Boldifications samples

Margaret Burnett, for CS/ECE 507, Fall 2016

These are excerpts of early drafts of this paper: Finding Gender-Inclusiveness Software Issues with GenderMag: A Field Investigation, Margaret Burnett et al., ACM CHI’16.

This is 2016 #6 from <http://web.engr.oregonstate.edu/~burnett/reprints.html>

# Background and Related Work

\*\*research shows gender differences (this parag might need rewording)\*\*

There is extensive research over the past decade establishing that the individual differences in how people use software features aimed at supporting problem-solving tends to cluster by gender, and that many such features are inadvertently designed around approaches favored more by males than by females [Beckwith et al. 2005; Beckwith et al. 2006; Beckwith et al. 2007; Borkin et al. 2013; Bruckman et al. 2002; Burnett et al. 2010; Burnett et al. 2011; Chang et al. 2014; Cao et al. 2010b; Dunwell et al. 2014; Grigoreanu et al. 2006; Grigoreanu et al. 2008; Grigoreanu et al. 2009; Grigoreanu et al. 2012; Hartzel 2003; Hou et al. 2006; Kulesza et al. 2011; Lee et al. 2014; Meyers-Levy and Maheswaran 1991; Piazza Blog 2015; Rode 2008; Rosson et al. 2010; Subrahmaniyan et al. 2008; Szafir and Mutlu 2012; Tan et al. 2003]. Studies like these are evidence of the need for a process to help software developers improve their products’ gender inclusiveness.

\*\*The gender differences are across platforms\*\* These findings of gender differences span numerous domains in which software provides a platform for problem solving. Spreadsheets are one example: one series of studies showed that females using spreadsheets employed new features for testing and debugging formulas significantly less than males [Beckwith et al. 2005; Beckwith et al. 2006; Beckwith et al. 2007], and several other spreadsheet studies have reported gender differences as well (e.g., [Grigoreanu et al. 2008, Grigoreanu et al. 2012, Hou et al. 2006, Surahmaniyan et al. 2008]. Gender differences have also been reported in people’s use of visualization features [Borkin et al. 2013; Tan et al. 2003], in the way they interact with online classwork platforms [Piazza Blog 2015], in the ways they design and debug web automations [Cao et al. 2010a; Rosson et al. 2010], in the types of barriers males vs. females run into when trying to figure out how to guide intelligent agents to better behavior [Kulesza et al. 2011], and in software developers’ use of a variety of programming tools [Burnett et al. 2010].

\*\*There are 2 ways people are trying to do something about it\*\*

Research to address the need for improving the gender-inclusiveness of software falls generally into four categories. The first is to develop demonstration software products, and the second 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.

\*\*The first way is…\*\*

Some examples of the “demonstration software” approach develop software that aim to appeal to females, without a goal of also appealing to males, since plenty of software products already target males. Kafir and Burke term this kind of approach “building new clubhouses” [Kafir and Burke 2014], as a counterpoint to the well-known work by Margolis and Fisher about “unlocking” the (boys-only) computing clubhouse [Margolis and Fisher 2003]. (From a feminist perspective, building new clubhouses can be said to reject the “let me in” implication that male values are the normative values, to be aspired to by everyone [Kafir and Burke 2014].) Examples of tech products designed specifically for females include Goldiblox and Storytelling Alice [Kelleher et al. 2007]. 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 extends the Alice programming language and environment by supporting storytelling through programming, which increased middle-school girls’ learning of computer programming [Kelleher et al. 2007].

\*\*Other example of demonstration products are like this....\*\*

Other examples of demonstration products aim to appeal to both males and females, often by removing barriers or enhancing features that tend to particularly affect one of the genders. This kind of approach has a pluralism theme, i.e., the idea that most individuals do not fit cleanly into a single gender bin, and that removing barriers can help everyone regardless of the gender bin with which they identify. Examples of this type of product include LilyPad [Buechley and Eisenberg 2008, Buechley and Hill 2010], Gidget [Lee et al. 2014], and StratCel [Grigoreanu et al. 2010]. LilyPad is a “maker” product with the same functionality as Arduino, but for wearable computing projects. Thus, it extends beyond activities that have been focused on programming and engineering to include sewing, textiles, and crafts, thereby adding inclusiveness by leveraging traditions of girls’ play worlds and the cultural practices of women’s crafting communities [Kafai and Burke 2014]. Gidget is a debugging game for novice programmers. Its gender inclusiveness comes from innovating certain programming environment characteristics. For example, it portrays the computer as fallible, personifies error messages, and presents explanatory help in ways that are compatible with both females’ tendency toward comprehensive information processing and males’ tendencies toward depth-first information processing [Lee et al. 2014, Lee and Ko 2011, Jernigan et al. 2015]. StratCel [Grigoreanu et al. 2010] is an add-on for Excel that supports problem-solving strategies statistically associated with females in addition to those statistically associated with males [Subrahmaniyan et al. 2008].

\*\*Empirical evidence: these products do help\*\*

Empirical research substantiates that many of these projects aimed at increasing the gender inclusiveness of software succeed at attaining this goal. The above projects’ effectiveness were all validated empirically. As additional examples, Tan et al. showed that displaying optical flow cues benefited both females and males in virtual world navigation [Tan et al. 2003]; Grigoreanu et al. showed how changes to spreadsheet features relating to confidence and feature support reduced gender gaps while improving everyone’s attitudes and feature usage [Grigoreanu et al. 2008]; and Jernigan et al. showed how a tool designed for a diversity of individual styles and situations enabled both female and male novice programmers to program as well with very little in-person help as participants in a previous year in which extensive in-person help was given [Jernigan et al. 2015].

\*\*demonstration projects like these are important but…\*\*

Demonstration projects like these are important for demonstrating not only that greater inclusiveness is possible in problem-solving, but also for providing examplars 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 very expensive, requiring extensive research, prototype building, and empirical work just to show that the underlying concept is valid. The developers of such projects than incur even more expense if their vision is to reach real users (as opposed to being just for research and proof-of-concept experiments).

\*\*scalability is the advantage of the “build a method” approach.\*\*

 Scalability is the advantage of the other kind of approach, namely developing new methods/practices for avoiding or identifying gender-inclusiveness issues in software. Methods can scale more readily than individual projects simply because they are devised for the sole purpose of influencing other projects. Design and decision-making processes are common 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 along the way. A few of these recommendations are including female customers in feature/product “pitches” to management, normalizing product decision vote-counting so that females’ representation is the same as males’, convening women-only early adopter panels, and so on [Williams 2014].

\*\*gendermag is different from these other methods in the following way, and this paper is different from an evaluation of them too…

\*\* The GenderMag method used in this paper’s field studies is not focused on adding gender inclusiveness to existing design processes as per the Williams set, but rather is a new evaluation process that specifically evaluates a software product’s gender inclusiveness. It gender-specializes the Cognitive Walkthrough (CW) technique [Wharton et al. 1994] and integrates it tightly with gender-specialized personas, as we detail in the next section. Although use of CWs has been empirically investigated before (e.g., [Blandford et al. 2008]), as has use of personas (e.g., [Matthews et al. 2014, Nielsen and Storgaard Hansen 2014, Turner and Turner 2011]), this paper is the first to empirically investigate the use of GenderMag’s unique specialization of these devices in the field. \*\*MMB: this last parag needs a bit of finetuning, but the essence is ok\*\*

# From: Empirical Methodology section

\*\*Participating organizations contacted us when they were interested in using the method and do a field study\*\*

We conducted xx\*\* field studies at various organizations to test the GenderMag method. We posted an announcement of GenderMag’s availability in kit form on our website [\*\*url], and also gave talks about it at interested conferences and meetings. In all cases, we explicitly included an invitation to interested parties to participate in field studies in order to refine the kit. Organizations contacted us if they were interested in participating in our field studies.

\*\*The roles of coordinator and researchers were…\*\*

A contact person at the organization acted as coordinator and was responsible for assembling a team, scheduling the study, selecting the task and prototype, and compiling the task list, ideal action sequences and tweaking the selected personas. A researcher assisted in compiling the task list, when necessary.

\*\*The field study procedures were largely similar to each other.\*\*

The evaluation team consisting of at least three persons used the method to walk through their task lists and prototype. The team appointed a facilitator and recorder, but all team members served as evaluators. The researcher introduced the method and evaluation procedures by stepping through the kit. At the end of the introduction, the facilitator took over and steered the session. At the end of the walkthrough, the team debriefed among themselves on the most important finding and the researcher then asked a set of semi-structured questions. The questions were structured around: 1) whether they found issues and might fix them, 2) what worked well and not so well with the method, the likelihood of using the method again, and 3) what worked well/not well with the personas and similarity to their users. Each session was video recorded. A researcher conducted follow-up interviews with some members of the teams two weeks later.

\*\*Although the procedures were similar, materials differed, because the method was refined iteratively\*\*

We kept the procedures for all field studies largely similar and followed the method’s step-by-step guidelines, but the materials differed slightly since the kit was refined iteratively after each field study. This affected the study outcomes for each field study.

\*\*The participants: The first two groups were related to each other\*\* The first two participating organizations were related to each other (contractor and client) and they both evaluated the same transportation system where operators capture accidents or traffic incidents on roads based on calls from citizens. One of the groups was the contracted developer of the other group.

\*\*The participants (cont.): The other two groups …\*\*

The remaining two participating…

# Results RQ2-RQ3: Company W

\*\*At first, Company W hadn’t decided to do it, but an internal champion got them on board. 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. In fact, 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.

\*\*Company W ended up doing both sessions with different teams and only one persona\*\* 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 50% overlap between the evaluators in the two sessions; with 3 people (two males and one female) attended both sessions. A Company W researcher observed the sessions along with two external researchers.

\*\*Company W conducted two sessions internally\*\*

Both sessions were deemed very useful by the participants, so they decided to conduct two more sessions. Although these two sessions were internal, without the 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. \*\*MMB to self: Tim is something new Company W brings to this paper, so be sure to make something of this\*\*

\*\*The teams evaluated a mobile app that was already in use\*\*

The teams evaluated a mobile app that was newly released, and still actively being improved. Each team member had a mobile phone with a freshly installed app for evaluation. The mobile devices had different versions of the \*\*Android?\*\* operating system and different levels of cellular service (some devices had no cellular service at all) to mimic a real world situation. \*\*MMB: parag needs finetuning, but okish

## RQ2: Company W & the Persona & facets

\*\*Abby was surprising to the developers\*\*

When working with the Abby persona, developers were 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."

**\*\*She made them think of their users in new ways, and as real people.\*\***

As W4m’s story shows, the Abby persona caused the teams to think of their users in new ways. Abby helped the team focus on designing software for their users, rather than for themselves or some ideal user:

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.

\*\*There was some danger of stereotyping, however.\*\*

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 after Session #2 that she had overheard male participants from Session #1 talking in the hallways, saying things like:

 “Today I learned that women are this and women are like that.”

\*\*To fix this, we developed two new personas\*\*To help guard against gender stereotyping, we had already begun developing two additional personas, Patrick and Patricia, described in Section \*\*personas. During follow-up interviews after Session #4, when the personas were complete, we talked about Patrick and Patricia and how they were different from Abby and Tim.

We expect Abby and Tim to be the most heavily used personas when time is a limiting factor, because they represent the extreme ends of the facet spectrum. \*\*CH to MMB: do we have evidence that providing the extreme examples covers a larger amount of issues? That might be a more compelling argument.\*\* When time isn’t a limiting factor, we expect Patrick and Patricia will be used as well, to cover a wider range of possible design issues:

W6m: I'll probably like Abby and Tim more than I'll like Pat and Pat because they stretch versus ... go towards ... the middle.

W4m (follow-up interview): “So, if you have only a limited amount of time for this kind of process then 2 [personas] is as much as you do. So, if you have a lot of time and you are really focused on this, and this is the biggest problem with our app, then definitely 3 or 4 would be to make sure you cover your bases.

## RQ3: Validity, Utility and Effectiveness

**\*\* Team identified different issues identified for Abby and Tim\*\***Company W identified 20 issues during the four sessions. Seven of these were gender-inclusiveness issues (Table \*\*w Issues): six gender-inclusiveness issues for Abby, one gender-inclusiveness issue for Tim. In addition, the team identified 13 other usability issues not tied to the facets. Figure \*\*w-facets shows that all five facets were considered when looking for issues.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User actions with gender-inclusiveness issues  | Session#:Persona | Notes | Data sources | Facets that found |
| Form | video | debrief | eval |
| 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 \*\*w Issues\*\*: The Company W team found 7 gender inclusiveness issues. 13 additional usability issues were not tied to gender inclusiveness. The debrief data source was a phone interview about sessions #3 and #4, which were performed without researchers present.



**Figure \*\*w-facets: 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.)**

**\*\*Using the method identified missing metrics\*\***

Interestingly, the method led the team to notice the absence of a performance metric to track their development progress:

W6m: What I have discovered outside of this exercise [the actual GenderMag session] is it's led to a metrics conversation about <application name>.

**\*\*Are the issues effective in leading to desirable behavior/attitude changes/what are the effects of the software change?\*\***

An important measure of the method’s utility is continued usage. Some of the Company W team started running sessions on other applications, in addition to the two internal sessions they ran after we left:

W3m: Well, I did some ... other testing [GenderMag sessions]...

**\*\*Team gets training to reuse the method in the future and used the method twice internally\*\*** In fact, W3m decided to give his entire testing team experience with the method, and to start using it more broadly:

W3m: We ...will probably 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".

**\*\*Fixing: distrib'd responsibility again (all that trouble <person> had trying to convince the designer).** As stated previously,one measure of the method’s utility is follow-through in terms of fixing issues found during the session. However, 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 make the gender-inclusiveness issue fixes that Company W's team envisioned. Company W’s team felt so strongly about the importance of three of the issues, they eventually convinced the original designers to fix them.

W6m summed up the team’s feelings about GenderMag:

W6m: [GenderMag] has already infected us.

…