From Gender Biases to Gender-Inclusive Design: An Empirical Investigation

Mihaela Vorvoreanu[1,2], **Lingyi Zhang**[2], **Yun-Han Huang**[2], **Claudia Hilderbrand**[3], **Zoe Steine-Hanson**[3], **Margaret Burnett**[3]

[1]Microsoft Research
Redmond, Washington, USA
Mihaela.Vorvoreanu@microsoft.com

[2]Purdue University
West Lafayette, Indiana, USA
zhan1540,huan1025@purdue.edu

[3]Oregon State University
Corvallis, Oregon, USA
minic,steinehz,burnett@eecs.orst.edu

ABSTRACT

In recent years, research has revealed gender biases in numerous software products. But although some researchers have found ways to improve gender participation in specific software projects, general methods focus mainly on *detecting* gender biases—not *fixing* them. To help fill this gap, we investigated whether the GenderMag bias detection method can lead directly to designs with fewer gender biases. In our 3-step investigation, two HCI researchers analyzed an industrial software product using GenderMag; we derived design changes to the product using the biases they found; and ran an empirical study of participants using the original product versus the new version. The results showed that using the method in this way did improve the software's inclusiveness: women succeeded more often in the new version than in the original; men's success rates improved too; and the gender gap entirely disappeared.

CCS CONCEPTS

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

KEYWORDS

Gender-inclusive software; gender biases; GenderMag

ACM Reference format:
Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, Margaret Burnett. 2019. From Gender Biases to Gender-Inclusive Design: An Empirical Investigation. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 14 pages. https://doi.org/10.1145/3290605.3300283

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300283

1 INTRODUCTION

Awareness of biases in software has increased in recent years. Some biases are embedded in data and algorithms that power a software product, out of the sight of regular users (e.g., [8, 30]). Others are more noticeable, permeating user experiences [7, 13, 58] and even the ways computing professionals think about their users [9]. Software biases are of critical importance: they can affect organizational and community health and even individuals' job prospects and outcomes in the criminal justice system [40].

Fortunately, the computing professions are awakening to problems like these, and new conferences and conversations are emerging about the need to address biases (e.g., [24, 34, 63]). Also emerging are research methods to help detect such biases (e.g., [25, 56]).

GenderMag [15] is one such method. GenderMag is a method in the cognitive walkthrough family that enables identification of gender biases in interfaces and in the workflows these interfaces make possible. By identifying issues that, for example, fail to support cognitive styles frequently associated with a gender, GenderMag can be used to detect gender biases in interfaces and workflows.

However, merely detecting gender biases does not *fix* a product. Thus, in this paper we used GenderMag to investigate whether and how biases *detected* can be leveraged into design *solutions* that make *software more usable and inclusive.*

Toward this end, we conducted a 3-step investigation using high-fidelity prototypes of Microsoft Academic (academic.microsoft.com), an academic search engine:

(Step 1) Two HCI researchers conducted a GenderMag analysis to identify gender biases in the software.

(Step 2) We used the results of Step 1 to directly derive redesigns for several aspects of the interface that Step 1 identified as problematic.

(Step 3) We ran a qualitative empirical study with 20 participants, to investigate whether the redesigned prototype was more usable and inclusive than the prototype we started with.

This paper makes the following research contributions: First, it is the first to empirically evaluate whether a gender bias detection method (GenderMag) can lead to design fixes that actually improve software's usability and inclusiveness. Second, it provides examples of GenderMag-inspired redesigns that improved usability and inclusiveness, and those that did not. Third, it empirically evaluates the accuracy of the detected issues and biases, in terms of false negatives (incompleteness) and false positives (identifying an issue when none was present). Finally, it illuminates the nuanced relationship between individual differences in cognitive styles vs. gender biases in software interfaces.

2 BACKGROUND AND RELATED WORK

2.1 Related Work

Prior research has established that individual differences in how people use software features tend to cluster by gender. Research spanning numerous software domains has shown that today's software mostly supports only the styles favored by men—for example in spreadsheets [4, 5, 6, 26, 28, 32, 60], visualization systems [7, 62], online classwork platforms [50], web and home automation [19, 51, 52, 55], intelligent agents and robots [41, 58], and programming tools [12, 47].

Research to improve gender inclusiveness in software falls generally into two categories. The first category contains *methods and practices* for avoiding or identifying gender inclusiveness issues in software. The advantage of methods is scalability: if the methods are 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 women in the decision-making processes that shape software [65]. The GenderMag method [14, 15, 31, 47] used in this paper is in this category.

The second category, *demonstration software projects* concretely improves software's fit to different genders. Some of these software projects aim to appeal specifically to women or girls, such as Goldiblox and Storytelling Alice [37]. 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 (men-only) computing clubhouse [45].

Other demonstration projects aim to support both men and women, often by removing barriers or enhancing features. This kind of approach has a pluralism theme such as advocated by Bardzell [3], i.e., the idea that most individuals do not fit neatly into a single gender bin [18], and that removing barriers 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 women's tendencies toward comprehensive information processing and men's tendencies toward selective information processing [35, 42, 43]. Another example is LilyPad [10, 11], a "maker" product with the same functionality as Arduino, but for wearable computing projects. It combines the "build it" tradition of boys' play worlds with craft traditions like sewing and textiles of girls' play worlds [36]. Another example is StratCel [27], an add-on for Excel that supports problem-solving strategies statistically associated with women in addition to those statistically associated with men [60].

Although gender inclusiveness in software is a relatively young area, we believe the time has come for a third category, one that bridges the first two. In this third category, HCI professionals would use a method from the first category to avoid or remove gender biases in software as in the second category. Researchers would then empirically evaluate the method's effectiveness by comparing users' experiences or performance with the original version of the software product vs. the new post-method version. The result brings the generality from the first category together with the concreteness and practical impacts from the second category.

This paper is in this third category: it evaluates the ability of a gender bias detection method (GenderMag) to improve the actual gender inclusiveness of a software product. This paper is, to our knowledge, the first to empirically investigate any gender inclusiveness method's effects on an industrial software product.

2.2 What is GenderMag?

Because this study investigates the effectiveness of the GenderMag method [14, 15, 31, 47], we briefly summarize it.

GenderMag (gendermag.org), short for "Gender Inclusiveness Magnifier" [15], integrates a specialized cognitive walkthrough with research-based personas that

Table 1. A summary of the facet values for each persona. This paper uses Abby (orange throughout this paper) and Tim (blue).

	Abby	Pat(ricia) & Pat(rick)	Tim
Motivations for using technology	Wants what the technology can accomplish.	Wants what the technology can accomplish.	Technology is a source of fun.
Computer Self-Efficacy (confidence) about using unfamiliar technology	Low compared to peer group.	Medium.	High compared to peer group.
Attitude towards Risk when using technology	Risk-averse.	Risk-averse.	Risk-tolerant.
Information Processing Styles for gathering information to solve problems	Comprehensive.	Comprehensive.	Selective.
Learning Styles for learning new technology	Process-oriented learner.	Learns by tinkering; tinkers reflectively.	Learns by tinkering (sometimes to excess).

capture individual differences in how people problem solve and use software features—differences that statistically cluster by gender. GenderMag has been used to detect gender biases in several commercial and open source software products (e.g., [12, 14, 22, 47, 57]).

The GenderMag method rests on five problem-solving facets (Table 1), and brings the facets to life with three multi-personas—"Abby", "Pat(ricia)/Pat(rick)", and "Tim". They are *multi*-personas in that their backgrounds, photos, job titles, education, etc., are customizable. The facets, however, are fixed. Abby's facet values (shown concretely in Figure 1) are more frequently seen in women than other genders, and Tim's facet values (see supplemental document) are more frequently seen in men than other genders. The Pats' (identical) facet values emphasize that differences relevant to inclusiveness lie not in a person's gender identity, but in the facet values themselves [31].

The personas and facets are integrated with a specialized cognitive walkthrough (CW) [64]. The specialization adds the cognitive facets above, to detect biases that occur at statistically different rates by gender. In the Methodology section, we detail the GenderMag analysis process in the context of the current investigation.

3 METHODOLOGY

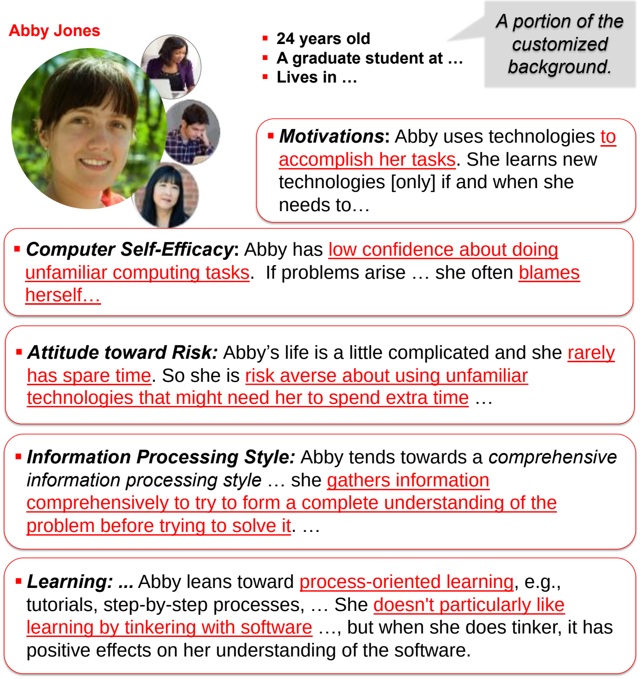
Our investigation followed a three-step process. In Step 1, two HCI researchers conducted a GenderMag analysis on a high-fidelity paper prototype of the product to identify gender bias issues. In Step 2, we redesigned the prototype to address the majority of those issues. In Step 3, we conducted a between-subject Wizard of Oz [39] study to empirically compare the two prototypes.

We conducted the investigation in the context of Microsoft Academic, an academic search engine that identifies, categorizes, and retrieves scholarly publications [49]. During the investigation, one of the authors was working with the Microsoft Academic team on Microsoft Academic's interface.

3.1 Step 1: GenderMag Analysis

For Step 1, we followed the GenderMag procedures described in [15, 17], which begin by choosing persona(s). We chose the Abby and Tim personas [17], because they represent opposite ends of the GenderMag facet value ranges, and customized their ages, occupations, hobbies, etc. to fit the product's target users and the study's context, a university in the U.S. (Figure 1; see the supplemental document for full details).

GenderMag walkthroughs use scenarios. For our scenarios (Table 2), we used the five main Microsoft Academic use cases, which capture the main functionality the product aims to support—searching, sorting, filtering, and citing papers; and claiming papers under an author profile. In total, our scenarios covered most of the Microsoft Academic interface. For a GenderMag analysis, scenarios must be broken down into subgoals. Subgoals provide



- 24 years old
- A graduate student at …
- Lives in …

- ***Motivations*:** Abby uses technologies to accomplish her tasks. She learns new technologies [only] if and when she needs to…
- ***Computer Self-Efficacy*:** Abby has low confidence about doing unfamiliar computing tasks. If problems arise … she often blames herself…
- ***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 …
- ***Information Processing Style:*** Abby tends towards a *comprehensive information processing style* … she gathers information comprehensively to try to form a complete understanding of the problem before trying to solve it. …
- ***Learning: …*** Abby leans toward process-oriented learning, e.g., tutorials, step-by-step processes, … She doesn't particularly like learning by tinkering with software …, but when she does tinker, it has positive effects on her understanding of the software.

Figure 1. Key portions of the Abby persona used. The complete personas are in the supplemental document.

digestible "abstract" sequences through the scenario (Table 2). A subgoal has concrete action sequences a product's owner/designer envisions users carrying out. GenderMag analysis asks whether users would want and be able to perform these actions.

Given these personas, scenarios, and subgoals, two analysts conducted the GenderMag walkthroughs on Microsoft Academic using the procedures in [15, 17]. They did one such walkthrough of each scenario from the perspective of the Abby persona, and separate walkthroughs of the same scenarios with the Tim persona. Both analysts were HCI researchers, but neither had ever done a GenderMag analysis before. The walkthrough process answered the following questions about each subgoal and about the actions a user would need to perform to accomplish those subgoals (italics added to show key differences from standard CWs):

SubgoalQ: Will <*Abby/Tim*> have formed this subgoal as a step to their overall goal? (Yes/no/maybe, why, *what facets are involved in your answer*).
ActionQ1: Will <*Abby/Tim*> know what to do at this step? (Yes/no/maybe, why, *what facets ...).*
ActionQ2: If <*Abby/Tim*> does the right thing, will s/he know s/he did the right thing and is making progress toward their goal? (Yes/no/maybe, why, *what facets....*).

Using this process, the analysts identified ten issues in total, and we followed up on six. The other four issues were not empirically viable: one issue did not yield itself to testing with a paper prototype and three were not part of core tasks in Microsoft Academic, so we could not motivate them well in our empirical setting.

Table 2. The scenarios and subgoals discussed in this paper, and the issues that arose in each. (The complete list of scenarios can be found in the supplemental document.)

Scenario	Subgoals	Led to
Scenario #1-TopInstitution: In <field>, look for the top-ranked institution, and cite the most cited paper it published within 15 years.	Subgoal #3: Find papers about <field> published in the top institution.	Issues 1 & 2
	Subgoal #5: Cite the paper.	Issue 3
Scenario #2-TopAuthor: Look for the author who published the most publications in <field> from University of Oxford and go to his/her author page.	Subgoal #3: Find the author who published the most publications, go to their page.	Issue 4
Scenario #4-ClaimPaper: Ask <persona> to pretend s/he is <name>, and declare the author as her/himself.	Subgoal #2: Start claiming, choose the right author.	Issue 5
	Subgoal #3: Review and submit claim.	Issue 6

3.2 Step 2: Facet-Driven Redesign

The outcomes of the analyses identify not only *where* an issue can arise, but *why* that issue might arise—what specific problem-solving facet(s) are not supported in the design. Thus, we used the facets identified in Step 1 as starting points for creating design remedies.

For example, the analysts identified an issue for Abby's self-efficacy facet that might affect her interaction with filters. To correct this issue, the team focused on Abby's low self-efficacy (Figure 1), and decided to add feedback when using the filters so that people like Abby would be less likely to believe a problem has arisen in applying the filters. Once the team agreed on redesign solutions, we generated a new high-fidelity paper prototype. For the rest of this paper, we refer to the original system's high-fidelity prototype as the "Original" version, and the high-fidelity prototype after redesign as the "post-GenderMag" version.

3.3 Step 3: Qualitative Empirical Study

In Step 3 we qualitatively analyzed 20 participants' use of the Original vs. post-GenderMag high-fidelity prototypes, with 10 participants per condition for each scenario (Table 2).

We recruited participants at a large U.S. university by advertising a 5-minute screening questionnaire via listservs and flyers. We targeted research-oriented people likely to do academic searches, namely faculty members and graduate students. The questionnaire used (with permission) a subset of an existing Microsoft survey [16] in which respondents self-assess their GenderMag facets with 9-point Likert questions (enumerated in the supplemental document).

Using respondents' questionnaire responses, we used criterion sampling to select 20 participants. The criteria were: use academic search engines regularly but had never used Microsoft Academic; provide a reasonable gender balance; and provide a diversity of facet values.

Eleven of the resulting participants identified as women and nine as men. Their ages ranged from 24 to 63 (median 27.5). Participants came from a variety of research disciplines: Computer graphics technology (3 women + 3 men), life sciences (2+1), engineering/math (1+2), interaction design (2+0), computer science (0+2), other technologies (1+1), education (1+0), and unknown (1+0). Participants in each discipline were divided approximately evenly across the two treatment groups, except that both of the interaction design participants ended up in the post-GenderMag treatment. Participants spanned a reasonably wide range of values for each of the five GenderMag facets. Figure 2 (Left) shows the number of facets each participant self-assessed on the Abby side of the grand medians (i.e., scoring closer to Abby than the grand median of that facet) and on the Tim side of the grand medians. Note that only three participants (1 Abby, 2 Tims) were "pure" Abbys or Tims; the other 17 participants had mixes of Abby and Tim facets.

We used a between-subject design, balancing participants in the Original vs. post-GenderMag treatments by gender, academic discipline, and GenderMag facets (Figure 2 (Right)). We conducted each session one participant at a time with a facilitator and an observer. The participants' tasks were to perform the same scenarios as used in the GenderMag analysis (Table 2). We collected audio recordings and observation notes from the sessions, and qualitatively analyzed participants' data one facet at a time. Since every participant has a value for each of the 5 facets, this means that each facet's data came from all 11 women and all 9 men.

3.4 Bringing the Three Steps Together

The GenderMag analysis of the Original prototype identified ten issues, and we investigated six, as explained above. For these six, we analyzed results strictly by GenderMag facet—not gender. When all analysis by facet was complete, we then reconsidered the results from a gender perspective, which we defer until Section 9.

4 RESULTS: OVERALL TASK SUCCESS

We begin by answering the central research question: if designers redesign around GenderMag-detected biases, can this actually make a more usable and inclusive piece of software?

As the results in Table 3 show, the usability answer is "yes": for every issue except Issue 3, fewer post-GenderMag participants than Original participants experienced action failures. (We counted an action failure whenever a minute passed without the participant figuring out the action/subgoal, or the participant asked for help; we then showed the participant what to do so they could continue with the rest of the task.) In total, post-GenderMag participants failed less than half as often as Original participants did.

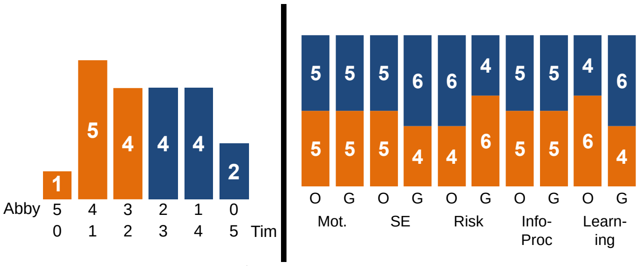


Figure 2. Participants' facet distributions. (Orange: more Abby facet values than Tim. Blue: more Tim than Abby.) (Left) Y-axis: # of participants (out of 20) who had each Abby vs. Tim facet combination (X-axis). (Right): The 20 participants' facet value distributions in the Original vs. post-GenderMag treatments, for each facet in Table 1.

Next, we consider the inclusiveness question at the level of facet values. If inclusiveness issues are found *for a particular facet value* using GenderMag, and then changes to fix the issues are made, does the software become more inclusive for people with *that facet value*? Table 4 answers this question.

The table lists the issues by the facet results of Step 1 (GenderMag analysis). A checkmark indicates that the

Table 3. Number of participants with action failures in the Original (n=10) vs. post-GenderMag (n=10) treatments.

Issues	Original	post-GenderMag
Issues 1 & 2	2	0
Issue 3	4	4
Issue 4	6	2
Issues 5 & 6	1	0
Total	13	6
Mean/median/mode	1.3/1/2	0.6/1/1

Table 4. Each row shows effects (Step 3) of the redesigns (Step 2), for each facet value identified in Step 1 for Abby (left symbols, in orange) or Tim (right symbols, in blue):

✓: in Step 2, redesigned the post-GenderMag version for that facet value.
+: post-GenderMag better: (fewer action failures than Original).
−: post-GenderMag treatment worse.
=: no effect.
***: no effect because no errors in either version.**

	Facets	Redesigned For		Effects	
Issues 1&2	Motivations	✓		+	+
	Self-Efficacy	✓		+	+
	Risk	✓		*	+
	Info-Process	✓	✓	+	*
	Learning	✓	✓	+	+
Issue 3	Motivations	✓		+	−
	Risk	✓		−	+
	Learning	✓		=	=
Issue 4	Risk	✓	✓	+	+
	Info-Process		✓	+	+
	Learning		✓	+	+
Issues 5&6	Self-Efficacy	✓	✓	+	*
	Risk	✓		*	+
	Info-Process	✓	✓	+	*
	Learning	✓		+	=

analysts found an issue for that facet value in Step1 and made design changes to the prototype (Step2) to improve support for that facet value. The right column indicates the empirical effects of that change on task performance.

For example, the orange checkmark in the top row shows that, for Issues 1&2, we redesigned the prototype to improve support for Abby's motivations facet value (task-oriented). The orange + shows a positive empirical effect from the redesign on Abbys, with post-GenderMag task-oriented participants failing less than Original task-oriented participants. The blue + in this row shows that, although the redesign was not targeting Tim's motivations (no blue checkmark in "Redesigned"), the post-GenderMag participants with Tim's motivations failed less than in the Original treatment—a nice side-effect of this particular redesign.

Other outcomes in Table 4 include no effect from the redesigns (=), post-GenderMag participants experiencing more failures than Original participants (−), and false positives: issues the analysts found for a facet that did not arise with any participant (*) which also could have no effect. Since the table includes only issues the analysts decided were issues, false negatives are not shown. (We will return to false positives and false negatives in later sections.)

To summarize Table 4, in a few cases the redesigns had no effects (3 for Abby's facets, 5 for Tim's) or had negative effects (1 Abby, 1 Tim). However, for most of the changes, the redesigns led to improvements for the facet values being targeted, and sometimes for the opposite values of those facets. In total, the redesign had 11 positive effects for Abby's facets and 9 for Tim's.

5 RESULTS: ISSUES 1 & 2: AN IN-DEPTH LOOK

What issues did the GenderMag analysis identify, what design remedies did we derive, and how did those remedies impact usability for different problem-solving facets? In this section we address these questions in depth.

5.1 Issues 1 & 2 GenderMag Analysis (Step 1)

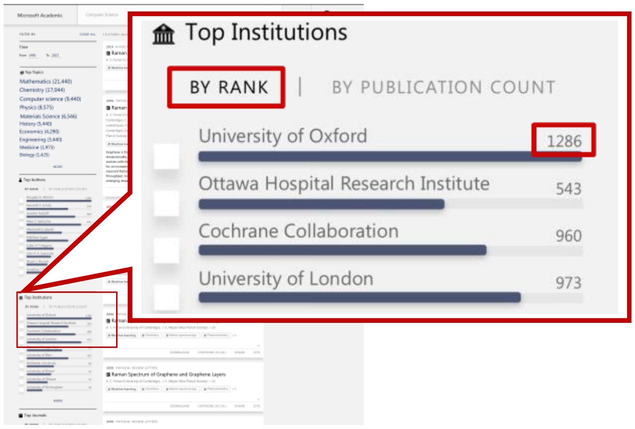
The analysts identified Issues 1 and 2 during Scenario 1's Subgoal 3 (recall Table 2). By Subgoal 3, the persona (Abby or Tim) had already searched for papers in the field of Cognition, and had sorted the results by number of citations. To then form Subgoal 3, Abby or Tim would have to understand that it is possible to filter the results down to papers published by authors at only the top institution.

To identify issues with this subgoal, the analysts considered each of Abby's and Tim's facet values. From these facets, the analysts decided that even forming this subgoal could be problematic for both Abby and Tim (Issue 1). To then perform the action, Abby/Tim would have to locate the filter, and the analysts decided this too could be a problem for both (Issue 2). Together, these issues involved all five of Abby's facets and two of Tim's, as we detail next.

Motivations (Abby): Some people are motivated to use technology mostly for the tasks it enables them to accomplish (like Abby), whereas others are motivated by their enjoyment of the technology for its own sake (like Tim) [12, 13, 29, 32, 37, 38, 46]. Because the Abby persona is task-oriented, she prefers using features that seem familiar and views learning new features as a detour from what she is trying to accomplish. The analysts' notes below show how they applied the motivations facet:

> Analyst Notes: "The filters on the left side are something Abby is not familiar with (motivations)… looks more like a data visualization. She might not know the institutions are ranked…" (Analyst is referring to the screen in Figure 3.)

Information Processing Style (Abby and Tim): The analysts also pointed out that Abby's information processing style was not well-served. People with Abby's comprehensive information processing style try to gather fairly complete information before proceeding with problem solving [1, 28, 48, 53]. The analysts noted that, although in some ways this style would help Abby, she would want more information about the feature than was given in the interface, such as whether "by rank" is a filter at all, how it works, and what the numbers are communicating. In contrast, people with Tim's selective information processing style try to follow the first promising information, then potentially backtrack. The



University of Oxford	1286
Ottawa Hospital Research Institute	543
Cochrane Collaboration	960
University of London	973

Figure 3. Screenshot of the Original prototype with a call-out of the section relevant to the Subgoal 3. The screen is about double the length shown, with more filtering options and results. The red boxes emphasize where Abby would have run into issues, namely with clicking on "By Rank", and the counts that look like a data visualization instead of a filter.

analysts also decided that this preference would also be encouraged by Tim's learning style, which is tinkering-oriented. Figure 4 illustrates their application to Tim.

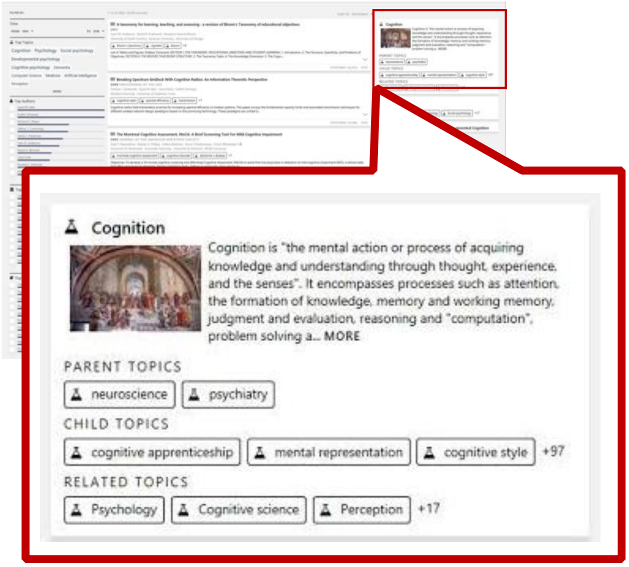
> Analyst Notes: "Since Tim leans towards a selective information processing and likes ... tinkering (learning style), he will click the 'cognition' topic..." (wrong action)
>
> Analyst Notes: "As Abby tends to process the information comprehensively, she will find the filter... [But because] the filter doesn't look like [a filter]..."

Attitude towards Risk (Abby): Abby is relatively risk-averse with new technologies; she does not like to waste time on a feature that might not do what she wants if she cannot predict its effects [21, 23]. In applying that facet here, the analysts decided that Abby's risk aversion would add to Abby's desire for more information before proceeding.

> Analyst Notes (continued from above Abby example): "...[But because] the filter doesn't look like [a filter]... she will be cautious [about] the filters."

Self-efficacy (Abby): Self-efficacy is a person's confidence about succeeding in a specific task [2]; *computer* self-efficacy is their confidence in performing computing tasks. Computer self-efficacy levels affect people's behavior with technology, such as which features they choose to use and how willing they are to persist with hard-to-use features [12, 13, 33, 50, 59]. People like Abby, who have lower computer self-efficacy than their peers, tend to believe that technology issues or the absence of feedback indicates that they are doing something incorrectly. The analysts decided that the lack of feedback from the action would interact with Abby's relatively low self-efficacy:

> Analyst Notes: "After Abby applies the filter, the result on the right will be refreshed. But there [is] no obvious [feedback] showing how the results linked to the filter ... As a result, Abby might not be ... sure that ... the filters she selected were applied (computer self-efficacy)."



Cognition

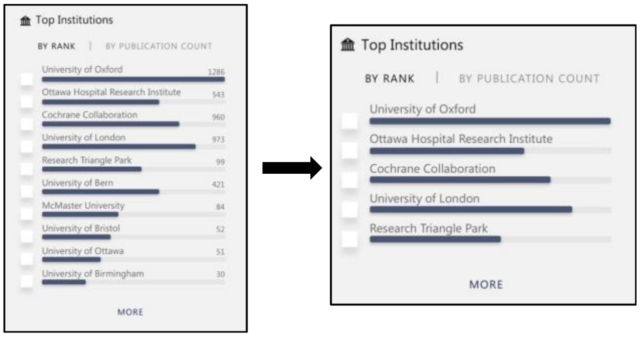
Cognition is "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses". It encompasses processes such as attention, the formation of knowledge, memory and working memory, judgment and evaluation, reasoning and "computation", problem solving a... MORE

Figure 4. Screenshot of the Original prototype. The analysts decided that Tim would notice the Cognition topic box (in the call out) before the filters (left), and click it right away.

Learning style (Abby and Tim): The analysts also noted that Abby's learning style came into play. Abby prefers to learn processes first, rather than tinkering with details first [12, 19, 20, 54], but the system's lack of information or hints about process does not support this style of learning. Tim's learning style (tinkering) could also lead him astray, exacerbating the influence of his selective information processing style, as discussed above.

5.2 Issues 1 & 2 Design Remedies (Step 2)

To derive design remedies for both Abby-like and Tim-like users, we started with the facets that identified the issues. As Figure 5 shows, we addressed two of Abby's facets—motivations and risk—by making the filters' appearance look more like filters (removing the number of publications on the right side of each bar, which drew attention away from the filter-action checkboxes). To help with Tim's most implicated facet, his selective information processing, we shortened the list (top 5 instead of top 10), so that Tim-like information processors could notice the institution filtering option without scrolling. Since the other facet that arose for Tim (a tinkering learning style) was magnifying the problems caused by his selective information processing style, we decided that resolving his most implicated facet would be sufficient to handle both. As Figure 6 shows, for the feedback problems particularly pertinent to Abby's computer self-efficacy, we added to the top of the filtered search results a list of filters that had been applied, as a way of providing confirmation to users that they had (or had not) accomplished their filtering goals.



Top Institutions

BY RANK | BY PUBLICATION COUNT

□ University of Oxford	1286
□ Ottawa Hospital Research Institute	543
□ Cochrane Collaboration	960
□ University of London	973
□ Research Triangle Park	99
□ University of Bern	421
□ McMaster University	84
□ University of Bristol	52
□ University of Ottawa	51
□ University of Birmingham	30

MORE

Top Institutions

BY RANK | BY PUBLICATION COUNT

□ University of Oxford
□ Ottawa Hospital Research Institute
□ Cochrane Collaboration
□ University of London
□ Research Triangle Park

MORE

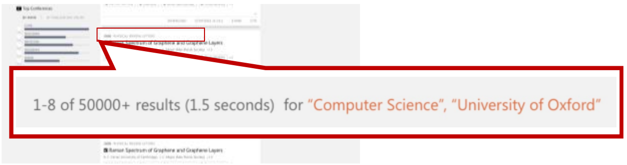
Figure 5. Issues 1 & 2 filtering redesign. (Left) Original: List of institutions with publication counts for each. (Right) post-GenderMag: Shorter list of institutions and removed the publication counts that drew attention away from the checkbox actionability.

5.3 Issues 1 & 2 Empirical Results (Step 3)

Original and post-GenderMag participants performed the same scenario as in the GenderMag analysis: selecting papers by authors affiliated with the top institution in a particular field (here, Computer Science). The Original participants' performance allowed us to evaluate the accuracy of the GenderMag analysis, and the comparison of Original with post-GenderMag enabled us to evaluate the effectiveness of the redesign.

As Figure 7 shows, two of the 10 Original participants, O2 and O9, had one or more action failures, as defined in the first Results section. (We will refer to specific Original participants as O1-O10, and post-GenderMag participants as G1-G10.) Participants O2 and O9 differed markedly in their problem-solving facets: participants' initial questionnaire responses showed that Participant O2 was an Abby4-Tim1 participant (four Abby facets, one Tim facet), and Participant O9 was an Abby1-Tim4 participant. Together, those two participants covered four of Abby's facets and four of Tim's facets.

Four of the facets that the analysts predicted for Abby and one of those predicted for Tim in the GenderMag analysis were present in the participants who experienced action failures. Even some of the successful participants had problems noticing that the "Top Institution" filtering option existed, as with this successful (no action failures) Original participant:



1-8 of 50000+ results (1.5 seconds) for "Computer Science", "University of Oxford"

Figure 6. Issues 1 & 2 post-GenderMag feedback redesign. We added feedback of what filters were applied.

	Original					post-GenderMag					
Participant	M	SE	R	IP	L	M	SE	R	IP	L	Participant
O1	-	-	-	-	-	-	-	-	-	-	G1
O2	●	●	■	●	●	-	-	-	-	-	G2
O3	-	-	-	-	-	-	-	-	-	-	G3
O4	-	-	-	-	-	-	-	-	-	-	G4
O5	-	-	-	-	-	-	-	-	-	-	G5
O6	-	-	-	-	-	-	-	-	-	-	G6
O7	-	-	-	-	-	-	-	-	-	-	G7
O8	-	-	-	-	-	-	-	-	-	-	G8
O9	■	■	■	●	■	-	-	-	-	-	G9
O10	-	-	-	-	-	-	-	-	-	-	G10

Figure 7. Results of Issues 1&2 by participant in order of most Abby facet values to least. Facet values of the 2 Original treatment participants with action failures are shown. ●: participant's facet is an Abby value; ■: participant's facet is a Tim value, -: no action failures.

Participant O4 (tracing along the filters with her finger): "How would I find the top institution?"
Observer notes: "The filter [panel] was too long ... took [O4] awhile to find the institution filter."

Participant O4's facet values help to demonstrate how certain facets helped participants solve problems. Participant O4 was an Abby3-Tim2 participant. Her comprehensive information processing style is evident in the above. That facet value served her well here, leading her to process enough of this option-dense screen to find the option she needed to succeed.

Ultimately, as Figure 7 shows, the redesign was successful—none of the post-GenderMag participants failed the task.

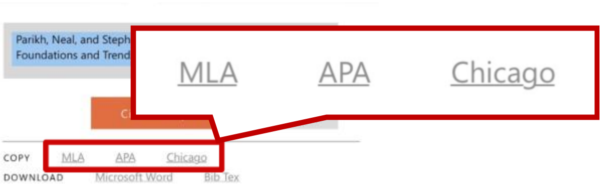
Observer note: "[G2] found the top institution filter easily."
Observer note: "[G10] found the filter easily and went for it directly."
Participant G3: "Top institution, yes this one…click here."

6 RESULTS: ISSUES 3 & 4: OPPOSITE TALES OF INCOMPLETENESS

Like other cognitive walkthroughs, a GenderMag analysis can miss some issues. In fact, errors of omission (false negatives) are common in cognitive walkthrough methods, with research reporting rates between 30% and 70%, depending on the analysts' expertise [44]. Our investigation's data likewise suggest that our analysts (who were not GenderMag experts) missed some issues in their GenderMag analysis. For example, the six "side-effect" improvements shown earlier in Table 4, show improvements for facets where no problem for those facets' values had been detected. Issues 3 & 4 illustrate opposite ways such false negatives played out in our data.

6.1 Issue 3: Not only Abby: And an unhappy ending

Issue 3 involved an action in Scenario 1's Subgoal 5 (Table 2), at which point the persona needed to copy a citation. The analysts identified ambiguous copy widgets and a lack of feedback as problematic for Abby's motivations, risk, and learning facets. To fix this issue in the post-GenderMag version, we added the orange feedback box and underlined the copy widgets as shown in Figure 8.



Parikh, Neal, and Steph
Foundations and Trend

MLA APA Chicago

COPY MLA APA Chicago
DOWNLOAD Microsoft Word Bib Tex

Figure 8. Issue 3's widget & post-GenderMag feedback redesign. We added the orange feedback box and underlined the copy widgets (see call out) to the dialog box.

Unfortunately, as the Original participants showed, the analysts missed some usability issues that turned out to affect participants' task performance in both the Original and post-GenderMag treatments. Specifically, the icon (not shown) to reveal the contents of Figure 8 was not noticeable and not obviously clickable in both versions. (The paper prototype probably accentuated these problems because of its lack of interactivity and of mouse-over actions.)

The post-GenderMag version did show a few signs of having improved Issue 3's feedback problem, such as with G10's appreciation of the feedback:

> Participant G10: "I will ... click on APA..." [A box says 'copied'] "Yeah I think it's copied now"

However, Original vs. post-GenderMag task performance did not reduce the number of action failures: the same number of participants had action failures in both conditions (Figure 9).

6.2 Issue 4: Not mainly Tim: But a happy ending

Issue 4 started out as almost a mirror image of Issue 3, with the analysts thinking Issue 4 mainly applied to Tim and erroneously missing most of Abby's facets. However, this issue turned out quite differently than Issue 3.

O1	●	●	●	●	●	●	●	●	■	●	G1
O2	●	●	■	●	●	●	●	●	■	●	G2
O3	-	-	-	-	-	-	-	-	-	-	G3
O4	●	■	●	●	■	-	-	-	-	-	G4
O5	-	-	-	-	-	-	-	-	-	-	G5
O6	-	-	-	-	-	■	■	●	●	■	G6
O7	-	-	-	-	-	-	-	-	-	-	G7
O8	■	●	■	■	■	■	■	■	●	■	G8
O9	-	-	-	-	-	-	-	-	-	-	G9
O10	-	-	-	-	-	-	-	-	-	-	G10
Participant	M	SE	R	IP	L	M	SE	R	IP	L	Participant
Original						post-GenderMag					

Figure 9. Results of Issue 3 by participant in order of most Abby facet values to least. Facets of the 4 Original and 4 post-GenderMag participants with action failures ●: participant's facet is an Abby value; ■: participant's facet is a Tim value, -: participant had no action failure.

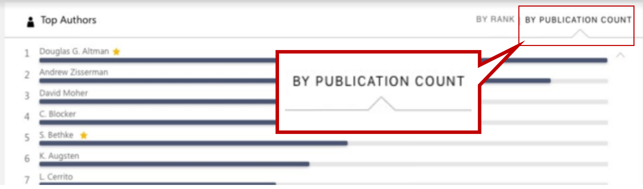

Top Authors
BY RANK | BY PUBLICATION COUNT
1 Douglas G. Altman ★
2 Andrew Zisserman
3 David Moher
4 C. Blocker
5 S. Bethke ★
6 K. Augsten
7 L. Cerrito
BY PUBLICATION COUNT

Figure 10. Issue 4 indicator redesign. We added a horizontal line with a carat pointing to the current sort (boxed in red, pointing to "By Publication Count").

Issue 4 arose during Scenario 2's Subgoal 3 (Table 2), at which point the persona would be trying to get to the author publishing the most papers in a particular field. The analysts thought the issue would affect three of Tim's facets (information processing, risk, and learning), but only one of Abby's (risk):

> Analyst notes: "Tim leans toward depth first ... (information processing style), ... constructing his own understanding (learning), and <doesn't> mind taking risks (...risk). Thus, he might click on the first author in the top author card without noticing the <toggle>."

To address this problem, the redesign—which focused on Tim—improved the visibility of the sorting toggle, by mapping it to the information in the tile (Figure 10).

Empirical data confirmed that the issue was problematic in the Original prototype: it arose for 6 of the 10 Original participants (Figure 11). These 6 Original participants also revealed that this issue would affect not only people with Tim facets, but also all of the Abby facets. For example, O4's interest in a clearer definition of "rank" was consistent with their comprehensive information processing style:

> Participant O4: "[confusion] So, by rank, what would that mean? I don't know... some kind of professional ranking?"

Despite the fact that the redesign focused mostly on Tim, the post-GenderMag version for Issue 4 improved task performance for all 5 Abby-like facets, and for 4 Tim-like facets (Figure 11).

7 RESULTS ISSUES 5 & 6: SOME SOLUTIONS ARE SIMPLE

Issues 5 and 6 were related and the same redesign addressed both. For these issues, the persona was in Scenario 4 to claim papers for him/herself. The analysts decided the

O1	●	●	●	●	●	-	-	-	-	-	G1
O2	●	●	■	●	●	-	-	-	-	-	G2
O3	-	-	-	-	-	-	-	-	-	-	G3
O4	●	■	●	●	■	-	-	-	-	-	G4
O5	●	■	■	●	●	-	-	-	-	-	G5
O6	-	-	-	-	-	-	-	-	-	-	G6
O7	-	-	-	-	-	■	●	■	●	■	G7
O8	■	●	■	■	■	-	-	-	-	-	G8
O9	-	-	-	-	-	-	-	-	-	-	G9
O10	■	■	■	■	■	■	■	■	■	■	G10
Participant	M	SE	R	IP	L	M	SE	R	IP	L	Participant
Original						post-GenderMag					

Figure 11. Results of Issue 4 by participant in order of most Abby facet values to least. Facets of the 6 Original and 2 post-GenderMag participants with action failures. ●: participant's facet is an Abby value; ■: participant's facet is a Tim value, -: participant had no action failures.

process was unclear, and also noticed problems with feedback:

Analyst notes: "After selecting the right [action], there's no feedback and instructions on what Abby should do next (learning)."

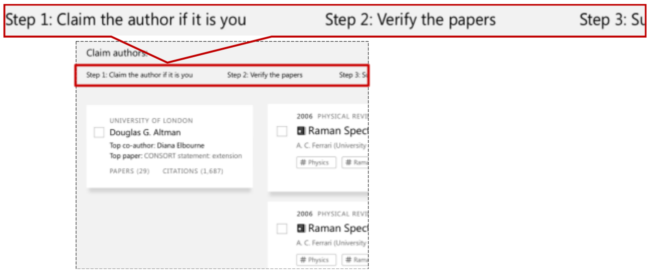
The analysts identified these as issues for several Abby and Tim facets. The redesign was straightforward: we added an explicit list of the steps to follow (Figure 12) and improved the feedback (not shown in the figure).

Empirically, only one Original participant failed the task, so the issue did not turn out to be as difficult as the analysts had feared. No post-GenderMag participants failed the task.

8 GENDERMAG'S ACCURACY: WERE THE ISSUES REAL?

Accuracy can be affected by false negatives (GenderMag missing some issues) or false positives (GenderMag reporting an issue where none exists). Since we have already discussed false negatives (Results Section 6), we turn now to false positives: if GenderMag identified an issue, was it really an issue?

As Table 5 shows in its rightmost column, the answer is yes. In this study, not a single false positive occurred. (However, in a few cases, the issues occurred for different facets than expected, which we discuss in the next section.) This high true positive rate is consistent with a prior GenderMag study reporting a true positive rate of 96% (false positive rate 4%) [15], and with a survey of cognitive walkthroughs reporting false positive rates of 5% or less [44].



Step 1: Claim the author if it is you

Step 2: Verify the papers

Step 3: Su

Claim authors:
Step 1: Claim the author if it is you
Step 2: Verify the papers
Step 3: S
UNIVERSITY OF LONDON
Douglas G. Altman
Top co-author: Diana Elbourne
Top paper: CONSORT statement: extension
PAPERS (29)
CITATIONS (1,687)
2006 PHYSICAL REVI
Raman Spect
A. C. Ferrari (University
# Physics
# Ram
2006 PHYSICAL REVI
Raman Spect
A. C. Ferrari (University
# Physics
# Ram

Figure 12. Issues 5 & 6 post-GenderMag redesign: Portion of the screen with the process enumerated (shown in red box).

Recall that the two HCI analysts doing the GenderMag analyses had never conducted a GenderMag session before. Thus, we can regard the GenderMag analyses as a conscientious first use of GenderMag by HCI researchers, not as an expert analysis. GenderMag's high true positive rate occurred *despite* their lack of experience with the method.

9 PERSPECTIVES: IT ISN'T ABOUT GENDER AND IT IS ABOUT GENDER

So, is GenderMag about gender? The previous sections say "no": the keys to more inclusive software lie not in someone's gender, but in the facet values themselves. As this answer makes clear, GenderMag can be used to find and fix inclusiveness issues without ever speaking of gender.

Here, we consider whether the answer is also "yes," that it *is also* about gender. First, we view this question from a data perspective: did the following hold true in our data?

if the facets analysts identified using GenderMag, actually did run into problems,

and the facets were (directly or indirectly) related to participants' genders,

then fixing issues tied with those facets should reduce gender gaps in participants' use of the product.

If: Whether the data met the "if" condition above is answered by Table 5. As Table 5 shows (second column from right), 75% of the issues' facet biases the analysts identified using GenderMag arose in Original participants who had those facet values. As for the remaining 25% the analysts predicted, at least one participant also had action failures in each of these (right-most column), but those participants had different facet values than those the

Table 5: GenderMag true-positive rate: 100% of the problems the analysts identified using GenderMag were experienced by participants in the empirical study. Even the facet values were reasonably well-matched: 75% of the facet values the GenderMag analysis had pointed to were experienced by participants with those facet values in the empirical study.

	GenderMag analysis: Facet Biases		Empirical: Facet biases validated		Empirical: Total facet biases validated	Empirical: Total facet problems validated
	Abby	Tim	Abby	Tim	Abby + Tim totals	Anybody (Abby or Tim)
Issues 1+2	5 facets	2 facets	4/5 facets	1/2 facets	5/7 facets (71%)	5/5 problems (100%)
Issue 3	3 facets	0	3/3 facets	-n/a-	3/3 facets (100%)	3/3 problems (100%)
Issue 4	1 facet	3 facets	1/1 facet	3/3 facets	4/4 facets (100%)	3/3 problems (100%)
Issues 5+6	4 facets	2 facets	3/4 facets	0/2 facets	3/6 facets (50%)	4/4 problems (100%)
				Totals	15/20 facet biases validated (75%)	15/15 problems validated (100%)

analysts had predicted. Thus, GenderMag's facet identification was reasonably accurate.

And: How these facets relate to the gender identifications of our participants is shown in Figure 13. Although there were wide differences among individual participants within gender, the distribution of the women skewed toward the "Abby" side of the facets, and the distribution of men skewed toward the "Tim" side of the facets.

Then: Given the accuracy of the facets identified and the relationships of facets to gender, the "then" is that fixing the issues identified with the facets should reduce gender gaps in men's and women's success with the software. And as Figure 14 shows, this is exactly what happened. In the Original prototype, women had twice as many failures as the men; but in the post-GenderMag version, the gender gap disappeared, and task performance improved for both the participating genders.

Moreover, note the lack of binary-ness in the data in Figure 13. This makes clear that software cannot be made "better" by having a "pink" version (supporting the 5 Abby facets in one version) and a "blue" version (supporting the 5 Tim facets in the other). In our data, such solutions would fit only three of the 20 participants: the one participant with all 5 Abby facets, and the 2 participants with all 5 Tim facets. Instead, as Figure 14 shows, to improve software's usability across genders, software needs inclusivity across the cognitive diversity that arises not only between/among different genders, but also within them.

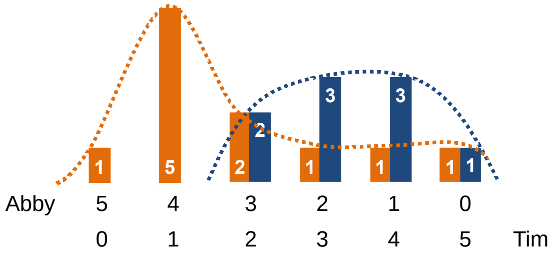


Figure 13. Y-axis: Counts of the 20 men and women participants by their facet values. (Same as Figure 2 but broken out by gender.) Orange: women, blue: men. X-axis: Abby=Abby Facets, Tim=Tim Facets. Example: the left bar says that the only participant with 5 Abby facets (0 Tim facets) was a woman; the right pair of bars says that one man and one woman had 5 Tim facets (0 Abby facets).

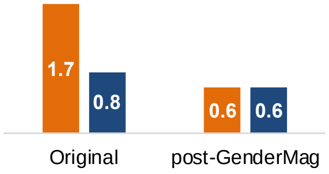


Figure 14. Average number of action failures per person by gender identification (orange: women, blue: men). In the Original version, women's action failure rates were over twice as high as men's; with the post-GenderMag redesign, all failure rates went down, and the gender gap disappeared.

These results and perspectives are consistent with feminist HCI theoretical thought. As Bardzell explained in her landmark paper at CHI'10 [3], third-wave feminism emphasizes attention to individual differences within genders, emphasizing pluralism over universality. GenderMag's core of individual differences in cognitive style embraces this notion of pluralism. This core of individual differences also supports non-binary notions of gender identification as per prominent feminist literature (e.g., [18, 52, 61]). Figure 13 helps to illustrate this point, as does the Pat persona (recall Table 1). Finally, as Rode and Poole illustrate through the lenses of several feminist theories [52], a person's construction and expression of their gender identity is often intimately intertwined with ways they feel about and interact with technology. The possible implications of how de-biasing software's user experiences might interact with how individuals construct and evolve their own gender identities in our technological world, is an interesting open question.

10 CONCLUDING REMARKS

In this paper, we have presented the first investigation into whether and how an HCI method to *detect* gender biases in software can *generate* more gender-inclusive designs. The method our study investigated was GenderMag; it was beyond the scope of this study to compare GenderMag with other usability inspection methods, or to generalize our results to other contexts. Also, other factors such as income, age, race, gender identification factors such as those discussed in [52], and algorithmic biases can influence software equity, and GenderMag does not address these.

That said, in our setting, GenderMag alone did lead to more inclusive and more usable designs, as our results show. Specifically:

- *Improved designs in total*: Participants in the post-GenderMag condition, who used the design changes derived from the GenderMag analyses, were more successful on almost every individual task than Original participants. In total, the post-GenderMag participants failed less than half as often as participants in the Original version.
- *Accurately found issues*: In 100% of the cases where analysts found an issue using GenderMag, the issue happened to one or more participants in the Original condition. Further, in most of the cases, the participants

experiencing the issues had the facet values analysts predicted.

- *The facets pointed to the fixes*: The method we used to achieve these results was to use the facets identified in Step 1 as starting points for creating the fixes (Step 2). The results suggest that this GenderMag-based sequence offers an effective, pre-user-study method to pinpoint and fix issues affecting gender- and cognitive-inclusivity.
- *From gender biases to gender-inclusiveness*: After the design changes, the Original version's gender gap in participants' failure rates entirely disappeared.

The results also illuminate the nuanced relationships between individual cognitive styles and gender. As such, the results have several direct implications for gender-inclusive design. First, designing for cognitive diversity improves software's gender inclusiveness. Second, it is neither necessary nor desirable to devise two (or more) gender-labeled versions of the same software to serve different genders.

Finally, many gender biases in software are cognitive biases, and many cognitive biases in software are gender biases. Thus, software designs that better support cognitive diversity also better support gender diversity—and improve the software for everyone.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1528061. We thank the Microsoft Academic team and visual designer Jin Kim. Without their hard work, the redesign would not have been possible.

REFERENCES

[1] Manon Arcand, Jacques Nantel. 2012. Uncovering the nature of information processing of men and women online: The comparison of two models using the think-aloud method. J. Theor. Appl. Elec. Comm. Res. 7, 2: 106-120.

[2] Albert Bandura. 1986. Social Foundations of Thought and Action. Prentice Hall.

[3] 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

[4] 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

[5] Laura Beckwith, Margaret M. Burnett, Valentina Grigoreanu, and Susan Wiedenbeck. 2006. Gender HCI: What about the software? IEEE Computer, 39(11), 97–101.

[6] 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

[7] 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

[8] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. Ethics and Information Technology 15(3), 209-227.

[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, 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.

[13] Margaret Burnett, Laura Beckwith, Susan Wiedenbeck, Scott Fleming, Jill Cao, Thomas Park, Valentina Grigoreanu, Kyle Rector. 2011. Gender pluralism in problem-solving software. Interacting with Computers 23, 5: 432–460.

[14] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender inclusiveness software issues with GenderMag: A field investigation. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2586-2598 Doi10.1145/2858036.2858274.

[15] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness, Interacting with Computers, 28(6), 760–787. DOI 10.1093/iwc/iwv046

[16] Margaret Burnett, Robin Counts, Ronette Lawrence, and Hannah Hanson. 2017. Gender HCI and Microsoft: Highlights from a longitudinal study. IEEE VLHCC, 139-143.

[17] Margaret Burnett, Simone Stumpf, Laura Beckwith, and Anicia Peters. 2018. The GenderMag Kit: How to Use the GenderMag Method to Find Inclusiveness Issues through a Gender Lens, http://gendermag.org, June 28, 2018.

[18] Judith Butler. 1999. Gender Trouble: Feminism and the Subversion of Identity. Routledge.

[19] 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

[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). ACM, New York, NY, USA, 674-686. https://doi.org/10.1145/2531602.2531660

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

[22] Sally J. Cunningham, Annika Hinze, and David Nicols. 2016. "Supporting gender-neutral digital library creation A case study using the GenderMag Toolkit," *Knowledge, Information, and Data in An Open Access Society*, vol. 10075, pp. 45-50, 2016.

[23] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jurgen Schupp, Gert G. Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. J. European Economic Association 9, 3: 522–550.

[24] FAT*, 2018. ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*). Retrieved Sept. 11, 2018 from https://fatconference.org/

[25] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In Proceedings ACM Foundations of Software Engineering, 498-510.

[26] 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

[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] 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.

[30] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr." In ACM CSCW, 1914-1933.

[31] Charles Hill, Maren Haag, Alannah Oleson, Chris Mendez, Nicola Marsden, Anita Sarma, and Margaret Burnett. 2017. Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? In Proc. CHI Conference on Human Factors in Computing Systems, 6658-6671. doi 10.1145/3025453.3025609.

[32] 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

[33] 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.

[34] IFIP News, 2018. ACM President Cherri Pancake Highlights Importance of Diversity in ICT, July 2018. Retrieved Sept. 11, 2018 from https://www.ifipnews.org/acm-president-cherri-pancake-highlights-importance-diversity-ict/

[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, 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.

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

[39] John F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems 2, 1: 26–41. https://doi.org/10.1145/357417.357420

[40] Keith Kirkpatrick. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? Communications of the ACM 59(10). 16-17.

[41] 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

[42] Michael Lee and Andrew Ko. 2011. Personifying programming tool feedback improves novice programmers' learning. In Proceedings of the ACM International Workshop on Computing Education Research (ACM ICER '11), 109-116. http://doi.acm.org/10.1145/2016911.2016934

[43] 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 2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '14), 57-64.

[44] 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.

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

[46] Jane Margolis, Jean J. Ryoo, Cueponcaxochitl D. M. Sandoval, Clifford Lee, Joanna Goode, and Gail Chapman. 2012. Beyond access: Broadening participation in high school computer science. ACM Inroads 3(4), 72-78.

[47] Christopher Mendez, Hema Susmita Padala, Zoe Steine-Hanson, Claudia Hilderbrand, Amber Horvath, Charles Hill, Logan Simpson, Nupoor Patil, Anita Sarma, and Margaret Burnett. 2018. Open Source barriers to entry, revisited: A sociotechnical perspective. ACM/IEEE ICSE 2018, 1004-1015.

[48] Joan Meyers-Levy and Barbara Loken. 2015. Revisiting gender differences: What we know and what lies ahead. J. Consumer Psych. 25, 1: 129-149.

[49] Microsoft Research, Microsoft Academic (product description). Retrieved Sept. 4, 2018 from https://www.microsoft.com/en-us/research/project/academic/

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

[51] Jennifer Ann Rode. 2008. An ethnographic examination of the relationship of gender & end-user programming. Ph.D Thesis. University of California Irvine, Irvine, CA.

[52] Jennifer A. Rode and Erika Shehan Poole. 2018. Putting the gender back in digital housekeeping. In Proceedings of the 4th Gender & IT Conference, Heilbronn, Germany, May 2018 (GenderIT'18), ACM, 12 pages. https://doi.org/10.1145/3196839.3196845

[53] 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.

[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 SIGCHI Conference on Human Factors in Computing Systems (CHI '09). ACM, New York, NY, USA, 419-422. DOI: https://doi.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] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In Data and Discrimination: Converting Critical Concerns into Productive Inquiry (a preconference at the 64th Annual Meeting of the International Communication Association).

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

[58] Dilruba Showkat, and Cindy Grimm. 2018. Identifying gender differences in information processing style, self-efficacy, and tinkering for robot tele-operation, International Conference on Ubiquitous Robots (UR), Hawaii, USA, 2018.

[59] 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.

[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 SIGCHI Conference on Human Factors in Computing Systems (CHI '08), ACM, 617-626.

[61] Lucy Suchman. Agencies in technology design: Feminist reconfigurations. In Proceedings of 5th European Symposium on Gender & ICT, Digital Cultures: Participation–Empowerment–Diversity. 2009.

[62] 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

[63] Verdict. 2018. Ethics, GDPR, and diversity: Verdict talks to the president of the Association for Computing Machinery, July 2018. Retrieved Sept. 11, 2018 from https://www.verdict.co.uk/association-for-computing-machinery-ethics/

[64] 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 & Sons, Inc., New York, NY, USA 105-140. http://dl.acm.org/citation.cfm?id=189200.189214

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