Is It Possible to Step into Someone Else's Shoes? An Empirical Study Investigating the Compatibility Between Cognitive Styles and GenderMag Personas

Carolina Yumi Fujii
Federal University of Technology - Paraná (UTFPR)
Campo Mourão, Brazil
carolinafujii@alunos.utfpr.edu.br

Igor S. Wiese
Federal University of Technology - Paraná (UTFPR)
Campo Mourão, Brazil
igor@utfpr.edu.br

Walter T. Nakamura
Federal University of Technology - Paraná (UTFPR)
Campo Mourão, Brazil
waltertakashi@utfpr.edu.br

Abstract

Software development is strongly affected by biases arising from team profiles, which may influence how people perceive, interpret, and interact with interfaces, creating usability barriers and exclusion. In this context, GenderMag is an established method for identifying usability issues related to gender inclusiveness by evaluating software from the perspective of cognitive personas representing different usage styles and behavioral profiles. However, evaluators' ability to effectively embody these personas during inspections, especially considering their own biases and cognitive profiles, remains underexplored. Therefore, this study aimed to analyze the relationship between evaluators' cognitive profiles and their influence on the use of personas during a GenderMag inspection. To this end, a controlled experiment was conducted with 27 Computer Science students from the Federal University of Technology - Paraná (UTFPR). Participants were divided into two groups based on their cognitive profiles; each group performed inspections using both personas, with half starting with Abby and the other half with Tim. After the inspections, the students completed a questionnaire to identify their perceptions regarding the difficulty of putting themselves in the personas' place and their views on GenderMag. The results indicated significant differences between the groups in the evaluation conducted with the Abby persona, but not in the evaluation conducted with the Tim persona. Participants with a Tim-aligned cognitive profile experienced greater difficulty putting themselves in Abby's place, leading to significantly different results than those of the group whose cognitive profile was aligned with Abby. These findings suggest that evaluators' cognitive profiles may influence the application of the GenderMag method, affecting the consistency and effectiveness of usability inspections. Our results contribute to a better understanding of the influence of cognitive profiles on inspections and support proposals for improving the application of the GenderMag technique.

Keywords

Gender inclusion, GerderMag, Usability inspection, Cognitive styles, Human-Computer Interaction

1 Introduction

In software engineering, the composition of development teams is a crucial factor that directly influences the quality and scope of the solutions they deliver. Teams are composed of professionals with distinct profiles, experiences, and worldviews, which can unintentionally introduce biases into software products. These biases, often unconscious, can manifest in functionalities, interfaces, and the system logic itself, reflecting the dominant perspectives of the team that developed it [25].

Among the various types of biases, gender biases have received increasing attention, as studies indicate differences in cognitive styles and problem-solving approaches between genders [4, 17, 30]. If not considered, these biases can result in products that serve a diverse user base unequally, compromising the software's usability and effectiveness for some users. The rapid technological evolution, with the rise of extended reality and the Internet of Things, only amplifies this challenge, making systems increasingly prevalent in daily life and requiring greater caution to ensure inclusion [1].

In this scenario, applying evaluation methods to identify and mitigate these biases throughout the development cycle becomes essential. Usability evaluation approaches play an important role in this process, as they allow for the identification of interaction problems during software development [9]. In this context, the GenderMag method was proposed as a usability inspection technique to identify gender-inclusion issues in software interfaces. The method is an adaptation of the *Cognitive Walkthrough* [24] and utilizes cognitive personas that represent different styles of interaction with technology. These personas are based on five cognitive facets (motivation, information processing style, computer self-efficacy, attitude toward risk, and learning style) that reflect patterns observed in empirical studies on software interaction [5, 16].

Despite its utility, the application of GenderMag directly depends on the interpretation of the evaluators conducting the inspection [25]. Since evaluators have their own experiences and cognitive styles, these characteristics may influence how the personas are represented during the evaluation [28]. Thus, the following research question emerged: Does assuming a GenderMag persona allow evaluators to accurately represent its cognitive profile, or do their answers remain biased by their own mental models?

To investigate this question, we analyzed the relationship between evaluators' actual cognitive profiles and their ability to apply GenderMag personas during usability inspections. Through a controlled experiment involving 27 computer science students, we employed a crossover design in which participants with distinct cognitive profiles performed inspections using both the Abby and Tim personas. By comparing quantitative inspection data with qualitative post-study perceptions, we evaluated the extent to which

individual cognitive characteristics interfere with the faithful representation of these personas and, consequently, with the effectiveness of the usability evaluation process. The primary contributions of this work are as follows:

- **Empirical evidence of evaluator bias:** Our study provides empirical evidence on how the evaluator's cognitive profile influences the application of the GenderMag method. We observed that evaluators often unintentionally attribute their own technical ease and mental models to personas that should exhibit different interaction patterns, highlighting how underlying cognitive biases can compromise usability inspections.
- **Asymmetry in persona simulation:** We identified that while the exploratory Tim persona is easily simulated, the risk-averse Abby persona poses a significantly higher cognitive challenge for evaluators, especially for those with a different profile.
- **Methodological insights for inclusive design:** We offer practical recommendations for improving the reliability of GenderMag, such as the need for cognitive diversity within evaluation teams and enhanced training to mitigate the "evaluator effect."
- **Differentiated perception of interface cues:** Our results suggest that distinct cognitive profiles prioritize different interface elements. While exploratory profiles (Tim) accept simple visual cues as sufficient proof of progress, risk-averse profiles (Abby) demand explicit feedback and clearer terminology to mitigate uncertainty.

2 Background

2.1 Gender Bias in Software

Software development is influenced not only by technical factors but also by social and cultural aspects. Professionals involved in the design and implementation process bring their own experiences, beliefs, and values, which can impact design decisions and, consequently, how interfaces are constructed [23]. As a result, systems may end up favoring certain user profiles over others.

In the context of gender inclusion, studies indicate that digital interfaces may reflect biases by assuming that all users possess the same cognitive styles, motivations, or levels of familiarity with technology [5]. These biases can arise early in development and are often exacerbated by the lack of diversity within engineering and design teams. Consequently, users who do not match the presumed profile may face usability barriers or interaction difficulties.

Given this scenario, promoting diversity in software development is not merely a social issue but also a strategy to improve system quality. More diverse teams tend to better understand the needs of different users and produce more inclusive solutions [18].

2.2 GenderMag

GenderMag (Gender Inclusiveness Magnifier) is a usability inspection method developed to identify problems related to gender inclusion in software interfaces [5]. The method is an adaptation of the *Cognitive Walkthrough* and uses cognitive personas to represent different styles of interaction with technology, thereby helping create more inclusive software [14]. It is based on five cognitive dimensions, called facets, which reflect patterns observed in empirical studies on software interaction [16]: (i) motivation for using technology, (ii) information processing style, (iii) computer self-efficacy, (iv) attitude toward risk, and (v) learning style for new technologies. These dimensions are operationalized through personas that represent different cognitive user profiles.

2.2.1 GenderMag Personas. GenderMag uses three main personas to represent different styles of interaction with technology [5]:

- **Abby**: a results-oriented user with lower confidence in unfamiliar technologies, who prefers learning through structured instructions and tends to avoid risks.
- **Pat**: an intermediate individual with a moderate level of confidence in technology and a more reflective approach to exploring systems.
- **Tim**: a user with high technological self-efficacy, a greater propensity for risk, and an exploratory learning style.

These personas enable evaluators to analyze usage scenarios from different cognitive perspectives, facilitating the identification of usability barriers that may affect specific user groups.

2.2.2 GenderMag Application. During the method's application, evaluators walk through usage scenarios from the personas' perspectives and answer a series of structured questions about their interactions with the system. This process allows for the identification of potential "gender bugs," that is, usability obstacles that may impact users with different profiles in different ways, that might otherwise go unnoticed, even when the inspector does not share the same gender identities as the personas [16].

3 Related Work

The software engineering community has long recognized that development and testing processes are not purely mechanical, but are deeply influenced by human factors. The effectiveness of software inspection methods, for instance, is intrinsically linked to the performance of the human inspectors who apply them. Thus, understanding the factors that influence this performance is crucial for improving the reliability of quality assurance processes [26].

The subjectivity inherent in usability and accessibility inspections is well documented as the *evaluator effect*. Jacobsen et al. [12] and Hertzum and Jacobsen [11] provided foundational evidence that different evaluators, even when using the same method on the same interface, tend to identify substantially different sets of problems and vary in their severity judgments. Similarly, Ling and Salvendy [13] have shown that the evaluator's cognitive style can affect how heuristics are applied and how problems are recognized during an inspection. The influence of inspectors' expertise has also been investigated. Yesilada et al. [32] and Brajnik et al. [3] investigated how the professional background of inspectors, comparing experts and non-experts, affects the identification of accessibility barriers. Their findings suggest that while technical expertise enhances the detection of certain flaws, it does not eliminate the subjectivity inherent in the evaluation process, indicating that other individual characteristics may be at play.

Recent research has shifted focus toward the underlying cognitive mechanisms that influence software practitioners. Mohanani

et al. [15] conducted a systematic mapping study that identified dozens of cognitive biases affecting various software engineering activities, from requirements gathering to testing. One such bias is the *confirmation bias*, which Çalıklı and Bener [6] demonstrated can significantly impact software quality, as developers often seek information that confirms their existing mental models rather than objectively testing for defects. Furthermore, individual differences in cognitive processing have been explored in specific technological contexts. Belk et al. [2] showed that users' cognitive styles directly influence their interaction with security-related tasks, such as authentication. This suggests that cognition is not only a factor for the end-user but also a critical variable for the practitioner designing or evaluating the system.

Despite extensive literature on the evaluator effect related to expertise [3, 12] and the identification of cognitive biases in general software engineering tasks [6, 15], there is a gap in the literature regarding persona-based inspection methods. To the best of our knowledge, this is the first study to investigate the intersection of an evaluator's cognitive profile and their ability to perform a persona-based inspection. While existing literature focuses on expertise or user behavior, our work addresses how the cognitive distance between the evaluator and the persona affects the reliability of gender-inclusive methods such as GenderMag. Our study builds upon the premise that if cognitive styles influence interaction [2] and biases affect technical judgment [6], then the cognitive distance between the inspector and the persona may be a hidden variable in the GenderMag method. By investigating this intersection, we aim to extend the understanding of the evaluator effect from a strictly technical perspective to a deeper cognitive dimension.

4 Methodology

This study employed a controlled experiment in an academic setting to investigate evaluators' ability to adequately represent the GenderMag method personas during usability inspections. We adopted a paired-comparison (cross-design), with one factor and two treatments [31]. Participants were initially divided into two distinct groups based on their cognitive profiles. Each participant performed usability inspections with both GenderMag personas (Abby and Tim). To mitigate order effects, within each group, half of the participants began the inspection with the Abby persona, while the other half started with the Tim persona.

4.1 Procedure

The experiment was designed to compare participants with distinct cognitive profiles as they analyze an interface from the perspective of different personas. Usability evaluation tasks were conducted, and data-collection instruments were used to analyze both task performance and participant perceptions. This approach allowed identification of differences in task execution and in how each group interpreted the persona's characteristics during the analysis. Figure 1 illustrates the steps of the research.

The study was conducted with third-semester students of the Bachelor of Computer Science program at UTFPR, within the context of the Human-Computer Interaction course. Before conducting the experiment, two introductory classes on the method were taught, covering the cognitive facets, application examples, and a practical activity. This stage aimed to ensure a minimum level of understanding of the method, thereby reducing potential difficulties arising from unfamiliarity with the technique.

First, participants completed a characterization questionnaire to obtain background and social demographic information. They also completed an informed consent form that explained the study's goals, their voluntary participation, and how the data would be handled and analyzed. Next, they completed the Cognitive Profile Questionnaire [30] to identify which of the personas they are more aligned with. The responses were aggregated by facet and analyzed using the median, as recommended by previous studies [5, 10]. Based on these medians, we categorized each participant into the most compatible persona (either Abby or Tim). For each of the five cognitive facets, participants with scores below the median were classified as Abby for that facet, whereas those with scores above the median were classified as Tim. Scores equal to the median were classified as Pat. After all five facets had been classified individually, participants were assigned to the group whose combination of facet classifications showed the greatest compatibility with either Abby or Tim. In the few cases where participants presented an equal distribution of facet classifications (e.g., Abby, Tim, and Pat), they were allocated to the groups to preserve sample balance [10].

After identifying their profiles, we divided the participants into two groups: Group A, comprising those more aligned with Abby's cognitive style, and Group B, comprising those more aligned with Tim. Within each group, we randomly assigned participants to start the inspection with either Abby or Tim. This strategy was adopted to counterbalance and systematically mitigate potential order, learning, or fatigue effects, ensuring that the experience gained in the first inspection did not bias the results of the second.

On the experiment day, the participants performed the inspection at a computing lab. The participants received a description of the personas and the links to the inspection in Google Forms. After completing the inspection with both personas, participants completed a post-study questionnaire to assess their perceptions of the inspection process with each persona, the adequacy of the scenarios for these personas, and the GenderMag method. All activities were conducted in the classroom during regular course hours, and the content covered was integrated into the curricular components outlined in the course syllabus.

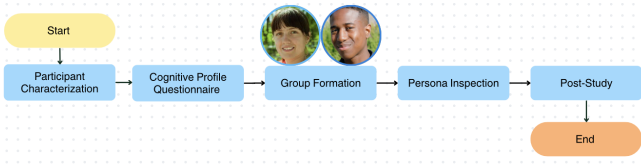

Figure 1: Overview of the research procedure.

4.2 Artifacts

To conduct the experiment, three main artifacts were used: (i) a cognitive profile identification questionnaire, (ii) inspection forms based on the GenderMag personas, and (iii) a post-study questionnaire to evaluate the participants' perception of the process. The artifacts are available in [8].

4.2.1 Cognitive Profile Questionnaire. The GenderMag Facets Survey [30] consists of 20 statements evaluated on a 9-point Likert scale, where 1 corresponds to "strongly disagree" and 9 to "strongly agree."

4.2.2 Inspection Forms with Personas. In this stage, the participants performed two usability inspections using the Abby and Tim personas as references. All participants received the standard description and profile of each persona as available on the GenderMag project website[1], translated into Brazilian Portuguese.

The analyzed scenario was the file submission workflow in the Moodle Learning Management System. The task was divided into two main sub-goals: adding the activity file and finalizing the activity submission. Figure 2 presents one of the interfaces used during the execution of the task.

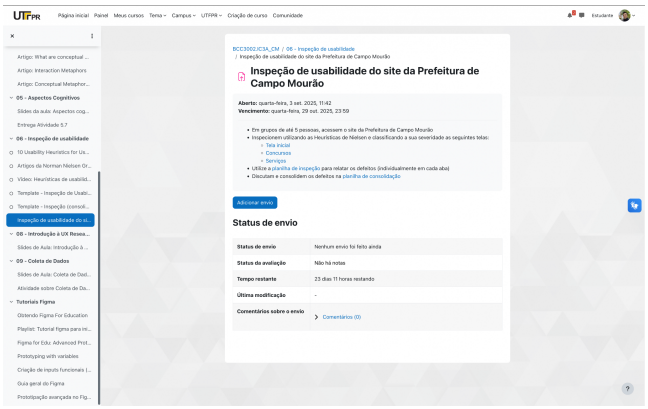

Figure 2: Interface for Sub-objective 1: Add the activity file.

Each step was accompanied by screenshots of the interface and structured questions from the GenderMag method, organized according to the inspection's cognitive flow. The forms were identical in content and structure, differing only in the persona used as a reference, allowing comparisons of interpretation and decision-making when adopting distinct cognitive profiles. For each action, participants evaluated whether the persona would perform it (Yes, Maybe, or No) and justified their answers using the method's cognitive facets. Additionally, they indicated which facets were considered in the decision and which interface elements contributed to or hindered the persona's understanding of the task.

After performing each action, participants were presented with a new screenshot, along with "post-action" questions asking whether the persona would know they are doing the right thing and making progress toward their goal, and why. This process allowed analysis of how different cognitive profiles interpret and make decisions when interacting with the interface, especially regarding perceptions of progress and feedback.

4.2.3 Post-Study Questionnaire. After completing the inspections, participants responded to a post-study questionnaire to evaluate their overall experience with the method and their perceptions of each persona. The instrument included quantitative scale-based questions, in which participants indicated their level of ease in thinking like Abby and Tim, as well as open-ended questions designed to explore the suitability of the scenario for the cognitive profiles, the main challenges faced during the task, and suggestions for improving the use of GenderMag.

[1]https://gendermag.org/custom_persona.php

4.3 Data Analysis

The results of the experiment were analyzed quantitatively and qualitatively, enabling an understanding of both the objective differences between groups and participants' subjective perceptions while using the GenderMag method.

In the quantitative analysis, the frequencies of "Yes," "Maybe," and "No" responses to the 14 inspection form questions were considered. Each participant performed two inspections (one with the Abby persona and one with the Tim persona), and their responses were compared between the two groups previously formed based on their cognitive profiles (Group A and Group B). We also applied statistical tests to assess statistical significance and effect size. First, we assessed the normality of the distributions using the Shapiro–Wilk test, which is recommended for small samples. Then, we applied the appropriate test based on the data distribution.

For the qualitative analysis, we used Thematic Analysis to interpret participants' open-ended responses in the post-study. The textual responses were read in their entirety, coded, and grouped into recurring themes, allowing for the identification of significant patterns in the participants' perceptions. This process was initially performed by one of the study's authors and subsequently reviewed by another researcher with experience in qualitative analysis. The identified discrepancies were discussed until consensus was reached, resulting in refined codes and greater consistency in the definition of the final themes.

5 Results

5.1 Participant's Profile

The research sample consisted of 27 participants, of whom 8 were female and 19 were male. Regarding age, a higher concentration was observed within the typical age range of university students, with 68% of participants aged 19-21. Regarding occupation, 23 participants identified as students only. Other occupations included full-stack developer (2), IT technician (1), and software engineer (1). Regarding education level, 26 participants were undergraduate students, and one was a graduate student.

All participants had introductory knowledge of usability stemming from the course content, including evaluation heuristics, user-centered design principles, and general notions of interface inspection. Only a small portion, equivalent to 7.4%, claimed to have some experience in real projects. Regarding the GenderMag method, none of the participants had prior experience with its application outside the academic context. All had contact with the method exclusively through practical activities carried out in the classroom, which provided a similar level of familiarity among participants and contributed to a more balanced execution of the experiment.

5.2 Cognitive Profile of Participants

Of the 27 participants, 12 were classified as having a profile closer to Abby, 13 to Tim, and 2 as having mixed characteristics, i.e., two

facets related to Abby, two with Tim, and one with Pat. These two participants were distributed equally between the two groups to maintain sample balance. Thus, subsequent analyses were conducted with 13 participants in the Abby group and 14 in the Tim group.

5.3 Cognitive Profile vs. Persona Used

To compare the results between the two groups, we first analyzed the data distribution. The Shapiro-Wilk test indicated that for all analyzed variables (Yes, Maybe, and No), at least one of the groups presented a non-normal distribution, with p-values < 0.05, evidencing a violation of the normality assumption (data presented in Tables 1 and 2). Thus, we selected the non-parametric Mann-Whitney U test to compare the responses between groups A and B.

Table 1: Descriptive statistics for evaluation with Abby.

	Yes		Maybe		No	
	A	**B**	**A**	**B**	**A**	**B**
Valid	13	14	13	14	13	14
Missing	0	0	0	0	0	0
Mean	7.615	10.29	4.923	2.857	1.462	0.857
Std. Deviation	3.501	2.367	2.691	2.107	1.713	1.406
Shapiro-Wilk	0.918	0.925	0.947	0.921	0.757	0.638
P-value of Shapiro-Wilk	.238	.260	.551	.228	.002	< .001
Minimum	2.000	7.000	1.000	0.000	0.000	0.000
Maximum	13.00	14.00	9.000	6.000	5.000	4.000

Table 2: Descriptive statistics for evaluation with Tim.

	Yes		Maybe		No	
	A	**B**	**A**	**B**	**A**	**B**
Valid	13	14	13	14	13	14
Missing	0	0	0	0	0	0
Mean	12.31	11.93	1.692	1.929	0.000	0.143
Std. Deviation	1.843	2.018	1.843	1.859	0.000	0.363
Shapiro-Wilk	0.849	0.847	0.849	0.854	NaN[0]	0.428
P-value of Shapiro-Wilk	.028	.020	.028	.025	NaN[0]	< .001
Minimum	8.000	7.000	0.000	0.000	0.000	0.000
Maximum	14.00	14.00	6.000	6.000	0.000	1.000

[0] All values are identical

The results demonstrated significant differences between the groups when participants inspected using the Abby persona, but not when they used the Tim persona. These differences become more evident when observing the boxplots for the "Yes," "Maybe," and "No" responses. In the condition where participants used the Abby persona (Figure 3), it is noted that Group B (Tim profile) presented a significantly higher frequency of "Yes" responses compared to Group A (Abby profile) ($U = 48$, $p = .038$), with a large effect size ($r_{rb} = 0.473$). This result suggests that participants with the Tim profile, characterized by a more direct, solution-oriented approach, tended to believe that Abby would complete the actions easily, as they would. This finding highlights the difficulty these participants faced in detaching from their own mental models to simulate Abby's more cautious and detail-oriented nature.

Regarding the "Maybe" responses, Group A presented a significantly higher frequency than Group B ($U = 132$, $p = .048$), also with a large effect size (r_{rb} = -0.451). This inclination toward uncertainty suggests that Group A participants, because they possess a style closer to the inspected persona, were more sensitive in identifying interface ambiguities that would cause hesitation in Abby. Thus, "Maybe" functioned as an indicator that these evaluators perceived that Abby would not necessarily fail the task but would experience doubt or slower decision-making before acting.

Regarding the "No" responses, Group A was more likely to believe that the Abby persona would not be able to perform the actions. However, the difference was not statistically significant ($U = 118$, $p = .169$). A possible explanation for this result is that Group A's perception of difficulties ended up concentrating in the "Maybe" category, indicating that interface problems were seen as points of uncertainty rather than impossible barriers to overcome. Furthermore, the low occurrence of "No" responses in both groups suggests that the usability problems found were not severe enough to generate a consensus of total failure, which kept the groups' results statistically close despite the observed tendency.

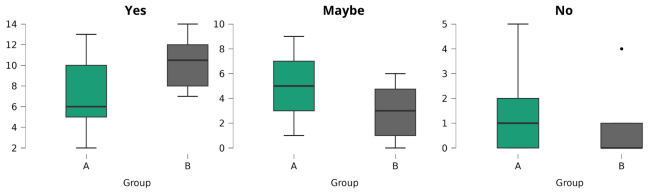

Figure 3: Distribution of responses for the Abby persona.

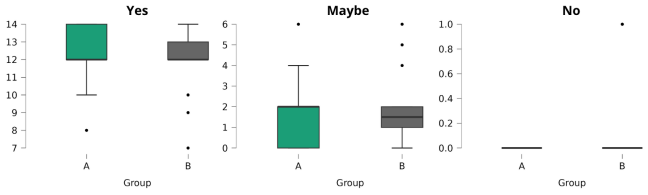

Figure 4: Distribution of responses for the Tim persona.

On the other hand, when participants performed the inspection using the Tim persona (Figure 4), no significant differences were observed between the groups ($p > .05$ across all responses). This result suggests that both Abby-profile and Tim-profile participants were able to represent Tim similarly. One possible interpretation is that the Tim persona (more exploratory, spontaneous, and experimentation-oriented) may have been perceived as more intuitive or easier to simulate, even for individuals who do not fully share this cognitive style. Another possibility is that the analyzed scenario favored the exploratory approach typical of Tim, thereby reducing the impact of cognitive differences between the groups.

Thus, the results indicate that the inspector's cognitive profile interferes with the application of GenderMag, but in an asymmetrical way. Simulating the Abby persona is more difficult when the evaluator has an opposing cognitive style, whereas representing Tim requires less cognitive effort and is more natural for most participants. These findings reinforce the importance of considering

the evaluators' cognitive profiles when applying GenderMag, as identification with the persona can directly influence the decisions made during the inspection process.

5.4 Putting Oneself in the Persona's Shoes

The post-study questionnaire revealed differences in the ease with which participants were able to put themselves in the place of the Abby and Tim personas during the application of the GenderMag method. As shown in Figure 5, Group B (Tim profile) found it more difficult to represent the Abby persona, with 71.4% of participants indicating some degree of difficulty. Participants from Group A (Abby profile) also found it difficult to put themselves in Abby's shoes, but to a lesser extent (53.9%).

On the other hand, representing Tim was perceived as a relatively easy task by both groups (Figure 6). A curious point is that Group A (Abby profile) reported greater ease in putting themselves in Tim's place (84.7%) than Group B itself (78.6%), which shares the persona's profile. This may indicate that Tim's exploratory behavior is seen as a "navigation ideal" or a simplified pattern of use, making its simulation less exhausting. While Abby requires the evaluator to anticipate problems and risks, Tim allows a more relaxed, reactive posture, which explains why even those with an opposite style feel comfortable in his role.

These perceptions reinforce the idea that there is an unequal cognitive load when applying GenderMag. The difficulty Group B reported when simulating Abby helps explain the previous quantitative results. As they found it harder to "get into character," they ended up projecting their own ease of use into their answers. Thus, the success of the method depends not only on the chosen persona but also on the capacity and effort the evaluator is willing to exert to break away from their own way of thinking.

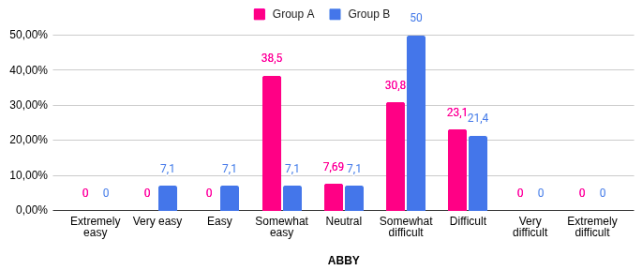

Figure 5: The perceived ease of adopting Abby's perspective.

The results of the qualitative analysis reveal an asymmetry in the cognitive effort required for persona representation, suggesting that the effectiveness of the inspection is intrinsically linked to the proximity between the evaluator's profile and the persona's facets. Figure 7 presents the codes and themes resulting from the analysis of the open-ended questions regarding participants' perceptions of the possibility of assuming the persona's role during the inspection.

In the case of the Abby persona, representation is predominantly hindered by her own intrinsic attributes, such as a more conservative, insecure, and anxious nature. For evaluators who do not share this profile, a rupture emerges between the "self" and the persona, imposing a high mental effort that results in overthinking and difficulty predicting actions and reactions in uncertain environments.

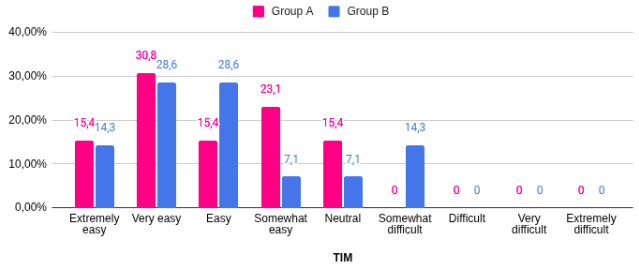

Figure 6: The perceived ease of adopting Tim's perspective.

Participant P2, for instance, reports that "evaluating the interface as Abby involves many 'what ifs.' There are operations that seem very obvious and logical, but as Abby, I needed to stop and reflect on whether there were counter-indicators for performing each action, things that would make me afraid of making a mistake. Inspecting as Abby meant adopting an insecure stance toward the software." This quote indicates that such a scenario forces the inspector to adopt a position of constant vigilance, having to "brake" their own intuition to emulate Abby's hesitation. Curiously, this same insecurity and the persona's step-by-step exploration method serve as facilitators for two participants, providing predictability in Abby's behavior. According to these participants, understanding her risk aversion enables the inspector to anticipate where interface barriers will be most critical.

In contrast, the Tim persona is perceived as significantly more intuitive to represent. This perception reflects attributes that facilitate fluidity in inspection, such as his high technological self-efficacy and tendency to learn by trial and error. Tim's exploratory behavior and confident profile create predictability, which simplifies the evaluator's task and allows for a freer, less taxing navigation. Much of this ease stems from direct identification with the persona or Tim's similarity to the profiles of people close to the participants, making the simulation's mental model more familiar. However, this proximity also presents challenges, as it may lead to the projection of one's cognitive profile during the inspection. Participant P23, for example, reports a similarity between Tim and their own cognitive profile, making it "much easier to imagine how it would be in these situations," while simultaneously feeling the need to "maintain a bit of extra caution to ensure it was still Tim interacting with the site" and not the participant themselves.

These findings reinforce the quantitative analysis, showing that applying GenderMag is not a neutral process and depends on the cognitive distance between the evaluator and the persona being represented. While Abby requires the construction of a cautious mental model that is attentive to details the evaluator usually ignores, Tim requires an exercise of control so that familiarity does not mask potential difficulties that could affect the persona.

5.5 Perception of Scenario Suitability According to Cognitive Style

The analysis of responses to the scenario suitability question revealed unequal perceptions of adequacy for the two personas. Figure 8 reveals a mixed perception regarding suitability for Abby,

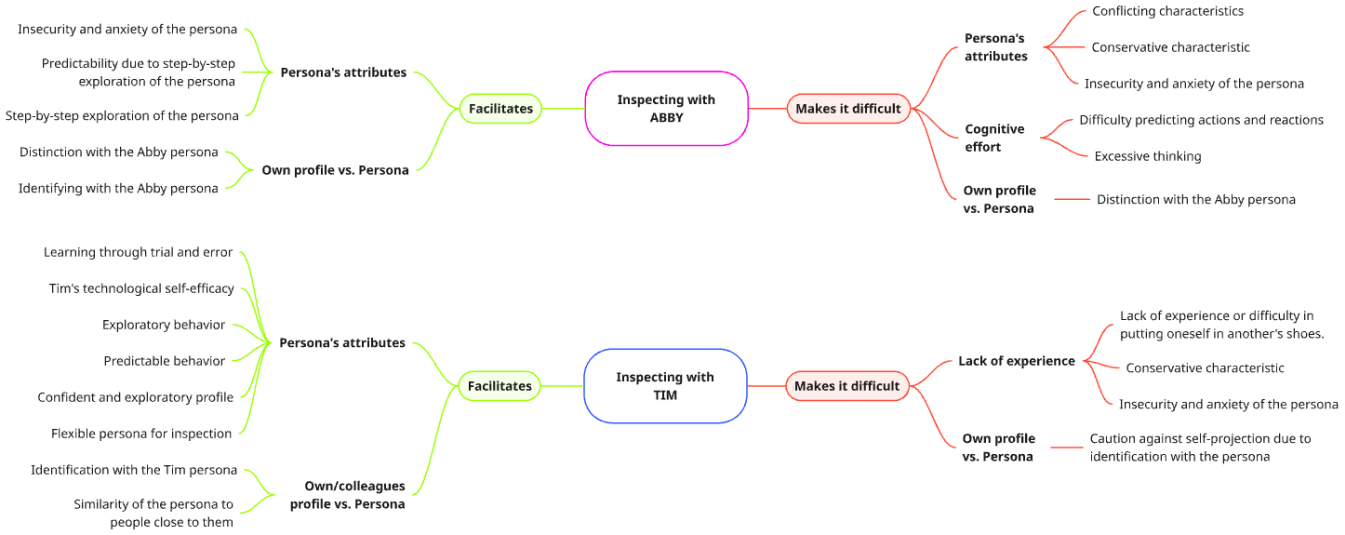

Figure 7: Themes and codes emerged from the qualitative analysis on perceptions about putting oneself in the persona's shoes.

even within the group of participants with a similar cognitive profile (Group A). This result highlights the difficulty in assessing whether a given interface or set of actions meets the needs and characteristics of users with a cognitive style similar to Abby's. Participants from both groups justified the low scores by pointing to the interface's lack of support for dealing with the "unknown." Some participants from Group A emphasized that the Moodle interface is "too simple," failing because it does not include explanatory texts that would assist Abby during her first contact with the task. Meanwhile, Group B emphasized that the absence of explicit confirmation messages and the lack of clarity in buttons and labels generate insecurity and fear of making mistakes. Overall, for the participants, Abby's effectiveness in the scenario depends on an interface that safely guides the user toward the correct path, which was not observed in the analyzed context, where elements such as the "Save changes" button were described as ambiguous.

On the other hand, Figure 9, referring to Tim, shows a more homogeneous distribution, with a predominance of perceptions that the scenario and tasks are suitable for this profile across both groups. Participants in Group A observed that, even in the face of interface defects or unclear titles, Tim's behavior of "just jumping in and testing" makes any obstacle surmountable. According to Group A, Tim's prior knowledge of technology and his familiarity with system logic allow him to quickly assimilate functions, even without detailed instructions. Group B also corroborated this view, describing Tim as a curious user who remains engaged through the exploration of functionalities. His trial-and-error method of exploration makes him more resilient to ambiguities.

The variation in responses regarding the Abby persona highlights the subjective nature of emotional states such as insecurity and risk aversion. Unlike Tim's exploratory profile, which is guided by a trial-and-error logic that is easier to measure, Abby's reactions depend on how each individual interprets the level of hesitation the interface generates. What one participant from Group A considers an acceptable obstacle, another may interpret as a total stopping

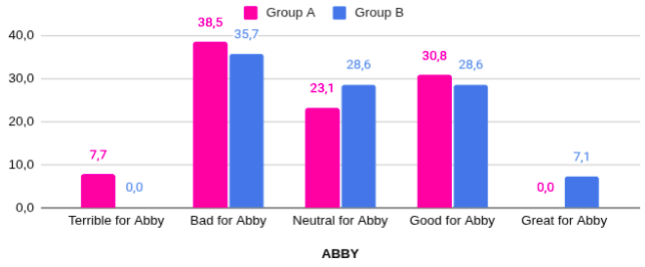

Figure 8: Perception of the suitability of the scenario for Abby.

Figure 9: Perception of the suitability of the scenario for Tim.

point, making Abby's evaluation significantly more subjective and less consensual than Tim's.

To deepen understanding of the differences in focus between the two groups, we analyzed responses to the question of whether Abby would know she did the right thing and is progressing toward her goal in "*Action 5: Click on 'Submit this file*"' (Figure 10). We selected this item because it showed the greatest variation in responses between the two groups.

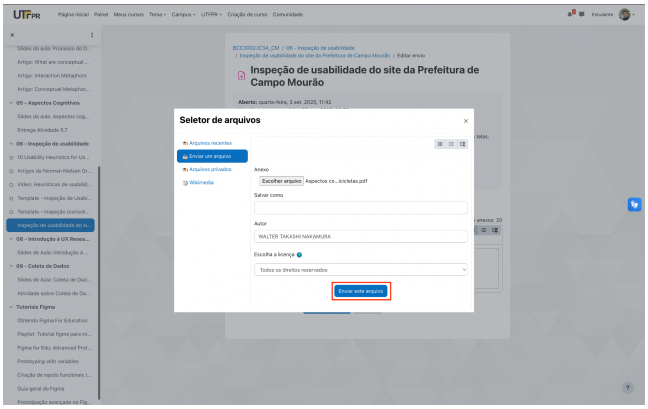

Figure 10: Screenshot regarding the 'Submit this file' action.

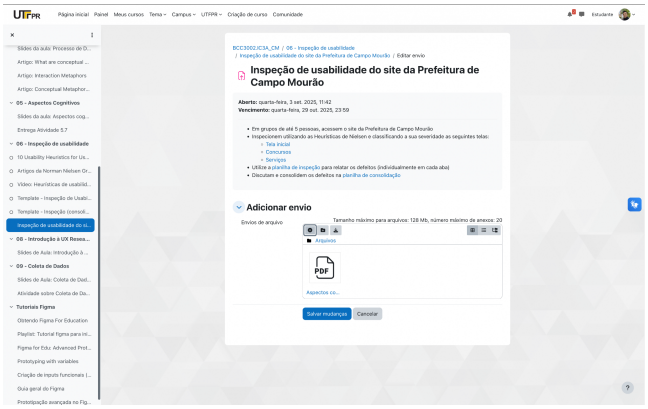

Figure 11: Screenshot after clicking on the "Submit this file" button.

The analysis revealed discrepancies among participants regarding whether Abby would know she was progressing toward her goal after performing the action. Participants in Group A (Abby profile) had greater difficulty considering the interface sufficient to confirm progress, frequently pointing out the need for more explicit feedback or clearer success messages. This behavior aligns with the cognitive characteristics associated with the Abby persona, such as higher risk aversion and lower tolerance for ambiguity.

The thematic analysis reinforced this pattern by revealing two themes in the Abby persona inspection: *Signs of Progress and Confirmation* and *Uncertainty and Need for Support*. Within the first theme, the file name stood out as the primary success indicator, followed by partial confidence generated by the file icon and the visibility of progress upon viewing the attachment. However, these elements were not considered sufficient to completely eliminate insecurity. In the second theme, codes frequently emerged related to the lack of initial guidance, recurring doubts about the submitted content, confusion with the "Save changes" button (see Figure 11), the need for clearer terminology, and persistent uncertainty even after viewing the file. These findings demonstrate that Abby tends to seek more explicit confirmations and additional support before considering the task complete.

Conversely, participants in Group B (Tim profile) tended to interpret the scenario more directly, considering the presence of the attached file on the interface as a sufficient indication of progress. This type of interpretation reflects typical characteristics of the Tim persona, such as greater exploratory confidence and higher tolerance for ambiguity during interaction.

Similarly, the thematic analysis of the Tim persona also revealed the themes *Signs of Progress and Confirmation* and *Uncertainty and Need for Support*, yet with distinct interpretations. Among the signs of progress, the visibility of progress when viewing the file was most frequently highlighted, alongside the recognition of the screen as a familiar element, the file name as a success confirmation, and the perception of step-by-step learning during the interaction. Regarding the theme of uncertainty, although there were mentions of seeking visual security and of expectations for submission feedback, these manifestations were less intense and accompanied by greater acceptance of ambiguity. This indicates that Tim can advance with more confidence even without explicit confirmations, using simple visual cues as sufficient evidence of progress.

Despite this general pattern, an atypical case was observed in Group B, in which a participant responded similarly to the Abby profile, indicating uncertainty regarding the action's conclusion. This result suggests that, although the cognitive profile influences evaluations, it does not completely determine participant behavior, as cognitive styles can manifest more flexibly in certain contexts.

Overall, the results indicate that the cognitive facets associated with the personas influence how participants interpret interface feedback. While the scenario seemed more intuitive for the Tim profile, participants with characteristics closer to Abby demonstrated a greater need for explicit confirmation of progress. This finding reinforces the importance of considering cognitive differences when applying the GenderMag method, as identification with the persona can directly influence the evaluations made during the inspection.

5.6 Perceptions of GenderMag

The participants' post-study perceptions of GenderMag highlight both its positive aspects and limitations, which can be synthesized into three themes: structure, cognitive load, and potential for refinement. In general, they emphasized that the methodological clarity and the step-by-step inspection script facilitate the organization of reasoning, helping to understand the interface's action flow in a logical manner. This perception of utility is reinforced by the ease of personification enabled by the guided questions, which encourage a more careful analysis and orient the evaluation toward different cognitive styles, making the inspection process more tangible for the evaluator.

However, this structuring entails practical challenges, especially regarding the cognitive effort required to detach from the "self." Participants reported a recurring difficulty in distancing themselves from their own technical knowledge to faithfully represent the personas, with an involuntary return to their own way of thinking being common when answering questions. This became more challenging when viewed from Abby's perspective. Participant P5, for

instance, reported that due to her low self-confidence, they struggled to choose between "Yes," "No," and "Maybe," i.e., whether it was a fear or a lack of information that would lead to a "no," or just hesitation, that would lead to a "maybe." Additionally, perceived ambiguity at certain stages and the repetition of questions throughout the process were identified as factors that increase cognitive load, making the inspection long and exhaustive and potentially compromising the precision of the method's application.

In light of these points, the improvement opportunities suggested by the participants focus on process optimization and greater instructional and didactic support. The suggestions include enriching the personas' profiles with more detailed descriptions and developing more objective instructions that include practical examples to facilitate the correlation between cognitive facets and design decisions. Ultimately, although GenderMag is perceived as a robust and structured method, the results indicate that strengthening the clarity of the personas and reducing redundancies in the steps are essential paths toward making its application more fluid, accessible, and efficient for new evaluators.

6 Discussion

The results of this study demonstrate that the cognitive profile of evaluators directly influences the application of GenderMag, affecting both the judgment of persona actions and the perceived ease of adopting each profile. It was observed that evaluators with a cognitive profile closer to Tim's experienced greater difficulty analyzing the system from the perspective of the Abby persona, suggesting a greater effort was required to understand behaviors and decisions distinct from their own cognitive patterns. These findings can be explained through the lens of cognitive biases that affect decision-making and perspective-taking, specifically social projection, egocentric bias, and anchoring bias.

The phenomenon of social projection [19] occurs when individuals assume others share their problem-solving approaches. It is closely related to the egocentric bias, which describes the tendency to over-rely on one's own perspective when interpreting others [20]. In our study, Tim-aligned evaluators likely projected their high technological self-efficacy onto Abby, assuming she would easily overcome interface barriers and thus underestimating usability issues for risk-averse profiles. This is compounded by the anchoring bias [29], as evaluators rely on their own mental models as a starting point. Despite being instructed to simulate Abby, Tim-profile evaluators likely adjusted insufficiently from this anchor [7], leaving their judgments influenced by their exploratory mindset. Although recent studies indicate that actively adopting another person's perspective helps overcome egocentric bias in visual tasks [22] and in manual actions [21], this flexibility does not seem to carry over to the software inspection context, where simulating an entirely different cognitive profile imposes substantial cognitive load, allowing egocentric bias to persist even when evaluators are explicitly instructed to adopt another perspective.

The inspector's cognitive profile can distort results, leading to evaluations that reflect less of the persona's expected behavior and more of the evaluator's actual behavior. As a consequence, an analysis supposedly guided by the persona's perspective may, in practice, reproduce individual biases. This implies that different evaluators may reach distinct conclusions, not because they interpreted different aspects of the interface, but because they possess different cognitive styles. Although GenderMag was designed to mitigate inequalities and reduce biases in software development, our results suggest that the method itself may also be subject to these biases, as the evaluator, while attempting to "embody" the persona, still carries their own cognitive tendencies. Thus, a practical consequence is that GenderMag may not reach its full potential if the inspector's cognitive profile is not taken into account. In real software development environments, this interference can lead to the failure to identify usability problems related to distinct cognitive profiles, thereby perpetuating the very biases the method aims to correct. If a team composed mostly of profiles similar to Tim's evaluates a system, for example, it is possible that difficulties faced by users closer to Abby's profile may not be adequately identified; a situation that aligns with literature on gender inequality in technology. Based on the findings of this study, we outline several implications and recommendations for practitioners and researchers.

6.1 Implication for Practitioners

Our findings have direct and practical implications for software engineering practice. While automated UI testing is effective for verifying the functional correctness of user interfaces, it cannot determine whether interface elements are understandable, intuitive, or appropriate for users with different cognitive profiles. Persona-based inspections fill this critical gap in the quality assurance process. When a problem is identified during these inspections, it is documented as a bug report or issue, creating technical debt that the development team must address and incorporate into the software maintenance cycle.

Given this impact on the development workflow, our results reinforce the critical importance of building cognitively diverse development and evaluation teams. When teams are homogeneous, the predominant cognitive biases can unintentionally influence the modeling of features and the system's underlying logic. Consequently, these teams may overlook usability defects that disproportionately affect users with different cognitive characteristics.

Furthermore, it is necessary to mitigate evaluator bias during the quality analysis stage. Our results show that inspection outcomes can be compromised by the cognitive distance between the inspector and the persona being considered, as practitioners often struggle to set aside their own technical expertise. To address this, practitioners should utilize personas that are less superficial; instead of merely listing general difficulties, personas should detail specific domains of competence, such as which older technologies the user is familiar with. This depth helps inspectors ground their evaluations in what the user would actually do, particularly for users who lack confidence or avoid taking risks.

Finally, integrating persona-based inspections into fast-paced industrial workflows, especially within agile contexts, requires reducing the inspector's cognitive workload. Participants described the evaluation process as massive and repetitive, which can hinder its practical adoption. To overcome this, organizations need a dedicated support infrastructure and tools that optimize workflows. For instance, evaluation systems could continuously display persona information to prevent context switching, and similar sub-goals

could be grouped to increase efficiency. AI-driven solutions to automate preliminary analyses would also make inclusive software inspections truly scalable across the industry.

6.2 Implication for Researchers

For the research community, our results highlight that the empathy gap is a challenge when using GenderMag. We found that the participants had difficulty putting themselves in the shoes of personas like Abby, who are very cautious and have characteristics different from their own. This highlights the need to develop new ways of training people or to provide cognitive scaffolding. There is an opportunity to explore how instructional supports or contextual tips can help the researcher better interpret the subjective motivations behind each persona's actions, reducing ambiguity in matters that seem trivial to an expert but complex to the end user. Future research could investigate how richer narratives or AI-driven support might help evaluators keep their own experiences and biases from getting in the way of their analysis of gender issues.

In line with this, another promising research avenue lies in automating the Cognitive Walkthrough process, potentially using AI. Given that the participants perceived the process as repetitive and overwhelming, researchers could investigate how Large Language Models (LLMs) might assist in generating justifications or performing preliminary analyses of the interface. This could make GenderMag more scalable and better suited to projects in industries.

Finally, our findings also point to an opportunity to refine the GenderMag method's instruments. The current set of answers (Yes, No, Maybe) was perceived by some participants as insufficient to understand how characters like Abby make decisions, especially in situations involving hesitation or a lack of information. Researchers could explore the impact of more granular scales, such as "Probably Yes" or "Probably No." This scale would allow researchers to distinguish between an inherently intuitive design and one that the user can barely navigate, demonstrating great hesitation. By capturing these degrees of certainty, researchers could identify cases where a cautious persona like Abby can technically succeed but is just one small obstacle away from abandoning the task altogether.

7 Threats to Validity

This section discusses the study's potential threats to validity and the mitigation strategies we have in place.

Conclusion Validity. The sample size ($n = 27$) may limit statistical power. To mitigate this, we employed the Shapiro-Wilk test and non-parametric tests (Mann-Whitney U) appropriate for our data distribution. Qualitative reliability was ensured through Thematic Analysis with cross-researcher triangulation.

Internal Validity. To mitigate learning effects from performing two consecutive inspections, we used a counterbalanced crossover design for the persona order. Fatigue was minimized by sizing tasks to fit a regular class period. Furthermore, prior GenderMag training could influence inspection behavior. We mitigated this by providing standardized training (identical instructor, materials, and exercises) to all participants, although we cannot entirely rule out its underlying influence on how personas were interpreted.

Construct Validity. Cognitive profiling relied on the validated GenderMag Facets Survey, though subjective Likert scale interpretations may vary. Excluding the Pat persona (to maintain sample balance) limits the breadth of cognitive diversity analyzed. To minimize inspection bias, we used identical forms and standardized screenshots of a real-world platform (Moodle).

External Validity. Undergraduates may not fully represent experienced professionals, though they serve as suitable proxies for junior developers in initial empirical studies [27]. Additionally, relying on a single system (Moodle) and a specific task (file submission) limits generalizability. Future replications should involve diverse populations, varying GenderMag experience levels, and more complex interfaces.

8 Conclusion

This study showed that the participants' cognitive profiles influence the application of the GenderMag method. Quantitative analysis revealed significant differences between groups defined by cognitive facets, especially in the "Yes" and "Maybe" responses, indicating that participants with distinct cognitive styles interpret the actions a persona would perform during the inspection differently. Qualitative analysis reinforced this finding, with many participants reporting difficulty distancing themselves from their own way of thinking and noting greater ease in adopting the perspective of the persona whose cognitive profile was most similar to their own. Furthermore, the scenario used was perceived as more compatible with Tim's exploratory style, whereas Abby's was considered more difficult to represent, requiring greater attention, caution, and clarity in the task flow.

These findings have implications for software development. If a development team predominantly possesses a specific cognitive style, the inspection results may reflect this bias, failing to reveal problems faced by users with distinct profiles. Thus, the application of GenderMag, when performed without cognitive diversity among evaluators, may not fully achieve its goal of identifying barriers related to cognitive differences, many of which are also associated with gender inequalities in the development and use of technology.

Based on these findings, our study points to promising avenues for future work. Among them, the need to deepen evaluator training in persona interpretation and in distancing themselves from their own cognitive styles during inspection stands out. Another possibility is investigating hybrid approaches that combine human evaluation with intelligent systems that reinforce persona characteristics and provide decision-making support. Future studies could also explore varied scenarios, more detailed persona profiles, and replications with participants of different experience levels, broadening the generalization of the results and contributing to the continuous improvement of GenderMag as a technique to support inclusive interface design.

Artifact Availability

The artifacts, raw data, and the analysis files are available in [8].

Acknowledgements

Igor Wiese acknowledges CNPq (409359/2024-6, 444802/2024-0) and Fundação Araucária/Government of Paraná (PRD2023361000043).

References

[1] Silvia Abrahão et al. 2025. Software Engineering by and for Humans in an AI Era. *ACM Transactions on Software Engineering and Methodology* 34, 4 (2025), 1–45. doi:10.1145/3715111

[2] Marios Belk, Christos Fidas, Panagiotis Germanakos, and George Samaras. 2017. The interplay between humans, technology and user authentication: A cognitive processing perspective. *Computers in Human Behavior* 76 (2017), 184–200. doi:10.1016/j.chb.2017.06.042

[3] Giorgio Brajnik, Yeliz Yesilada, and Simon Harper. 2011. The Expertise Effect on Web Accessibility Evaluation Methods. *Human–Computer Interaction* 26, 3 (2011), 246–283. arXiv:https://doi.org/10.1080/07370024.2011.601670 doi:10.1080/07370024.2011.601670

[4] Margaret Burnett, Robin Counts, Ronette Lawrence, and Hannah Hanson. 2017. Gender HCl and microsoft: Highlights from a longitudinal study. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 139–143. doi:10.1109/VLHCC.2017.8103461

[5] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Laura Beckwith, and Anicia Peters. 2016. GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.

[6] Gül Çalıklı and Ayşe Başar Bener. 2013. Influence of confirmation biases of developers on software quality: an empirical study. *Software Quality Journal* 21, 2 (June 2013), 377–416. doi:10.1007/s11219-012-9180-0

[7] Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology* 87, 3 (2004), 327. doi:10.1037/0022-3514.87.3.327

[8] Carolina Yumi Fujii, Igor Scaliante Wiese, and Walter Nakamura. 2026. *Supplementary material for the paper "Is It Possible to Step into Someone Else's Shoes? An Empirical Study Investigating the Compatibility Between Cognitive Styles and GenderMag Personas"*. doi:10.5281/zenodo.21418774

[9] Carina S. González-González. 2023. Introducing Speculative Design, Gender, and AI in Lean UX for HCI Education. In *International Conference on Human Computer Interaction*. ACM. doi:10.1145/3612783.3612801

[10] M. M. Hamid et al. 2024. How to Measure Diversity Actionably in Technology. In *Equity, Diversity, and Inclusion in Software Engineering*. Apress, 469–485.

[11] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human–Computer Interaction* 13, 4 (2001), 421–443. arXiv:https://doi.org/10.1207/S15327590IJHC1304$_0$5 doi:10.1207/S15327590IJHC1304_05

[12] Niels Ebbe Jacobsen, Morten Hertzum, and Bonnie E. John. 1998. The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42, 19 (1998), 1336–1340. arXiv:https://doi.org/10.1177/154193129804201902 doi:10.1177/154193129804201902

[13] Chen Ling and Gavriel Salvendy. 2009. Effect of Evaluators' Cognitive Style on Heuristic Evaluation. *International Journal of Human-Computer Studies* 67, 4 (2009), 382–393.

[14] Nicola Marsden and Maren Haag. 2016. Evaluation of GenderMag Personas Based on Persona Attributes and Persona Gender. In *Human-Computer Interaction: HCII 2016 Posters, Part I*, Constantine Stephanidis (Ed.). Springer, Cham, 122–127. doi:10.1007/978-3-319-40548-3_20

[15] Rahul Mohanani, Iflaah Salman, Burak Turhan, Pilar Rodríguez, and Paul Ralph. 2020. Cognitive Biases in Software Engineering: A Systematic Mapping Study. *IEEE Transactions on Software Engineering* 46, 12 (2020), 1318–1339. doi:10.1109/TSE.2018.2877759

[16] Emerson Murphy-Hill et al. 2024. GenderMag Improves Discoverability in the Field. In *International Conference on Software Engineering*. ACM.

[17] Lyndsey O'Brien, Tanjila Kanij, and John Grundy. 2025. Assessing gender bias in the software used in computer science and software engineering education. *Journal of Systems and Software* 219 (2025), 112225. doi:10.1016/j.jss.2024.112225

[18] H. S. Qiu and M. Connell. 2022. Seeking New Measures for Gender Bias Effects in Open Source Software. In *Cooperative and Human Aspects of Software Engineering*. ACM.

[19] Jordan M Robbins and Joachim I Krueger. 2005. Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and social psychology review* 9, 1 (2005), 32–47. doi:10.1207/s15327957pspr0901_3

[20] Michael Ross and Fiore Sicoly. 1979. Egocentric biases in availability and attribution. *Journal of personality and social psychology* 37, 3 (1979), 322. doi:10.1037/0022-3514.37.3.322

[21] Steven Samuel, Geoff G Cole, Madeline J Eacott, and Rebecca Edwardson. 2026. How to eliminate (and even reverse) egocentric bias in perspective taking. *Quarterly Journal of Experimental Psychology* 79, 2 (2026), 338–354. doi:10.1177/17470218251341289

[22] Steven Samuel, Anna Frohnwieser, Robert Lurz, and Nicola S Clayton. 2020. Reduced egocentric bias when perspective-taking compared with working from rules. *Quarterly Journal of Experimental Psychology* 73, 9 (2020), 1368–1381. doi:10.1177/1747021820916707

[23] M. T. S. Santos, Isabela Gasparini, and Luciana B. Frigo. 2024. Análise da Participação Feminina nos Cursos da Área de Computação: Um Olhar sobre os Projetos do Programa Meninas Digitais. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*. SBC, Rio de Janeiro, 2259–2272.

[24] Arun Shekhar and Nicola Marsden. 2018. Cognitive walkthrough of a learning management system with gendered personas. In *Proceedings of the 4th Conference on Gender & IT (GenderIT '18)*. 191–198. doi:10.1145/3196839.3196869

[25] Christian Alex de Souza da Silva, Tayana Uchôa Conte, and Wilson Silva Prata. 2025. Revolucionando a inclusão digital: uma avaliação híbrida de interfaces potencializada pelo GPT-4. *Revista P2P & Inovação* 12, 1 (2025). doi:10.21728/p2p.2025v12n1e-7624

[26] Dag I. K. Sjoberg, Tore Dyba, and Magne Jorgensen. 2007. The Future of Empirical Methods in Software Engineering Research. In *Future of Software Engineering (FOSE '07)*. 358–378. doi:10.1109/FOSE.2007.30

[27] Dag I. K. Sjøberg et al. 2005. A Survey of Controlled Experiments in Software Engineering. *IEEE Transactions on Software Engineering* 31, 9 (2005), 733–753.

[28] Kethelen do Carmo Souza, Genildo Gomes, Bruno Gadelha, and Tayana Conte. 2024. Utilizando o GenderMag para aprimorar UX em plataformas educacionais. In *Proceedings of the 27th Ibero-American Conference on Software Engineering (CIBSE)*. Sociedade Brasileira de Computação, Curitiba, PR, Brazil, 396–403. doi:10.5753/cibse.2024.28466

[29] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131. doi:10.1126/science.185.4157.1124

[30] Mihaela Vorvoreanu, Lauren Wilcox, Yvonne Morley, Denae Ford, Amber S. Horvath, and Margaret Burnett. 2019. From Gender Biases to Gender-Inclusive Design: An Empirical Investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. doi:10.1145/3290605.3300283

[31] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media. doi:10.1007/978-3-642-29044-2

[32] Yeliz Yesilada, Giorgio Brajnik, and Simon Harper. 2009. How much does expertise matter? a barrier walkthrough study with experts and non-experts. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) *(Assets '09)*. Association for Computing Machinery, New York, NY, USA, 203–210. doi:10.1145/1639642.1639678