Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways?

Charles Hill[1], Maren Haag[2], Alannah Oleson[1], Chris Mendez[1], Nicola Marsden[2], Anita Sarma[1], and Margaret Burnett[1]

[1]Oregon State University
Corvallis, Oregon, USA
{hillc,olesona,mendezc,sarmaa,burnett}
@eecs.oregonstate.edu

[2]Heilbronn University
Heilbronn, Germany
{maren.haag,nicola.marsden}
@hs-heilbronn.de

ABSTRACT

Personas often aim to improve product designers' ability to "see through the eyes of" target users through the empathy personas can inspire—but personas are also known to promote stereotyping. This tension can be particularly problematic when personas (who, of course as "people" have genders) are used to promote gender inclusiveness—because reinforcing stereotypical perceptions can run counter to gender inclusiveness. In this paper we explicitly investigate this tension through a new approach to personas: one that includes multiple photos (of males and females) for a single persona. We compared this approach to an identical persona with only one photo using a controlled laboratory study and an eye-tracking study. Our goal was to answer the following question: is it possible for personas to encourage product designers to engage with personas while at the same avoiding promoting gender stereotyping? Our results are encouraging about the use of personas with multiple pictures as a way to expand participants' consideration of multiple genders without reducing their engagement with the persona.

Author Keywords

GenderMag; Gender; Stereotypes; Personas; Lab study

ACM Classification Keywords

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

INTRODUCTION

Prior research spanning multiple diverse age groups and populations shows that certain differences in the ways people use software tend to cluster by gender. In addition, there is evidence to suggest that many software products are not designed to take these differences into account. This means some users might find that these software products don't really meet their needs. Today's software designers and developers cannot afford to ignore the needs of any portion of their programs' user bases, given the sheer number of competing programs available to users.

The GenderMag method (Gender Inclusiveness Magnifier) [10] is a relatively new software inspection method that aims to help software creators identify gender-inclusiveness issues in their technologies. Emerging research indicates that GenderMag is very effective at unearthing these issues: while using GenderMag to evaluate their products, development and design teams from 3 different U.S. companies found surprisingly high numbers of gender-inclusiveness issues in their own software—25% of features evaluated had gender-inclusiveness issues [11, 44].

GenderMag is based in part on gendered personas, which raises the possibility of unintended stereotyping. Stereotyping is an ingrained human characteristic [75], so we cannot hope to stamp it out entirely. The issue we focus on is *techno*-stereotyping: if GenderMag's personas increased adverse techno-stereotyping of women, the method's components would be working against each other. Further, although GenderMag's main focus is to represent diverse problem-solving strategies, framing the method as a way to find gender-inclusiveness issues makes the concept of gender highly salient. This salience of gender per se automatically leads to gender stereotype activation [37]. Thus, GenderMag participants presented with a persona named Abby that includes a picture of a woman might techno-stereotype Abby with regards to her traits in using software or might inaccurately assume that "all women" use software similarly to Abby.

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

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025609

study, triangulating our results with eye tracking. We structure our investigation around two research questions:

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

BACKGROUND AND RELATED WORK

Background: The GenderMag Method

GenderMag's foundations lie in research that shows people's problem-solving strategies tend to cluster by gender. GenderMag focuses on five facets of problem-solving that have been found in literature that cluster by gender. The method uses faceted personas to give life to these facets and embeds the personas' usage in a facet-focused specialization of the Cognitive Walkthrough (CW) [74, 79]. The five facets are:

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

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

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

Risk: Prior research shows that females tend statistically to be more risk-averse than males when dealing with software [26], surveyed in [78], and meta-analyzed in [18]. Risk aversion may impact users' decisions regarding which features of software to use.

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

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

GenderMag combines the use of these personas with a specialized Cognitive Walkthrough (CW). The CW is a long-standing software inspection method used to uncover usability issues for users new to a program or feature [79]. During a GenderMag CW, evaluators step through a detailed use case (a goal and a list of actions) from the perspective of one of the personas and answer CW questions with respect to the five gendered problem-solving facets.

Related Work

Personas

Personas were created and developed by Cooper as a way to channel, clarify, and understand a user's goals and needs [20]. Today, personas are widely used in industry: sometimes simply to convey users' needs during software design, such as during informal role-playing tests or ideation [34, 59, 63, 67]. Recounted benefits of using personas include inducing empathy towards users [1] and facilitating communication about design choices [67]. Reasons cited for these benefits are: (a) that personas focus issues [45], (b) they provide uniform language to talk about the user and their needs [58], (c) they reduce conflict over what the user's perceived goals are [1], and (d) they summarize data about users in a relatable and concise format [35].

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

Issues that have been reported with personas include the following: practitioners not believing personas are credible; finding personas to be abstract, misleading, or impersonal; and seeing the personas' personifying details as irrelevant [17, 59]. Furthermore, research suggests that personas are most often used by the people that created them, in part because they have firsthand knowledge of the persona's intent and formalized training on personas in general [59]. On the

other hand, people who have not helped create the persona seem to prefer the raw data behind it, and are less likely to use the persona in design decisions [59]. We have also observed tensions between UX designers and software developers in which designers feel they must justify their personas' validities [55]. In addition, these findings suggest that software developers may have trouble empathizing with personas and that, for the persona to be accepted by developers, it must either be grounded in empirical work or a mainstream stereotype of a subset of users.

A persona photo or picture is part of most persona descriptions. Practitioners appreciate these images because they feel that they personalize the personas, although some worry that the photos might carry stereotypes [77]. To our knowledge, there have not been any studies tracking the actual use of persona pictures. Photos in person descriptions from other domains, such as resumes, receive a considerable amount of attention: for instance, eye tracking of LinkedIn profiles showed that recruiters spent almost one fifth of the time looking at the picture [29].

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

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

Related Work on Stereotypes

Personas foreground people and rely on impressions made based on the persona description. Their appeal lies in the fact that person perception is something all of us do automatically. It happens intuitively and has an imperative flavor, giving people the feeling of understanding a person. To a large extent, person perception relies on automatic processes [3, 38, 43], i.e., it happens without conscious endorsement. This makes it prone to biases like stereotypes that can distort social judgment [30]. Gender is a major source of bias in person perception, linked to prescribing certain roles and traits [8, 50], and feminine attributes or qualities displayed by females tend to be devalued [4, 31]. Gender is also closely linked to the two basic dimensions that we rely on to judge other people: When we meet someone, we intuitively make judgments of their warmth and their competence [22, 23, 32].

Personas inherit the tensions and biases regarding the perception of others [55]. A content analysis of personas in use showed that male and female personas tended to be presented as equally competent, but tended to rely on stereotypes regarding the warmth dimension [57]. As the GenderMag personas are designed with the explicit aim of highlighting gender differences, they cannot prevent being subject to the very processes that people employ in gendered perceptions. Studies using the GenderMag personas found that the presence of masculine problem-solving facets led people to attribute higher competence to the personas with those facets, even though each GenderMag persona is carefully designed to display equal competence to the others [56]. Therefore, there is a need for further research on what gender-inclusiveness can look like with regard to personas and how personas can be designed to alleviate the stereotyping associated with gender.

Cognitive Walkthroughs

As mentioned before, specialized Cognitive Walkthrough (CW) forms the foundation of the GenderMag method. The most up-to-date comprehensive study of CWs we could locate is the 2010 survey by Mahatody et al. [53]. Their survey describes many variations of the CW introduced by Lewis [52] and updated by Wharton et al. [79]. Later adaptations to the CW include such variations as having users in the CW during the process [36] or incorporating theories of cognition [28, 70]. Other modifications of the CW focus on solving problems identified with the classic CW process [71, 74].

One of the earliest responses to CW issues outside of revisions to the original method was that of Spencer's streamlined CWs [74]. Their work identified constraints of CWs that reduced the utility of the process in practice. After identifying these issues, Spencer changed the CW method in ways that attempted to fix these problems. Streamlined CWs reduced the number of questions in the CW to relieve the issues Spencer found.

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

METHODOLOGY

Our manipulation consisted of presenting the GenderMag Abby persona with either a single or four different pictures (Figure 1). Abby was specifically designed for use with the GenderMag method, and (with the single picture) has been

employed by various companies that used GenderMag. For the four-picture treatment (Figure 1), we added three pictures to the persona to show that a persona with these problem-solving facets could possess socio-demographic attributes different from the young, white, female Abby on the original persona. Since Abby focuses on facets that have been shown to affect women more than men [10], the manipulation depicted more women than men. We added a footnote to the manipulated persona explaining that Abby represents users with motivations/attitudes and information/learning styles similar to hers and offered a link to find further information. For brevity, we refer to the manipulated version of Abby as multiAbby, and the non-manipulated version as soloAbby.

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

In order to answer our research questions, we ran two studies. Study 1, conducted at a university in Germany, used eye tracking to analyze participants' gaze on different parts of the persona description sheet. Study 2, based at a university in the US, examined the effect of the manipulation both with use in actual GenderMag sessions (referred to as *GenderMag* condition) as well as in sessions where participants only viewed the persona (referred to as *PersonaOnly* condition) and gave their impressions on Abby and her problem-solving facets.

Study 1 (Eye tracking) Methodology

Procedure

Participants of Study 1 were a convenience sample of professionals in the field of software development, research, and management. They were not compensated for their participation. 14 professionals (5 females, 9 males) participated in the eye-tracking study, filled out the questionnaire (Section Study 2 Methodology), and were debriefed. We instructed all participants in the same way at the beginning of the experiment about the usage of the devices and the procedure. We then presented the GenderMag persona "Abby" on a screen. In a between-participants design with two levels, participants were randomly assigned to the presentation with one vs. four pictures. Seven participants saw the persona description with one picture, seven participants saw it with four pictures. We collected eye-tracking data using a Tobii X60 Eye Tracker with preliminary data analysis in Tobii Studio, and further analysis was performed in SPSS Statistics 22. We placed the eye tracker approximately 70 cm distance from the participant's eyes, and the vertical angle that the screen made from the participant's view was less than 35°.

Data analysis

In order to analyze participants' gaze on the different parts of the persona description and pictures, we defined areas of interest (AOI). Ten AOIs were defined for the one-picture condition (name, picture Abby, age/employment etc., abstract, background and skills, and the five facets, i.e., motivations, computer self-efficacy, attitude towards risk, information processing style, tinkering). For the four-picture condition, we defined 14 AOIs: three for the additional pictures and one for the footnote that was included to explain the usage of the four pictures (see Figure 2). As the dependent variable, we used the time spent inspecting the AOI, i.e., the duration of the visit in seconds. We measured the duration spent on each facet, accounting for length of facet description by measuring the duration of gaze per word in each facet. We also looked at the sequence in which the participants fixated the AOIs. Additionally, we used the number of fixations to create a visual overview ("heat map") of the gaze's dynamic.

Study 2 Methodology

Participants in Study 2 were mostly students at a University in the US, though we did not limit participation to only students. We recruited participants by emailing announcements and distributing and posting flyers around campus and the surrounding areas. All participants were over 18 years old. The final participant count was 36 females and 36 males, spanning many age groups, academic majors and statuses.

Figure 1: The pictures on the persona profiles. SoloAbby participants viewed personas with the large picture (left), and multiAbby participants viewed all four as shown here.

Figure 2: Areas of interest (AOIs) on multiAbby. SoloAbby's AOIs were equivalently adapted to the gaze without the AOIs covering the three extra pictures and the footnote.

In a 2x2 between-participants design, independent variables included the picture manipulation (one vs. four pictures) and the use of the persona in a GenderMag session (with GenderMag vs. PersonaOnly). Participants were randomly assigned to treatment sessions. Results across Study 1 (Germany) and Study 2 (US) did not show any significant differences. The GenderMag vs. PersonaOnly results were not significantly different either. Therefore, In the results sections we report aggregated results with N = 86:

- Group A: soloAbby, PersonaOnly, n = 24 (7 Germany, 17 US);
- Group B: multiAbby, PersonaOnly, n = 23 (7 Germany, 16 US);
- Group C: soloAbby, with GenderMag, n = 18 (US);
- Group D: multiAbby, with GenderMag, n = 21 (US).

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

PersonaOnly Sessions

We presented Groups A and B with the Abby persona and then asked participants to "get to know" her. Participants in these groups read the persona silently and could ask the researcher for clarification if they had questions about Abby or her problem solving-traits.

After participants read the persona, we removed the persona and they were given the first questionnaire (described in section Data Analysis). Following the questionnaire, Abby was returned to them and they were asked to fill out a second questionnaire (also described in section Data Analysis).

GenderMag Sessions

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

We gave participants a brief introduction to the method before they began, during which they studied the persona (similar to the PersonaOnly condition). During the walkthrough, participants evaluated a feature of a popular word processing software using GenderMag from Abby's perspective.

We observed the roughly 1 hour sessions, and we also video-recorded (or audio-recorded, if the participants did not consent to video-recording) and later transcribed each session. After performing the GenderMag walkthrough, participants were given the same questionnaires as groups A and B.

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

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

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

Data Analysis

Gender Stereotyping

To measure gender stereotyping, we measured to what extent participants applied traditional feminine attributes to Abby. This dependent variable was operationalized with two instruments yielding a total of 17 items: a short version [76] of the Bem Sex-Role Inventory BSRI [7], and warmth/competence questionnaire of the stereotype content model SCM [33].

To elicit the extent to which the participants were stereotyping, we compared the scores of the BSRI and the SCM with results found in representative samples: For the BSRI the test values were based on Donnelly and Twenge [25] who found women's feminine role at M = 5.0, women's masculine role at M= 4.9, men's feminine role at M = 4.6, men's masculine role at M = 5.1. For the SCM we used Asbrock's [2] means of women's warmth at M = 5.6 and competence at M = 4.2, and men's warmth at M = 4.2 and competence at M = 5.6 (transformed from a five to a seven point Likert scale by proportional transformation [19]).

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

Facet Perception

We operationalized facet perception by applying the facet attributes of the GenderMag persona description to the persona [56]. For each facet, two items had been developed. For example, motivation (task orientation) was measured with "spends money on technology because new technology is fun or cool" (reverse code) and "spends time or money on technology mainly to accomplish some work or task goal". The results for facet perception were submitted to ANOVA with the manipulation of the pictures (one vs. four pictures) and the use in a GenderMag session (with GenderMag vs. PersonaOnly) as between-participant factors.

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

To determine whether or not a facet was recalled correctly, we determined acceptable Likert scale answers based on the

descriptions of the facets in the persona: participant responses should reflect the descriptions of the facets in the persona. For instance, the description of Abby's attitude towards risk contains the phrase *"Abby is risk averse when she uses computers to perform tasks."* Therefore, the *"tries to avoid risk"* item should be rated as greater than 4 (i.e., on the Likert scale: moderately, very, or extremely) to be correct.

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

Gendering of Search Scope

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

Engagement with Persona

The transcripts of the GenderMag sessions were analyzed to identify indicators of participant engagement with Abby. We operationalize *"engagement with the persona"* similarly to past literature [34]: we measured the invocation of Abby in relation to the number of conversational turns. We split the transcripts by conversational turn (as has been done in [34, 44]). We then counted the number of times the persona was invoked during each conversational turn. To be conservative, we didn't count invocations if the participant was reading a question from the CW forms.

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

Confusion regarding Persona

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

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

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

RESULTS

RQ1: How do people gender-stereotype personas in the context of gender-inclusiveness?

Stereotyping

To measure stereotyping of the persona, we asked participants to rate the Abby persona on traditional feminine and masculine attributes in Bem's Sex Role Inventory (BSRI), and asked them to evaluate her warmth and competence. Over all participant groups, Abby's BSRI-masculine score ($M = 3.6$, $SD = 0.80$) was lower than her BSRI-feminine score ($M = 4.5$, $SD = 0.61$; $p = .000$). Abby's SCM scores for competence ($M = 4.4$, $SD = 0.84$) were lower than for warmth ($M = 4.9$, $SD = 0.77$; $p = .000$).

To quantify whether or not participants saw Abby as traditionally feminine or masculine, we compared our results with BSRI scores found in current studies in which participants rated real people's gender roles. These studies yield

women's masculine role at M = 4.9 and feminine role at M = 5.0; men's masculine role at M = 5.1 and feminine role at M = 4.6 [25]. We analyzed our BSRI scores in a one sample t-test with the reference scores, and found that Abby's BSRI-masculine score is lower than women's or men's masculine scores (p = .000). Abby's BSRI-feminine score is lower than women's BSRI-masculine score (p = .000) and there was a tendential difference to women's BSRI-feminine score (p = .060) – i.e., feminine traits and masculine traits our participants attributed to the GenderMag persona are lower than all the BSRI and scores used as a reference (see Table 1).

We performed a similar comparison for the SCM (warmth and competence). For the SCM, we used a baseline of women's competence at M = 4.2 and warmth at M = 5.6, and men's competence at M = 5.6 and warmth at M = 4.2 [2]. The one sample t-test yielded significant differences to the test value of 4.2 (p = .018 for women's competence [2] and p = .000 for men's warmth [2]); and for the test value 5.6 (p = .000 for men's competence [2] and p = .000 for women's warmth [2] i.e., the warmth and competence our participants attributed to Abby was in between the average warmth and competence typically attributed to men and women.

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

Engagement

We measure engagement by how often Abby was invoked during discussion in the GenderMag sessions. Past non-GenderMag work [34] indicates that personas were used between 2% and 10% of conversational turns, and a previous GenderMag study [44] showed invocation rates of up to 23%. We use a similar metric to measure engagement with the persona: we counted the conversational turns between participants during which the participants invoked the Abby persona. We found that our participants invoked Abby during 34% of conversational turns (Table 2).

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

Table 1: Abby's scores compared to reference scores (N=86).

RQ2: Can we reduce stereotyping by introducing a diverse "cast" of personas all representing a single persona's traits, and does that negatively affect engagement, learning, or turbulence?

Stereotyping

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

Gendered search scope: The t-test comparing responses from multiAbby participants and soloAbby participants showed that for the "friends unlike Abby", the participants of the multiAbby condition named female "friends unlike Abby" at M = 37 % (SD = .45), whereas, the participants of the soloAbby condition named female "friends unlike Abby" at M = 26 % (SD = .33, p = .001), i.e., significantly more female "friends unlike Abby" were named by the participants in the multiAbby condition. There was no significant difference between the responses for "friends like Abby".

Learning

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

To determine whether or not a facet was recalled correctly, we compared the results to the Likert scale answers that would correctly reflect the facets in the persona. Both groups generally recalled facets correctly, with the exception of one information processing style item ("selective in dealing with information" was M = 5, SD =1.55).

Turbulence

We identified implicit and explicit turbulences from the GenderMag transcripts. Recall that we define Explicit turbulence as a participant expressly stating they were confused by the persona, and Implicit turbulence as a participant implying they were confused by the persona without stating it outright.

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

Table 2: Invocations of Abby in this study vs. other work.

We identified a total of four instances of explicit turbulences in the transcripts; we found 216 instances of implicit turbulences across both groups (see Table 3 for more details). We performed Fisher's exact test to determine whether or not either treatment experienced more turbulence. No significant differences could be found regarding multiAbby vs. soloAbby for either implicit or explicit turbulence.

Engagement

By Conversational Turn: We performed Fisher's exact test to determine whether or not either treatment referred to Abby more often. No significant differences could be found regarding multiAbby vs. soloAbby. Participants in the multiAbby group invoked Abby during 35.95% of their conversational turns, while soloAbby participants invoked Abby in 28.38% of their turns (Table 4).

Eye Tracking: The analysis of the eye-tracking data shows which parts of the persona the participants engaged with visually. Participants read through the text from top to bottom and looked at the AOI covering the footnote only at the end, and not immediately after viewing the pictures. With respect to the overall time that participants spent looking at the persona description, the manipulation did not show any significant differences: the average duration of the visits of all the areas of interest (AOIs) for N = 14 was M = 140.36 seconds and SD = 25.42 for multiAbby, and M = 145.35 seconds (SD = 35.63) for soloAbby. The aggregated gaze distribution over all participants in the multiAbby condition is shown in Figure 3, and for soloAbby in Figure 4.

The results show differences in the duration that participants look at background/skills and at the name of the persona: In the multiAbby condition participants looked at the background/skills longer (M = 32.72 seconds, SD = 3.15) than in the soloAbby condition (M = 30.10 seconds, SD = 9.27, p = .010). Also, multiAbby participants tended to look at the name longer (M = 0.80 seconds, SD = 0.50) than the participants in the soloAbby condition (M = 0.25 seconds, SD = 0.22, p = .056).

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

To account for different lengths of facet descriptions, we measured gaze duration per word as a measure of the time spent on each facet. A significant difference between soloAbby and multiAbby could not be found. The average durations that the participants spent on each facet can be seen in Table 5. The average time per word spent on the facets was higher for computer self-efficacy compared to all other facets ($p < .05$), the comparison of the other facets did not yield significant results.

The average time soloAbby participants spent on the picture (M = 2.18 seconds, SD = 3.35) did not differ significantly from the time multiAbby participants looked at the four pictures altogether (M = 1.72 seconds, SD = 2.10). In fact, participants in both groups looked at the picture for less than 2% of their total time. Male and female participants did not differ significantly (women: M = 2.92 seconds, SD = 3.82; men: M

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

Table 3: Mean instances of turbulence per participant.

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

Table 4: How often participants in each group invoked Abby.

Figure 3: Eye fixations, aggregated over all participant in multiAbby eye-tracking condition (n=7; darkest red = 30.16 counts)

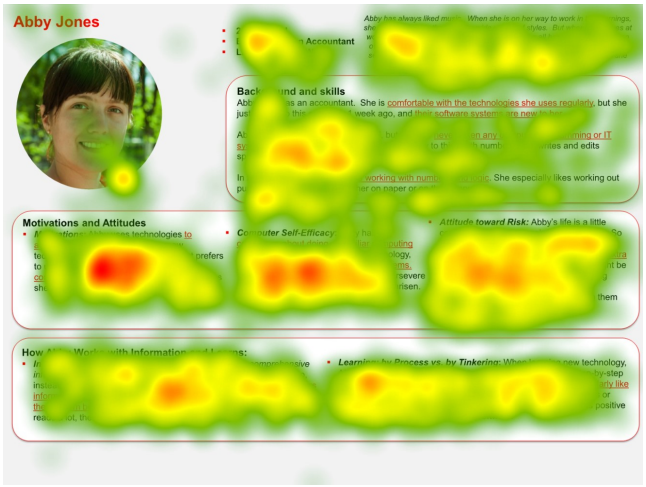

Figure 4: The soloAbby heat map

= 1.41 seconds, SD = 1.90). Participants in the multiAbby condition spent a similar amount of time reading the footnote (M = 2.17 seconds, SD = 2.77). Of these participants, one person did not look at the footnote at all. The sequence of the eye fixations showed that the other ones (n = 6) looked at it after looking at all the other AOIs.

DISCUSSION

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

Stereotyping

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

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

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

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

Table 5 Mean durations of gaze (seconds) on facets. *=different from all other facets (p < .05)

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

Believability

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

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

Education

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

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

Engagement

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

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

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

Four pictures or one?

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

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

Why did participants spend so little time looking at the picture? Perhaps they were aware that the picture on the persona is just an illustration not conveying any "real" information about a person – an arbitrary picture of a person, meant to illustrate and underline the persona description. Thus, illustrating a persona description with four instead of one picture accentuates the message that Abby could be any age, any ethnicity, and/or any gender. This may explain why multiAbby participants named significantly more female "friends *un*like Abby" than the soloAbby participants did.

CONCLUSION

In this paper, we present the first investigation of multiple unrelated portraits on a persona and measure participants' perceptions of this modified persona. This paper provides evidence to suggest that people's perceptions of personas are perhaps not as straightforward as they seem. We saw evidence to support this from multiple perspectives:

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

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

Pictures: Participants looked at Abby's picture - whether solo or multi - for less than 2% of the time they spent looking at the persona description. We expected the picture to receive a much larger portion of participants' attention. This suggests that the participants realized that the persona's appearance was not an important aspect of the persona.

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

ACKNOWLEDGMENTS

This work was funded in part by NSF 1253786, 1314384, 1528061, and 1559657; the Brigitte-Schlieben-Lange-Programm/Ministry of Science, Research, and Arts Baden-Württemberg, Germany; and Federal Ministry of Education and Research, Germany (BMBF), FKZ 01FP1603.

REFERENCES

1. Tamara Adlin and John Pruitt. 2010. *The Essential Persona Lifecycle: Your Guide to Building and Using Personas.* Morgan Kaufmann/Elsevier, San Francisco, CA.

2. Frank Asbrock. 2010. Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology* 41, 2: 76–81.
3. John A. Bargh. 2013. *Social psychology and the unconscious: The automaticity of higher mental processes.* Psychology Press.
4. Manuela Barreto, Naomi Ellemers. 2015. Detecting and Experiencing Prejudice: New Answers to Old Questions. *Advances in Experimental Social Psychology* 52: 139-219.
5. Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. 2005. Effectiveness of end-user debugging software features: Are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '05), 869-878. http://doi.acm.org/10.1145/1054972.1055094
6. Laura Beckwith, Cory Kissinger, Margaret Burnett, Susan Wiedenbeck, Joseph Lawrance, Alan Blackwell, and Curtis Cook. 2006. Tinkering and gender in end-user programmers' debugging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '06), 231-240. http://doi.acm.org/10.1145/1124772.1124808
7. Sandra L. Bem. 1981. *Bem Sex-Role Inventory: Professional Manual*. Consulting Psychologists Press, Palo Alto, CA.
8. Marilynn B. Brewer. 1988. A dual process model of impression formation. In *Advances in Social Cognition* (Vol. 1), Thomas K. Srull and Robert S. Wyer (eds.). Psychology Press, New York, 1-36.
9. Margaret Burnett, Laura Beckwith, Susan Wiedenbeck, Scott D. Fleming, Jill Cao, Thomas H. Park, Valentina Grigoreanu, and Kyle Rector. 2011. Gender pluralism in problem-solving software. *Interacting with Computers* 23, 5: 450–460.
10. Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers*, online January 2016. doi:10.1093/iwc/iwv046.
11. Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender inclusiveness software issues with GenderMag: A field investigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '16), 2586-2598. http://doi.acm.org/10.1145/2858036.2858036.2858274
12. Margaret Burnett, Scott D. Fleming, Shamsi Iqbal, Gina Venolia, Viyda Rajaram, Umer Farooq, Valentina Grigoreanu, and Mary Czerwinski. 2010. Gender differences and programming environments: Across programming populations. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (ESEM '10), 28. http://doi.acm.org/10.1145/1852786.1852824
13. Patricia Cafferata and Alice M. Tybout. 1989. *Gender Differences in Information Processing: A Selectivity Interpretation, Cognitive and Affective Responses to Advertising*. Lexington Books.
14. Jill Cao, Kyle Rector, Thomas H. Park, Scott D. Fleming, Margaret Burnett, and Susan Wiedenbeck. 2010. A debugging perspective on end-user mashup programming. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing* (IEEE '10), 149-156.
15. J. Cassell. 2002. Genderizing HCI. In *The Handbook of Human-Computer Interaction,* M.G. Helander, T.K. Landauer, and P.V. Prabhu (eds.). L. Erlbaum Associates Inc., Hillsdale, NJ, 402-411.
16. Shou Chang, Vikas Kumar, Eric Gilbert, and Loren Terveen. 2014. Specialization, homophily, and gender in a social curation site: findings from Pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (CSCW '14), 674-686. http://doi.acm.org/10.1145/2531602.2531660
17. Christopher N. Chapman and Russell Milham. 2006. The personas' new clothes: methodological and practical arguments against a popular method. *Human Factors and Ergonomics Society Annual Meeting* 50: 634-637.
18. Gary Charness and Uri Gneezy. 2012. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization* 83, 1: 50–58.
19. Andrew Colman, Claire Norris, and Carolyn Preston. 1997. Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports* 80: 355-362.
20. Alan Cooper. 2004. *The Inmates Are Running the Asylum*. Sams Publishing.
21. Constantinos K. Coursaris, Sarah J. Swierenga, and Ethan Watrall. 2008. An empirical investigation of color temperature and gender effects on web aesthetics. *Journal of Usability Studies* 3, 3: 103-117.
22. Amy Cuddy, Susan Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology* 40: 61-149.
23. Amy Cuddy, Susan Fiske, Virginia Kwan, Peter Glick, Stephanie Demoulin, Jacques-Philippe Leyens, et al. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology* 48, 1: 1-33.
24. Soussan Djamasbi, Marisa Siegel, and Tom S. Tullis.

2012. Faces and viewing behavior: An exploratory investigation. *AIS Transactions on Human-Computer Interaction* 4, 3: 190-211.

25. Kristin Donnelly and Jean M. Twenge. 2016. Masculine and feminine traits on the Bem sex-role inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*: 1-10. http://dx.doi.org/10.1007/s11199-016-0625-y.
26. Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association* 9, 3: 522–550.
27. Alan Durndell and Zsolt Haag. 2002. Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. *Computers in Human Behavior* 18, 5: 521–535.
28. Joel U. Eden. 2007. Distributed cognitive walkthrough (DCW): a walkthrough-style usability evaluation method based on theories of distributed cognition. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition* (C&C '07), 283-283 http://doi.acm.org/10.1145/1254960.1255019
29. Will Evans. 2012. Eye-tracking online metacognition: cognitive complexity and recruiter decisionmaking. Retrieved September 9, 2015 from http://cdn.theladders.net/static/images/basicSite/pdfs/TheLadders-Eye-Tracking-StudyC2.pdf
30. Susan Fiske. 2000. Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology* 30, 3: 299-322.
31. Susan Fiske. 2015. Intergroup biases: a focus on stereotype content. *Current Opinion in Behavioral Sciences* 3: 45-50.
32. Susan Fiske, Amy Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences,* 11, 2: 77-83.
33. Susan Fiske, Amy Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82, 6: 878-902.
34. Erin Friess. 2012. Personas and decision making in the design process: an ethnographic case study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1209-1218.
35. Kim Goodwin. 2009. *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Wiley, Indianapolis, IN.
36. Toni Granollers and J. Lorés. 2004. Incorporation of users in the evaluation of usability by cognitive walkthrough. In *HCI Related Papers of Interacción*. 243-255.
37. Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1: 4.
38. Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97, 1: 17-41.
39. Valentina Grigoreanu and Manal Mohanna. 2013. Informal cognitive walkthroughs (ICW): paring down and pairing up for an agile world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 3093-3096. http://doi.acm.org/10.1145/2470654.2466421
40. Jonathan Grudin. 2006. Why personas work: The psychological evidence. In *The Persona LifeCycle: Keeping People in Mind Throughout Product Design*, John Pruitt and Tamara Adlin (aut). Morgan Kaufmann Publishers.
41. Jonas Hallström, Helene Elvstrand, and Kristina Hellberg. 2015. Gender and technology in free play in Swedish early childhood education. *International Journal of Technology and Design Education* 25, 2: 137-149.
42. Kathleen Hartzel. 2003. How self-efficacy and gender issues affect software adoption and use. *Communications of the ACM – Why CS students need math* 46, 9: 167–171.
43. E. Tory Higgins. 1996. Knowledge activation: Accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles*, Guilford Press, New York, 133-168.
44. Charles Hill, Shannon Ernst, Alannah Oleson, Amber Horvath and Margaret Burnett. 2016. GenderMag Experiences in the Field: The Whole, the Parts, and the Workload. In *Proceedings of the IEEE Symposium on Visual Languages and Human-centric Computing* (IEEE '16).
45. Karen Holtzblatt, Jessamyn B. Wendell, and Shelley Wood. 2004. *Rapid Contextual design: A How-to Guide to Key Techniques for User-Centered Design*. Morgan Kaufmann, San Francisco, CA.
46. Weimin Hou, Manpreet Kaur, Anita Komlodi, Wayne G. Lutters, Lee Boot, Shelia R. Cotten, Claudia Morrell, A. Ant Ozok, and Zeynep Tufekci. 2006. Girls don't waste time: Pre-adolescent attitudes toward ICT. In *Proceedings of the CHI EA Conference on Human Factors in Computing Systems* (CHI '06), 875-880. http://doi.acm/org/10.1145/1125451.1125622
47. Ann H. Huffman, Jason Whetten, William H. Huffman.

2013. Using technology in higher education: The influence of gender roles on technology self-efficacy. *Computers in Human Behavior* 29, 4: 1779–1786.

48. Tejinder K. Judge, Tara Matthews, and Steve Whittaker. 2012. Comparing collaboration and individual personas for the design and evaluation of collaboration software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1997-2000. http://doi.acm.org/10.1145/2207676.2208344
49. Caitlin Kelleher. 2009. Barriers to programming engagement. *Advances in Gender and Education* 1, 1: 5-10.
50. Mary E. Kite, Kay Deaux, and Elizabeth L. Haines. 2008. Gender stereotypes. In *Psychology of women: A handbook of issues and theories* (2nd ed.), Florence L. Denmark and Michele A. Paludi (eds.). Greenwood Publishing Group, 205-236.
51. J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1: 159-174.
52. Clayton Lewis, Peter G. Polson, Cathleen Wharton, and John Rieman. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90), 235-242. http://doi.acm.org/10.1145/97243.97279
53. Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. 2010. State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human-Computer Interaction* 26, 8: 741-85.
54. Jane Margolis and Allan Fisher. 2003. *Unlocking the Clubhouse: Women in Computing*. MIT Press.
55. Nicola Marsden and Maren Haag. 2016. Stereotypes and politics: reflections on personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 4017-4031. http://doi.acm.org/10.1145/2858036.2858151
56. Nicola Marsden and Maren Haag. 2016. Evaluation of gendermag personas based on persona attributes and persona gender. In *HCI International 2016 – Posters' Extended Abstracts: 18th International Conference, HCI International 2016, Toronto, Canada, July 17-22, 2016, Proceedings, Part I*, Constantine Stephanidis (Ed.). Cham: Springer International Publishing, 122-127.
57. Nicola Marsden, Jasmin Link, and Elisabeth Büllesfeld. 2015. Geschlechterstereotype in Persona-Beschreibungen. In *Mensch und Computer 2015 Tagungsband*, Sarah Diefenbach, Niels Henze and Martin Pielot (eds.), Stuttgart: Oldenbourg Wissenschaftsverlag, 113-122.
58. Adrienne L. Massanari. 2010. Designing for imaginary friends: information architecture, personas, and the politics of user-centered design. *New Media & Society* 12, 3: 401-416.
59. Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1219-1228. http://doi.acm.org/10.1145/2207676.2208573
60. Philipp Mayring. 2014. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Beltz, Klagenfurg.
61. Joan Meyers-Levy and Barbara Loken. 2015. Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology* 25, 1: 129-149.
62. Joan Meyers-Levy and Durairaj Maheswaran. 1991. Exploring differences in males' and females' processing strategies. *Journal of Consumer Research* 18, 1: 63–70.
63. Lene Nielsen and Kira S. Hansen. 2014. Personas is applicable: a study on the use of personas in Denmark. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 1665-1674. http:/doi.acm.org/10.1145/2556288.2557080
64. Lene Nielsen, Kira S. Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *International Journal of Sociotechnology and Knowledge Development* 7, 1: 45-61.
65. Anne O'Leary-Kelly, Bill Hardgrave, Vicki McKinney, and Darryl Wilson. 2004. The influence of professional identification on the retention of women and racial minorities in the IT workforce. In *NSF Info. Tech. Workforce & Info. Tech. Res. PI Conf.*: 65-69.
66. Piazza Blog. 2015. STEM confidence gap. Retrieved September 24th, 2015 from http://blog.piazza.com/stem-confidence-gap/
67. John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences* (DUX '03), 1-15. http:/doi.acm.org/10.1145/997078.997089
68. René Riedl, Marco Hubert, and Peter Kenning. 2010. Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of EBay offers. *MIS Quarterly* 34, 2: 397-428.
69. Daniela Rosner and Jonathan Bean. 2009. Learning from IKEA hacking: I'm not one to decoupage a tabletop and call it a day. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 419-422. http://doi.acm.org/10.1145/1518701.1518768

70. Hokyoung Ryu and Andrew F. Monk. 2004. Analysing interaction problems with cyclic interaction theory: Low-level interaction walkthrough. *Psychology Journal* 2, 3: 304-330.
71. Andrew Sears. 1997. Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction* 9, 3: 213-234.
72. Steven J. Simon. 2001. The impact of culture and gender on web sites: an empirical study. *The Data Base for Advances in Information Systems* 32, 1: 18-37.
73. Anil Singh, Vikram Bhadauria, Anurag Jain, and Anil Gurung. 2013. Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets. *Computers in Human Behavior* 29, 3: 739–746.
74. Rick Spencer. 2000. The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '00), 353-359. http://doi.acm/org/10.1145/332040.332456
75. Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? *Design Studies* 32, 1: 30-44.
76. Afshin Vafaei, Beatriz Alvarado, Concepcion Tomás, Carmen Muro, Beatriz Martinez, and Maria Victoria Zunzunegui. 2014. The validity of the 12-item Bem Sex Role Inventory in older Spanish population: An examination of the androgyny model. *Archives of gerontology and geriatrics* 59, 2: 257-263.
77. Gabriela Viana and Jean-Marc Robert. 2016. The practitioners' points of view on the creation and use of personas for user interface design. In *Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I*, Masaaki Kurosu (Ed.). Cham: Springer International Publishing, 233-244.
78. Elke U. Weber, Ann-Renée Blais, and Nancy E. Betz. 2002. A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making* 15, 4: 263-290.
79. Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. 1994. The cognitive walkthrough method: A practitioner's guide. In *Usability Inspection Methods*. John Wiley, NY, 105-140.