what else? Male and female end users' debugging strategies. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08), ACM, 617-626.
61. Desney S. Tan, Mary Czerwinski, and George Robertson. 2003. Women go with the (optical) flow. In Proceedings of the SIGCHI Conference on Human Factos in Computing Systems (CHI '03), 209-215. http://doi.acm.org/10.1145/642611.642649
62. 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 and Decision Making 15, 263-290.
63. Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. 1994. The cognitive walkthrough method: A practitioner's guide. In Usability Inspection

Methods, Jakob Nielsen and Robert L. Mack (Eds.). John Wiley, NY. 105-140

64. Gayna Williams. 2014. Are you sure your software is gender-neutral? Interactions 21, 1 (January 2014), 36–39.

65. Robert K. Yin. 2009. Case Study Research: Design and Methods (Fourth Edition). Sage Publications.