Semi-Automating (or not) a Socio-Technical Method for Socio-Technical Systems

Christopher Mendez, Zoe Steine Hanson, Alannah Oleson, Amber Horvath,
Charles Hill, Claudia Hilderbrand, Anita Sarma, Margaret Burnett
Oregon State University
Corvallis, Oregon, USA 97330
{mendezc,steinehz,olesona,horvatha,hillc,minic,anita.sarma,burnett}@oregonstate.edu

***Abstract*—How can we support software professionals who want to build human-adaptive sociotechnical systems? Building such systems requires skills some developers may lack, such as applying human-centric concepts to the software they develop and/or mentally modeling other people. Effective socio-technical methods exist to help, but most are manual and cognitively burdensome. In this paper, we investigate ways semi-automating a socio-technical method might help, using as our lens GenderMag, a method that requires people to mentally model people with genders different from their own. Toward this end, we created the GenderMag Recorder's Assistant, a semi-automated visual tool, and conducted a small field study and a 92-participant controlled study. Results of our investigation revealed ways the tool helped with cognitive load and ways it did not; unforeseen advantages of the tool in increasing participants' engagement with the method; and a few unforeseen advantages of the manual approach as well.**

Keywords—GenderMag, gender inclusiveness, socio-technical

I. Introduction

How should software professionals go about building human-adaptive socio-technical systems? Because socio-technical systems are systems in which humans are intrinsic parts of the system, building such systems effectively requires (1) human-centric concepts, in part (2) to model human behavior—but some developers may not have these skills.

A spectrum of methods—which are themselves socio-technical—exist to help, by integrating (1) and (2) into the design and/or implementation phases of building such systems. Examples include Heuristic Evaluation [43], Cognitive Walkthroughs [37, 58, 60], personas [2, 25, 31], and GenderMag [12]. Teams of software professionals can work together using these socio-technical methods to evaluate socio-technical systems in the design and/or implementation phases of building such systems.

However, methods like these are cognitively heavy, requiring software developers to immerse themselves in perspectives of people different from themselves. This is especially cognitively difficult for modeling people *very* different from themselves—such as having a different gender, as is the case when using the GenderMag method [12, 30].

This raises the question of whether semi-automating such a method might ease developers' cognitive burden. To investigate this question, we built a Chrome-based web extension for GenderMag called the GenderMag Recorder's Assistant. The tool semi-automates evaluating any prototype/mockup viewable in a Chrome browser: e.g., web-based apps (mobile or desktop), html mockups, etc.

To use the Recorder's Assistant, a software team navigates via the browser to the app or mockup they want to evaluate, then starts the tool from the browser menu. The main sequence is to view a persona (Fig. 1(c)) and proceed through the scenario of their choice from the persona's perspective, one action at a time. At each step, the tool's "context-specific capture" captures screenshots about the action the team selects (Fig. 1(a)), and records the answers to questions about it (Fig. 1(b)). The tool saves this sequence of screenshots and questions/answers to form a gender-bias "bug report."

Through these mechanisms, the Recorder's Assistant aims to reduce the cognitive load for software professionals working with GenderMag in three ways: visually marking the user action that software professionals are currently considering (Fig. 1 (a), box around the action "click on shift"); guiding the software professionals through the GenderMag questions, including a checklist of the persona's facets to be considered (Fig. 1(b)); and keeping the software practitioners' chosen persona visible and quickly accessible (Fig. 1(c)).

Fig. 1: The Recorder's Assistant tool during an evaluation of a mobile time-and-scheduling app. (Left): The app being evaluated is displayed with (a) a rectangle around the action the evaluators are deciding if a user like "Abby" will take. (Right): A blow-up of portions of the GenderMag features for the app: (b) the GenderMag question the team is answering at the moment, including a checklist of Abby's facets; and (c) a summary of the persona the team has decided to use (in this case, Abby).

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

978-1-5386-4235-1/18/$31.00 ©2018 IEEE

But might a semi-automated tool like the Recorder's Assistant do more harm than good? One potential problem might be disengagement. That is, since only one member of the software team would actually navigate through the tool, the rest of their team might disengage and become distracted by other apps on their computers (e.g., email and messages). Another might be a decrease in accuracy, such as if the team starts checking off boxes (e.g., Fig. 1(b)) without thinking much about them, or becomes distracted by having to deal with the tool itself.

To investigate whether these issues would arise, we conducted two studies: a small field study at a technology company, and a mixed-methods laboratory study with 92 participants. The following research questions guided our investigation:

- *RQ-Cognitive*: What benefits and disadvantages can a tool like the Recorder's Assistant bring to software teams' *cognitive load* and *recording accuracy*?
- *RQ-Engagement*: Can such a tool manage to "do no harm" to software teams' *engagement*?

II. Background and Related Work

A. Background: The GenderMag Method

GenderMag [12] is a socio-technical method. Its "socio-" aspect is that a software team works together to use it. Its "technical" aspect is, of course, that what the team is using it for is to evaluate software. It integrates human-centric concepts and mentally modeling other people into the process of evaluating software as follows.

GenderMag's foundations lie in research on how people's individual problem-solving strategies sometimes cluster by gender. GenderMag focuses on five facets of problem-solving:

(1) Motivations: More women than men are motivated to use technology for what it helps them accomplish, whereas more men than women are motivated by their interest in technology itself [3, 8, 10, 16, 27, 32, 36, 38, 56]. (2) *Information processing styles*: Problem-solving with software often requires information gathering, and more women than men gather information comprehensively—gathering fairly complete information before proceeding—but more men than women use selective styles—following the first promising information, then backtracking if needed [14, 20, 41, 42, 49]. (3) *Computer self-efficacy:* Women often have lower computer self-efficacy (confidence) than their peers, and this can affect their behavior with technology [3, 4, 5, 8, 10, 24, 29, 33, 38, 44, 46, 57]. (4) *Risk aversion:* Women tend statistically to be more risk-averse than men [18, 23, 59], and risk aversion can impact users' decisions as to which feature sets to use. (5) *Styles of Learning Technology:* Women are statistically more likely to prefer learning software features in process-oriented ways, and less likely than men to prefer learning new software features by playfully experimenting ("tinkering") [5, 8, 15, 17, 32, 51]. Any of these differences in cognitive styles is at a disadvantage when not supported by the software.

GenderMag brings these facets to life with a set of four faceted personas—"Abby", "Pat(ricia)", "Pat(rick)" and "Tim" (Fig. 2). Each persona's mission is to represent a subset of a system's target users as they relate to these five facets.

GenderMag intertwines these personas with a specialized Cognitive Walkthrough (CW) [58, 60]. The CW is a long-standing inspection method for identifying usability issues for new users to a program or feature. In a GenderMag CW, evaluators answer a question about each subgoal one of the personas might have in a detailed use-case, and two CW questions about each action, using the persona's five facets. Further, because GenderMag specializes in inclusiveness, a GenderMag CW inclusively collects answers from multiple team members. The questions are:

SubgoalQ: Will <persona> have formed this subgoal as a step to their overall goal? (Yes/no/maybe, why)
ActionQ1: Will <persona> know what to do at this step? (Yes/no/maybe, why)
Action Q2: If <persona> does the right thing, will s/he know s/he did the right thing & is making progress toward their goal? (Yes/no/maybe, why)

The GenderMag Recorder's Assistant tool aims to facilitate the recording of these answers.

B. Background: Mentally Modeling People

The GenderMag method's effectiveness rests on enabling software professionals to mentally *model* other people, a capability called "Theory of Mind." Theory of Mind is cognitive perspective-taking: the innate human ability to reason and make inferences about another's feelings, desires, intentions, and goals [47, 53]. Theory of Mind is similar to empathy—but empathy is *emotional* perspective-taking, whereas Theory of Mind is *cognitive* perspective-taking,

An example of Theory of Mind is someone (say, a software developer) building a model in their brain of another person (say, a user) who is different from themselves, and then "executing" that model in a new situation to predict how that person will behave. GenderMag's personas are meant to facilitate developers' Theory of Mind modeling of their users.

C. Related Work

GenderMag as a method (unsupported by a tool) has had several evaluations. Marsden and Haag did an eye-tracking study on the GenderMag personas and found that people's understanding and recollection of the facets were not significantly affected

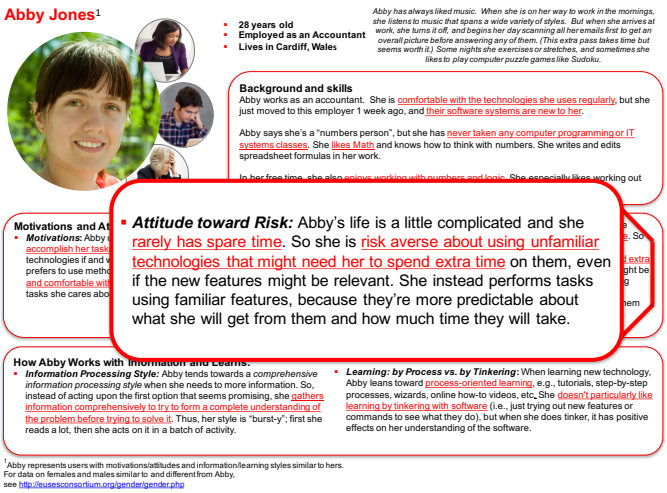

Fig. 2. Abby is a "multi-persona", meaning that she has multiple appearances and demographic portions of her are customizable [31]. One of the facets is blown up for legibility.

by the persona's picture (a favorable finding for these personas), but that people's perceptions of the persona's competence were affected by the picture (an unfavorable finding for these personas) [39]. A follow-up study investigated ways to mitigate this phenomenon, and found that "multi-personas"—in which a single persona shows pictures of different people the persona can represent—helped discourage gender stereotyping [31].

Evaluations of GenderMag's validity and effectiveness have produced strong results. In a lab study, professional UX researchers were able to successfully apply GenderMag, and over 90% of the issues it revealed were validated by other empirical results or field observations, with 81% aligned with gender distributions of those data [12]. In a field study using GenderMag in 2-to-3-hour sessions at several industrial sites [11, 30], software teams analyzed their own software, and found gender-inclusiveness issues in 25% of the features they evaluated. GenderMag has also been used to evaluate a Digital Library interface [21] and a learning management system [55], uncovering significant usability issues in both. In Open Source Software (OSS) settings, OSS professionals used GenderMag to evaluate OSS tools and infrastructure and found gender-inclusiveness issues in 32% of the use-case steps they considered [40]. Finally, in a longitudinal study at Microsoft, variants of GenderMag were used to improve at least 12 teams' products [9].

There is also related work on problems and/or tools on related methods, such as personas and cognitive walkthroughs. Personas were created and developed by Cooper as a way to channel, clarify, and understand a user's goals and needs [19]. Among the benefits claimed from using personas are inducing empathy towards users [2] and facilitating communication about design choices [48]. However, personas are not uncontroversial. Most pertinent to this paper is the issue of personas being ignored. For example, Friess reported that personas were referenced only 2% of the time in conversations regarding product decisions [25]. Friess also found that, even when evaluators used personas alongside CWs as focal points [25, 35], the personas were used only 10% of the time [25]. Thus, in this paper we measure engagement with the personas for both the tool and the paper method.

Regarding problems and tools for the other component of GenderMag, a specialized CW, Mahatody et al.'s [37] comprehensive literature survey of cognitive walkthroughs describes many CW variations, some of which focus on reducing problems with the classic CW [26, 54, 58] such as by reducing the time it requires. Niels et al. recommended that a tool for CWs might address issues like these by guiding the analyst through each CW step, in order to avoid missing steps and to more accurately record results, and to integrate a CW tool into a prototyping tool [34]. (The GenderMag Recorder's Assistant tool follows these recommendations.)

There is only a little work on creating such CW tools, but early in the lifetime of CWs, Rieman et al. created a tool with similar goals as the GenderMag Recorder's Assistant, in that it records the results of a human-run CW [50]. Their study found that analysts' predictions using the tool were accurate. However, their tool was based on an older, much more complex version of the CW, and was a stand-alone recorder, whereas the GenderMag Recorder's Assistant is integrated with the prototype being evaluated. Most pertinent to this paper, use of their tool was not compared to using a manual/paper version of the CW.

At the other end of the automation spectrum, a few researchers have created tools to automatically perform subsets of the cognitive walkthrough (e.g., [6, 7, 22]). Tools like these are different from the GenderMag Recorder's Assistant in that they handle only subsets of CWs, and are intended to *replace* humans in using such methods, whereas our investigation considers how to *support* humans using such methods. None of these works evaluates how using a tool impacts evaluators' effectiveness when software teams use a socio-technical method like GenderMag. That is the gap this paper aims to help fill.

III. Study #1: Initial Field Study

We began with a small field study to gain a real-world perspective. Two professional software developers at a West-Coast technology company, one man and one woman, conducted a GenderMag evaluation of one of their company's mobile printing apps (Fig. 3(left)) using the Recorder's Assistant tool. There are three roles in the process: *facilitator* (runs the walkthrough), *recorder* (records the results), and *evaluator* (answers the questions). One of the developers acted as both the facilitator and recorder, and both developers served as evaluators. We observed and video-recorded the session, which lasted about two hours. Both of the developers had prior experience using the (paper-based) GenderMag method.

Study #1 revealed evidence both against and for the tool reducing cognitive load. On the negative side, the tool sometimes distracted the participants from their evaluation task, essentially stealing cognitive cycles to think about the tool instead of the task when subtleties arose. For example:

West1 (minute 1:28): "ok, perform it…Ummmm, ok, what happened?" Researcher: "is it not letting you…oh here, hover over that…" Discovers duplicated screen shot had been entered, but it looks exactly the same, so looks like tool didn't respond.

Even so, the team's overall opinion was positive about the

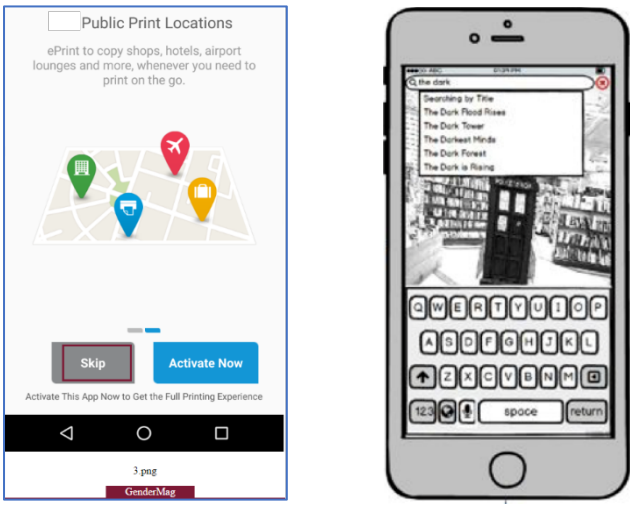

Fig. 3: (Left): A partial screenshot from the field study's mobile printing app; the red rectangle is around the "Skip" actions the participants are currently evaluating. (Right): One project a participant team brought to the controlled study was an "executable" mock-up of an augmented-reality bookstore navigator. Another was Fig. 1's mobile scheduling app.

tool's cognitive benefits. As Participant West1 put it:

West1 (minute 1:48, during debrief, in response to how it compared to paper): "…Way easier. This is way better than the paper version. … It keeps you focused."

For RQ-Engagement, the session also revealed both positive and negative effects. Participant West1 was always engaged—with his/her screen being projected s/he had little choice—but West2 occasionally disengaged from the task, catching up on email instead. Still, the brightly projected image captured both participants' gaze most of the time. More critically, perhaps because it included Abby pictures (as in Fig. 1), their evaluations consistently showed an Abby perspective. For example:

West1 (minute 33): "All indicators encourage the risk-averse user to push the activate button: skip is gray, and the text below …"
West2 (minute 34): " <if she hits skip> … <she doesn't> know what's going to happen."

Their faithfulness to an Abby perspective paid off in the kinds of insights Theory-of-Mind methods aim for, such as:

West1 (minute 37): "This <feature> is probably where we're losing half our <target users>."

Interestingly, when the participants needed to think more deeply about Abby, sometimes they chose to study the *paper* version of Abby rather than the version the tool was displaying on the projector—even though the displayed version had explicit links (Fig. 1(c)) to the full details.

West1 (minutes 19-20): "Motivation? Information Processing style?" Both turn to the paper description and start studying it. West1 reads aloud: "she prefers to use methods she is already familiar and comfortable with…" West1 turns back to screen and marks the facet. West2 (studies paper further): "Maybe information processing style." They both start reading aloud from the paper…
West1 (minute 36, looking at screen): "But Abby would read this, right?" Goes back to studying the paper.
West2: (minute 39): "Even though I've done GenderMag a couple of times, I still have to look at the paper."
West1: (minute 1:42, during debrief, when asked why referred to the paper persona): "I liked the ones up on the screen because they're very succinct… but sometimes I had to go back here <to the paper> because I thought there was something more, some detail that I wanted to consider."

These initial results, which are summarized in Table I, suggested the need for a more in-depth investigation, so we then conducted a controlled, mixed-methods laboratory study.

TABLE I: STRENGTH AND WEAKNESS EVENTS OBSERVED IN THE INITIAL FIELD STUDY.

	RQ-Cognitive	RQ-Engagement
Tool strengths	• Tool *"way easier."*	• Recorder fully engaged: Tool *"keeps you focused."*
Tool weaknesses	• Tool sometimes taxed cognition: e.g., *"ok, what happened?"*	• Non-recorder had laptop open, used it to multi-task. • Participants turned to Abby-on-paper, attended less to Abby-in-tool

IV. CONTROLLED STUDY METHODOLOGY

Study #2 used a between-subjects Tool vs. Paper design. We conducted it in two settings at a U.S. university: one setting primarily to collect quantitative data (classroom setting) and the other primarily to collect qualitative data (videorecorded in a lab). In both settings, teams of 2-4 participants performed GenderMag evaluations on their own software

A. Participants (both settings)

The 92 participants were junior and senior students recruited from two computer science courses. These courses enabled a controlled investigation with enough suitable teams for statistical power because: (1) the courses provided a reasonably large pool of software creators already on software teams of similar sizes. (2) These teams were in the process of creating software they cared about for their grades in these courses. (3) Their software was at a stage suitable for a GenderMag evaluation: mature enough to evaluate but early enough that changes could still be made inexpensively.

All students enrolled in the two courses performed the GenderMag evaluations as part of their coursework, but only teams who opted in are part of the reported study. That is, if a team opted into the study, their session outputs became part of our data; otherwise their outputs were used only for the class. Although a few participants had seen or used GenderMag before, their teams did not show any advantage from this: their teams' measures fell near the average (two slightly above, and two slightly below). Participant demographics are shown in Table II.

B. Procedures (both settings)

After the teams were randomly assigned to a treatment, they opted in or not as desired. As Table II shows, this process resulted in about half the participating teams performing their evaluations using the tool, and the rest using the paper materials from the GenderMag kit [13]. As in the field study, participant teams had a real stake in doing these evaluations, because they used the GenderMag method to find problems with their *own* software projects (e.g., Fig. 1(a) and Fig. 3 (right)), which they were developing over the course of the term.

To control variability, we pre-selected which persona—Abby—all teams would use. (If a team wanted to evaluate the software using a second persona, they could do so outside of the study session.) A few days before the sessions, we introduced

TABLE II: PARTICIPATING TEAMS, WITH 2-5 PARTICIPANTS PER TEAM, BY SETTING (COLUMNS) AND BY TREATMENT (ROWS: DARK ROWS ARE TOOL, LIGHT ARE PAPER). TOTALS: 41 TOOL PARTICIPANTS, 51 PAPER PARTICIPANTS.

	Classroom	Video lab	Treatment Totals
Number of teams	10 teams	2 teams	12 teams
	11 teams	3 teams	14 teams
Men	31 men	2 men	33 men
	31 men	7 men	38 men
Women	4 women	2 women	6 women
	9 women	2 women	11 women
Declined to state	0 people	2 people	2 people
	1 people	1 people	2 people
Had seen GenderMag before	4 people	0 people	4 people
	1 people	0 people	1 people

all teams to the Abby persona, and then told them to customize three fields of Abby—her age, place of residence, and occupation—to fit their own software project's target demographics. For example, GenderMag's prepackaged Abby is a 28-year-old accountant who lives in Wales, but among the teams' customizations were Abby as a 16-year-old Oregon high school student and as a 40-year-old Baltimore car mechanic.

We began with a brief tutorial on the GenderMag method. In the Tool treatment, we also helped participants set up the tool on their team's laptop, and briefly instructed them in how to operate the tool. The teams then performed their GenderMag evaluations, in which they used their customized Abby to walk through a use case they chose in their own software project. At each step, they answered the questions on the CW form (see the Background section) about whether and why Abby would act upon the "right" feature in the way they, the software's designers, had intended with their design. Finally, each participant filled out a NASA Task Load Index (TLX) questionnaire to report their impressions of cognitive load [28].

C. Treatments (Classroom): GenderMag via Tool vs. Paper

The classroom setting was two large classrooms (one room per treatment), each with multiple teams of 2-4 participants. The Tool teams walked through their use cases with their software prototypes embedded in the tool as in Fig. 1(a), and answered the questions as in Fig. 1(b). The Paper teams did the same things but without a tool: their prototypes were running on their laptops or on paper storyboards, but their CW questions were printed on paper with no limitations on what they could enter (e.g., no checkboxes) and unlimited space. In the Tool treatment, resources (forms, personas, etc.) were primarily computer-based, whereas in the Paper treatment, resources were primarily on paper. However, because some people prefer reading paper over screen and some prefer typing over writing, *both* treatments were allowed to add on use of paper or the computer for reading or writing. For example, some Tool teams turned to paper-based Abby, and some Paper teams typed their CW answers on their laptops using word processing.

D. Treatments (Lab): GenderMag via Tool vs. Paper

Participants in the lab setting followed the same procedures as in the classroom setting, but with their evaluations conducted one team at a time in a lab and videorecorded.

E. Data analysis (both settings)

We combined the settings for analysis. Qualitative data came primarily from the videorecorded setting's sessions. We transcribed the videos of each session, segmenting the resulting transcripts by conversational turn. We then qualitatively coded each conversational turn for the presence of the number of persona mentions within a conversational turn, any mentions of the persona's problem-solving facets (e.g., motivations, information processing style, etc.), and the presence of cognitive issues.

To measure how often participants explicitly referred to Abby, we coded each time a participant said "Abby", "she", or "her". To be conservative, we did not count instances of the participant simply reading the CW form questions aloud ("Would Abby have formed this subgoal as a step to her overall goal?").

We coded instances of facet engagement and of particular cognitive issues using prior works' GenderMag code sets for facets and cognitive issues [11, 30] (see the relevant results section for code set details). Two researchers independently coded 20% of the transcripts' conversational turns using these code sets and obtained 99% agreement (Jaccard index). The two researchers then split up the rest of the coding.

We also coded both settings' written CW forms for persona mentions and facet mentions using the same code sets as above. We segmented these forms by CW step (i.e., each new CW question started a new segment). Two researchers independently coded 20% of the data and reached 93% agreement (Jaccard index). The two researchers then split up the rest of the coding.

In total, we qualitatively coded 1681 conversational turns from the videorecorded setting and 392 CW form segments from both the videorecorded and classroom settings.

V. Results: Controlled Study

A. Results: Cognitive Load and Recording Accuracy

1) Participants' perceptions of cognitive load

To measure the 92 participants' perceptions of cognitive load, we used the NASA Task Load Index (TLX) questionnaire [28]. The TLX is a validated questionnaire with six questions, each answered on a scale from 1-21. Four of these questions measure perceived cognitive costs: how hard participants felt they had to work, how rushed the pace of the task was, how stressed they felt during the task's completion, and how high they felt the mental demand to be. The fifth question measures how successful they felt, and the sixth is on physical exertion.

The results of the participants' TLX responses were an interesting mix. As Fig. 4 shows, Tool participants felt that they had to work less hard (ANOVA, $F(1,90)=6.14$, $p=.0150$)—but also felt *more* stressed (ANOVA, $F(1,90)=6.4$, $p=.0129$). There were no differences between the two treatments in their perception of physical exertion, the amount of mental demand, or how rushed they felt, but Tool participants felt that they were less successful (ANOVA, $F(1,90)=4.2$, $p=.0445$).

The Tool participants' perception of working less hard is consistent with the Study #1 comment by participant West1, whose comparison of the tool with their prior experience with

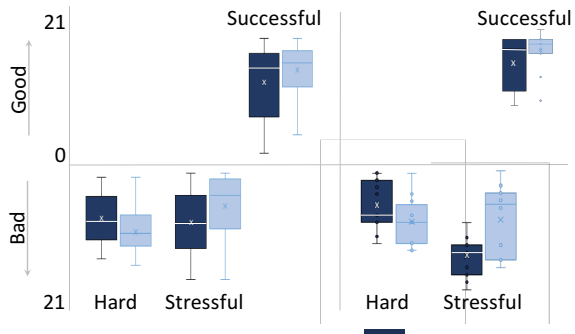

Fig. 4: TLX scores (out of 21). (Left bottom) Tool participants (N=41) felt the work was not as hard (down=Harder) as Paper participants did (N=51), but felt more "insecure, discouraged, irritated, stressed, annoyed" than Paper participants (down=more Stress). (Left top) Tool participants felt less successful than Paper participants (shown as TLX complement, so Up=more success). (Right) Tool recorders (N=11) did not work as hard as Paper recorders (N=12), but were much more stressed.

the paper version concluded that the tool was "way easier." Our interpretation is that the Tool participants' perception of working less hard was due to the tool keeping them on track when stepping through their prototypes, and also ensuring that their CW answers were tied to the actions they intended those answers for.

The fact that the Tool participants at the same time felt more stressed is also consistent with Study #1 data. Sometimes the tool behaved in ways that participants did not understand or had to be restarted, and this seems likely to have added stress. For example, one participant became confused by the tool's large collection of prototype screenshots:

Tool2-P1: And after that, oh my god…. I think there's too many screens.

Stress was particularly high for the Tool teams' recorders. Tool recorders had a median stress measure of 12, compared to only 5.5 for the Paper recorders. Ultimately, the cognitive cost of the stresses Tool participants reported may have played a part in their perceived lack of success, consistent with Schneider et al.'s findings that cognitive load interferes with Theory-of-Mind effectiveness [53].

2) *Actual Recording Accuracy*

However, Paper participants' perceptions of their own success were overly optimistic, or perhaps they simply discounted the importance of recording accuracy. We analyzed the verbalizations in the videorecordings for two types of recording errors: a team discussing a facet and deciding upon it verbally but the recorder omitting it, or the recorder including a facet that the team had not mentioned. The results showed that recording accuracy in both treatments was a bit problematic—but the videorecorded Tool teams recorded their facets more accurately than any of the Paper teams did, with Tool teams averaging 65% accuracy vs. Paper teams averaging only 35% accuracy.

3) *Two Cognitive Issues: "Where are we?" & Detours*

Prior work has reported accuracy issues in the GenderMag context to be disproportionately tied to two particular cognitive issues: *"Where are we?"* (participants losing track of which action with the prototype they are evaluating), and *detours* (participants digressing from the evaluation, such as getting sidetracked by talking about potential new features for their application) [30]. Thus, following the same procedures as this work, we investigated how often "where are we?" and detours arose for the teams in our study.

The "where are we" problems reported in the prior work rarely occurred, with only 6 instances in total out of a total of 1681 conversational turns, perhaps because the teams were smaller than those experiencing "where are we?" problems in prior work [30]. However, detours were problematic, with a total of 49 instances spanning over 12% of their conversational turns.

The detours were particularly problematic for Paper teams. As Table IV shows, the videorecorded Tool teams experienced fewer detours than Paper teams—especially lengthy detours. (Since "long" is a matter of judgment, we tried different threshold values, but they reveal similar patterns. Shown are the 5-turn and 10-turn thresholds.) Overall, the greater the number and/or length of detours, the more pervasive the inaccuracy problems. Note in Table IV's right three columns that, when Tool teams got sidetracked into detours, those teams recovered more quickly and got back on track, consistent with field study participant West1's comment that the tool "keeps you focused".

Table V summarizes the results of the Accuracy and Cognitive Issues subsections. Together with the summary of participants' perceptions of cognitive load (Table III), these results point out that (1) Theory-of-Mind modeling is hard work, and that (2) each of the Tool and the Paper approach have their own strengths in lightening the load.

B. *Results: RQ-Engagement*

1) *Persona Engagement By the Numbers*

GenderMag requires real engagement for participants to mentally build and then mentally "execute" models of people not necessarily like them. Thus, to measure engagement, we compared participants' explicit engagement with Abby (saying/writing "she", "Abby", etc.) in three ways: on teams' written forms, in their verbalizations, and against prior literature.

By all three measures, as Table VI and Fig. 5 (left) summarize, the teams were very engaged with the persona. This was

TABLE III: COGNITIVE LOAD SUMMARY: TOOL VS. PAPER PARTICIPANTS' PERCEPTIONS OF COGNITIVE LOAD.

	Hard work	Stress	Felt successful
Tool strengths	Not as hard (Study #1 & Study #2)		
Paper strengths		Less stressful (Study #2)	Felt more successful (Study #2)

TABLE IV: THE VIDEORECORDED TEAMS' ACCURACY PROBLEMS & COGNITIVE ISSUES IN 1681 CONVERSATIONAL TURNS, SORTED BY DEGREE OF INACCURACY (COLUMN 2). GRAY CHANGES AT 15%, 30%, ..., AND HIGHLIGHTS HOW DEGREE OF INACCURACY TENDED TO WORSEN AS DETOURS WORSENED. PAPER TEAMS TENDED TO HAVE MORE PROBLEMS WITH BOTH.

	Inaccurate recordings (% of facet instances)	Conversational turns spent in Detours + WAWs	"Long" detours (% of detour instances)		
			≥5 turns	≥10 turns	Mean length
Tool2	28%	7%	25%	13%	3.5 turns
Tool1	42%	9%	0%	0%	2.0 turns
Paper1	50%	22%	30%	10%	3.8 turns
Paper3	55%	18%	42%	21%	5.8 turns
Paper2	91%	6%	33%	33%	5.0 turns

TABLE V: SUMMARY OF TOOL VS. PAPER ACCURACY STRENGTHS. RECORDING ACCURACY HAS NO SHADING BECAUSE, ALTHOUGH TOOL WAS MORE ACCURATE THAN PAPER, NEITHER WAS STRONG.

	"Where are we?"	Detours	Recording accuracy
Tool strengths	Few problems (Study #2)	Shorter detours (Study #2)	Better recording accuracy (Study #2)
Paper strengths	Few problems (Study #2)		

true in both treatments: there was no significant difference between the Tool vs. Paper treatment, and both treatments' team engagement with Abby was comparable to prior literature.

However, one surprising similarity in Tool and Paper teams' engagement with Abby was *where* they looked when they wanted to remind themselves of Abby's attributes. Consistent with the Study #1 results, Tool participants often referred back to *paper* versions of Abby. For example:

Tool2-P2: (reads from paper) "Abby uses technology to accomplish her tasks, she learns new technologies when she needs to but prefers to use technology she's already comfortable with." (stops reading): "So yeah. Motivation…"

Tool1-P1: "Um," (reads from paper) "…gathers information to try to form a complete understanding" (stops reading). "Probably none of the above."

An arguably "ideal" level of engagement in a CW-based method like GenderMag would be for a team to refer to Abby at every step in their CW analysis. Remarkably, both the Tool and the Paper teams neared that ideal, referring to Abby in *almost every single segment*: in 94% and 97% of the CW steps, respectively (Table VI, bottom section).

2) *Facet Engagement By the Numbers*

Recall from Section II that the core of this method lies in its problem-solving facets. Thus, to measure engagement with these facets, we coded each of the teams' written CW forms for mentions of each of the five facets. (Duplicate mentions of the same facet were not counted.) As Fig. 5 (right) shows, the Tool teams mentioned significantly more facets per response than Paper teams did (Fisher's exact test, p=.0048, n=26).

3) *Depth of Engagement*

But did the Tool teams mark off checkboxes just because they were there, without really deciding on them? (Indeed, in a pilot of the field study on an earlier version of the GenderMag Recorder's Assistant that did not list "none of the above" as a checkbox option, participants did sometimes mark facets that they never discussed verbally or in writing.) To look for evidence of "brainful" engagement or lack thereof, we measured whether, for each facet they *checked off*, the videotaped teams gave other evidence of commitment to it via either a mention in their free-form response areas or a verbalization on the videorecordings. This measure showed engagement in 80-87% of the facets they marked.

An alternative lens on depth of engagement lies in what the teams actually said to one another about Abby. Some participants referred to Abby at a very surface level, with no information content about Abby. For example, in the quote below, the information content is not about Abby herself, but rather about the choices available:

Paper2-P2: "I'd say she will know what to do at this step because there's only 3 choices, 'yes', 'no', or 'cancel'…."

In contrast, some participants gave real attention to how Abby worked through the ways her facets led to her choices:

Tool2-P2: "And then I would also say willingness to tinker. Because she's not going to be willing to tinker with the screen to find out if it's the right screen or not."

To get a sense for teams' depth of engagement with Abby, we analyzed the videorecorded teams' verbalizations with explicit content about *both* Abby and her facets in a single conversational turn, like the one above. As Fig. 6 shows, the Tool teams showed much more evidence (via Abby-information content) of

TABLE VI: ENGAGEMENT: BOTH TOOL AND PAPER TEAMS MENTIONED ABBY AT RATES COMPARABLE TO PRIOR GENDERMAG RESULTS, AND BETTER THAN THE BEST PRIOR NON-GENDERMAG PERSONA RESULTS WE HAVE BEEN ABLE TO LOCATE. PRIOR RESULTS ARE SHADED. TOOL VS. PAPER RESULTS WERE NOT SIGNIFICANTLY DIFFERENT.

	Source	Explicitly mentioned persona (Abby)
Per conversational turn	Prior field work on personas [25]	…verbally in 10% of conversational turns
	Prior GenderMag field study (using paper) [30]	…verbally in 23% of conversational turns
	Prior GenderMag lab study (using paper) [31]	…verbally in 34% of conversational turns
	Tool teams	…verbally in 24% of conversational turns
	Paper teams	…verbally 28% of conversational turns
Per CW step	Prior GenderMag field study (using paper) [30]	…verbally while discussing 79% of the CW steps
	Tool teams	…written on 49% of CW steps, and …verbally in 94% of CW steps
	Paper teams	…written on 62% of CW steps, and …verbally in 97% of CW steps

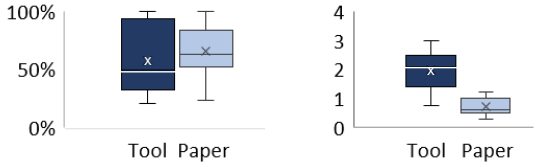

Fig. 5 Engagement: (Left:) Tool vs. Paper teams' mentions of Abby as a % of their written CW responses (no significant difference). (Right:) Number of facets per response: Tool teams mentioned significantly more facets/response.

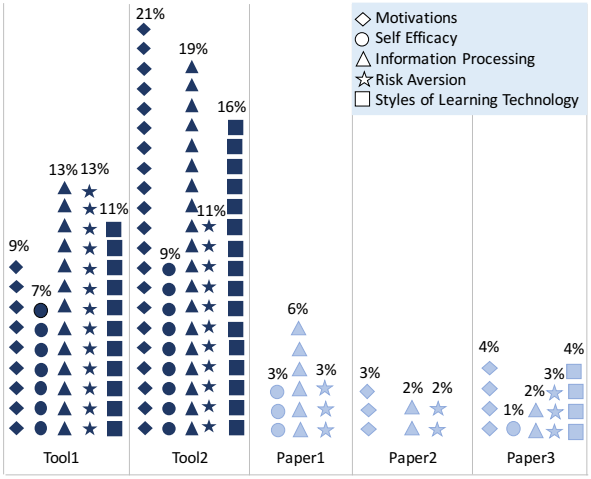

Fig. 6: How often each videorecorded Tool or Paper team verbalized Abby-information content, broken down here by facet, as a percent of their Abby-mentions. The Tool teams showed deeper engagement than the Paper teams.

engagement depth with Abby than the Paper teams did.

Given that the Paper teams' engagement was as strong as the strongest prior work we have been able to find on persona engagement, we expected a "ceiling" effect; i.e., we did not see room for much more engagement. However, the Tool teams surprised us. As Table VII summarizes, Paper teams were strong with engagement, but Tool teams were stronger.

VI. Discussion

Our results show that whether to "tool up" a sociotechnical Theory-of-Mind method like GenderMag is not a simple question. As Fig. 7 summarizes, our results revealed a checkerboard of complementary strengths in Paper vs. Tool.

A. *Are the strengths transferable…?*

Some of the strengths in supporting our participants may be due in part to the way each was presented (i.e., not inherent to tools or paper), and this suggests that tools or paper could obtain some of the strengths demonstrated by the other. As an example of tool-to-paper transferability, the tool's checkboxes seemed to remind participants of the facets. This could be implemented in the paper version by adding the same checkboxes to the paper form. An example of paper-to-tool transferability is that paper Abby made all of Abby's details readily available; this could be accomplished by adding a second display screen to a tool's set-up, so that Abby's complete details could always be displayed.

B. *…or Inherent?*

However, there are some strengths that may be inherent to what tool support vs. paper support can bring to a sociotechnical Theory-of-Mind method. For example, paper as a medium (1) brings less cognitive load, and cognitive load works against Theory-of-Mind [53]. Also, (2) the paper medium is tied to enhanced comprehension of written material [1], which is needed for empathy and engagement with personas like Abby, whose existence is solely in the form of a written description. This may explain why Tool teams so often turned to "paper Abby."

The Tool condition also brought key advantages to our participants that seem tied to the medium—e.g., the continually updated screen display. Recall that the Tool teams (with access to paper Abby) had greater depth of engagement with Abby than Paper teams did (also with access to paper Abby). We are again reminded of field study participant West1's observation that the tool "keeps you focused." The tool enabled a coordinated display of what exact action in the prototype was being evaluated involving what widget/feedback, and what had been said about it. This may help explain the Tool participants' rapid recovery from detours (recall Table IV).

TABLE VII: SUMMARY OF TOOL VS. PAPER ENGAGEMENT STRENGTHS IN SUPPORTING OUR PARTICIPANTS.

	Abby engagement	Facet engagement	Depth of engagement
Tool strengths	High (Study #1, Study #2)	Tool: more engagement (Study #2)	Tool: greater depth (Study #2)
Paper strengths	High (Study #2)		

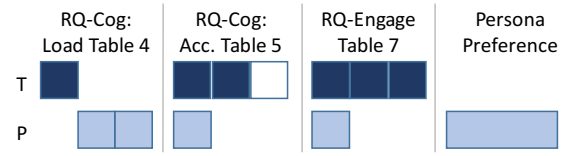

Fig. 7: Visual summary of Tool vs. Paper strengths (summarizes Tables 4, 5, and 7, plus the preference for paper-based Abby).

C. *Social aspects*

The social aspects of the tool seemed to help our participants with recording accuracy. GenderMag sessions occur in group settings (e.g., a conference room). In the paper-based method, one team member usually projects the prototype, and the rest of the team discusses the action they see playing out on the projector while the recorder somehow captures the discussion (using paper or word-processing on another computer). However, in the tool-based setting, the prototype step and CW questions are integrated on the projection screen, so the entire team can see what is being recorded for what action at the same time. This transparency may have been another reason for the Tool teams' better accuracy: if others are watching, a recorder may be more vigilant in capturing what they say, and other team members will have more opportunity to catch recording errors right away.

VII. Concluding Remarks

The complementary strengths of each medium suggests that the best ways to support developers' use of a method like GenderMag lie in strategic partnerships of tooling and paper.

The first category of strengths in Fig. 7, cognitive load, seems challenging to resolve because of the interdependencies among how stress (Paper was better), perceived ease of use (Tool was better), and feelings of success (Paper was better) interact with one another and with cognitive absorption/focus, engagement, and Theory-of-Mind processing [45, 52, 53]. How to go about resolving this tension is an open question.

The second and third categories yield more obvious possibilities. Accuracy needed work in both conditions (so no "good" choice here), but one commonality was a single recorder capturing everyone's ideas in real time. Perhaps distributing the recording to all team members and then sharing/combining what they wrote would improve accuracy on either medium. Engagement, on the other hand, was good in both conditions (no "bad" choice here). Still, the tool was better; perhaps the paper medium's facet engagement might be further improved by adding facet checkboxes to the paper forms, as mentioned earlier.

The fourth category, Personas, yields a clear choice. The participants' preferences, their ability to deeply engage with Abby, comprehend written material [1], and to learn and think about her facets [45], all point to paper personas.

Thus, the key is to find the right combinations of tools and paper to best support a sociotechnical Theory-of-Mind method like GenderMag, to enable software teams to create more human-centric, adaptable, and usable software for everyone.

The GenderMag Recorder's Assistant is an Open Source project, and we welcome contributions. To download it or contribute to it, go to http://gendermag.org.

REFERENCES

[1] R. Ackerman and T. Lauterman. Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. Computers in Human Behavior 28(5), pp. 1816-1828, 2012.

[2] T. Adlin and J. Pruitt, The Essential Persona Lifecycle: Your Guide to Building and Using Personas, Morgan Kaufmann/Elsevier, 2010.

[3] L. Beckwith and M. Burnett, Gender: An important factor in end-user programming environments? IEEE VL/HCC, pp. 107-114, 2004.

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

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

[6] M. Blackmon, P. Polson, M. Kitajima, and C. Lewis, Cognitive walkthrough for the web, ACM CHI, pp. 463-470, 2002.

[7] M. Blackmon, P. Polson, and C. Lewis, Automated Cognitive Walkthrough for the Web (AutoCWW), ACM CHI Workshop: Automatically Evaluating the Usability of Web Sites, 2002.

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

[9] M. Burnett, R. Counts, R. Lawrence, H. Hanson, Gender HCI and Microsoft: Highlights from a longitudinal study, IEEE VLHCC, pp. 139-143, 2017.

[10] M. Burnett, S. D. Fleming, S. Iqbal, G. Venolia, V. Rajaram, U. Farooq, V.Grigoreanu, and M. Czerwinski, Gender differences and programming environments: Across programming populations, IEEE Symp. Empirical Soft. Eng. and Measurement, Article 28 (10 pages), 2010.

[11] M. Burnett, A. Peters, C. Hill, and N. Elarief, Finding gender inclusiveness software issues with GenderMag: A field investigation, ACM CHI, pp. 2586-2598, 2016.

[12] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, GenderMag: A method for evaluating software's gender inclusiveness. Interacting with Computers 28(6), pp. 760-787, 2016.

[13] M. Burnett, S. Stumpf, L. Beckwith, and A. Peters, The GenderMag Kit: How to Use the GenderMag Method to Find Inclusiveness Issues through a Gender Lens, http://gendermag.org/ 2018.

[14] P. Cafferata and A. M. Tybout, Gender differences in information processing: a selectivity interpretation, in Cognitive and Affective Responses to Advertising, Lexington Books, 1989.

[15] J. Cao, K. Rector, T. Park, S. Fleming, M. Burnett, and S. Wiedenbeck, A debugging perspective on end-user mashup programming, IEEE VLHCC, pp. 149-159, 2010.

[16] J. Cassell, Genderizing HCI, In The Hand-book of Human-Computer Interaction, M.G. Helander, T.K. Landauer, and P.V. Prabhu (eds.). L. Erlbaum Associates Inc., pp. 402-411, 2002.

[17] S. Chang, V. Kumar, E. Gilbert, and L. Terveen, Specialization, homophily, and gender in a social curation site: findings from Pinterest, ACM CSCW, pp. 674-686, 2014.

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

[19] A. Cooper, The Inmates Are Running the Asylum, Sams Publishing, 2004.

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

[21] S. Cunningham, A. Hinze and D. Nichols, Supporting gender-neutral digital library creation: A case study using the GenderMag Toolkit, Digital Libraries: Knowledge, Information, and Data in an Open Access Society, pp. 45-50, 2016.

[22] A. Dingli and J. Mifsud, USEFul: A framework to mainstream web site usability thorugh automated evaluation, Int. J. Human Computer Interaction 2(1), pp. 10-30, 2011.

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

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

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

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

[27] J. Hallström, H. Elvstrand, and K. Hellberg, Gender and technology in free play in Swedish early childhood education, Int. J. Technology and Design Education 25(2), pp. 137-149, 2015.

[28] S. Hart, and L. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, Advances in Psychology 52, pp. 139–183, 1988.

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

[30] C. Hill, S. Ernst, A. Oleson, A. Horvath and M. Burnett, GenderMag experiences in the field: The whole, the parts, and the workload, IEEE VL/HCC, pp. 199-207, 2016.

[31] C. Hill, M. Haag, A. Oleson, C. Mendez, N. Marsden, A, Sarma, and M. Burnett, Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? ACM CHI, pp.6658-6671, 2017.

[32] W. Hou, M. Kaur, A. Komlodi, W. Lutters, L. Boot, S. Cotten, C. Morrell, A. Ant Ozok, and Z. Tufekci, Girls don't waste time: Pre-adolescent attitudes toward ICT, ACM CHI, pp. 875-880, 2006.

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

[34] N. Jacobsen, and B. John, Two case studies in using cognitive walkthrough for interface evaluation (No. CMU-CS-00-132), Carnegie-Mellon Univ School of Computer Science, 2000.

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

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

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

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

[39] N. Marsden and M. Haag, Evaluation of GenderMag personas based on persona attributes and persona gender, HCI International 2016 – Posters' Extended Abstracts: Proceedings Part I, pp. 122-127, 2016.

[40] C. Mendez, H. S. Padala, Z. Steine-Hanson, C. Hilderbrand, A. Horvath, C. Hill, L. Simpson, N. Patil, A. Sarma, M. Burnett, Open Source barriers to entry, revisited: A sociotechnical perspective, ACM/IEEE ICSE 2018.

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

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

[43] J. Nielsen, Enhancing the explanatory power of usability heuristics. ACM CHI '94, pp.152-158, 1994.

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

[45] S. Oviatt. Human-centered design meets cognitive load theory: Designing interfaces that help people think. In Proceedings of the 14th ACM International Conference on Multimedia, pp. 871-880, 2006. https://doi.org/10.1145/1180639.1180831

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

[47] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavior & Brain Sciences* 1(4), pp. 515-526, 1978.

[48] J. Pruitt and J. Grudin, Personas: practice and theory. ACM DUX, pp. 1-15, 2003.

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

[50] J. Rieman, S. Davies, D. Hair, M. Esemplare, P. Polson and C. Lewis, An automated cognitive walkthrough, ACM CHI, 1991.

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

[52] R. Saadé and B. Bahli. The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: An extension of the Technology Acceptance Model. Information & Management 42(2), pp. 317-327, 2005.

[53] D. Schneider, R. Lam, A. Bayliss, P. Dux. 2012. Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science* 23(8), pp. 842-847, 2012.

[54] A. Sears, Heuristic walkthroughs: Finding the problems without the noise, Int, J. Human-Computer Interaction 9(3), pp. 213- 234, 1997.

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

[56] S. Simon, The impact of culture and gender on web sites: An empirical study, The Data Base for Advances in Information Systems 32, pp. 18-37, 2001.

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

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

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

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