Empirical Software Engineering manuscript No.
(will be inserted by the editor)

Are there gender differences when interacting with social goal models?

A quasi-experiment

Catarina Gralha · Miguel Goulão · João Araújo

Received: date / Accepted: date

Abstract *Context:* Research has shown gender differences in problem-solving, and gender biases in how software supports it. GenderMag has five problem-solving facets related to gender-inclusiveness: motivation for using software, information processing style, computer self-efficacy, attitude towards risk, and ways of learning new technology. Some facet values are more frequent in women, others in men. The role these facets may play when interacting with social goal models is unexplored. *Objectives:* We evaluated the impact of different levels of GenderMag facets on creating, modifying, understanding, and reviewing iStar 2.0 models. *Methods:* We performed a quasi-experiment and characterised 180 participants according to each GenderMag facet. Participants performed creation, modification, understanding, and reviewing tasks on iStar 2.0. We measured their accuracy, speed, and ease, using metrics of task success, time, and effort, collected with eye-tracking, EEG and EDA sensors, and participants' feedback. *Results:* Although participants with facet levels frequently seen in women had lower speed when compared to those with facet levels more often observed in men, their accuracy was higher. There were also statistically significant differences in visual and mental effort, and stress. Overall, participants were able to create, modify, and understand the models reasonably well, but struggled when reviewing them. *Conclusions:* Participants with a comprehensive information processing style and a conservative attitude towards risk (characteristics frequently seen in female) solved the tasks with lower speed but higher accuracy. Participants with a selective information processing style (characteristic frequently seen in males) were able to better separate what was relevant from what was not. The complementarity of results suggests there is more gain in leveraging people's diversity.

Keywords social goal models · iStar 2.0 · biometrics · gender

Catarina Gralha, Miguel Goulão, João Araújo
NOVA LINCS, Department of Computer Science
Faculty of Science and Technology
Universidade NOVA de Lisboa
E-mail: acg.almeida@campus.fct.unl.pt,
{mgoul, joao.araujo}@fct.unl.pt

1 Introduction

Research into gender differences has determined that individual characteristics in how people solve problems often cluster by gender [5,77]. In software systems, it is common to have features inadvertently designed to be more supportive of problem-solving processes typically followed by males than by females [36,88]. Awareness of these gender biases within software systems has increased in recent years [30,81], and analysing gender differences with software is important. If males and females work differently with software systems, tools, and other artefacts, such as requirements models, these differences could reveal a need to change the artefacts or the way they are dealt with, by taking this new knowledge into account. Designing software systems to be more gender-inclusive can benefit all problem solvers, regardless of their gender [89,46].

GenderMag (**Gender** Inclusiveness **Mag**nifier) [11] aims to help software practitioners evaluate their software system from a gender-inclusiveness perspective. It has five problem-solving facets related to gender-inclusiveness, that have been extensively investigated in the literature: *i)* motivation for using the software (e.g., [83,43,27,12]); *ii)* information processing style (e.g., [14,54,59,53]); *iii)* computer self-efficacy (e.g., [25,42,2,44]); *iv)* attitude towards risk (e.g., [96,23,18]); and *v)* ways of learning new technology (e.g., [43,5,70,9]). Some facet values are more frequent in women, others in men. GenderMag proposes personas to bring those facets to life. Although GenderMag has been used in HCI and design (e.g., [8,95]), the role its facets may play when building social goal models is mostly unexplored.

In this paper, our goal is to report the impact of differences in the levels of each GenderMag facet, when stakeholders perform creation, modification, understanding, and reviewing tasks on iStar 2.0 models [21]. iStar 2.0 is an evolution of *i** [100], a goal-driven modelling language used to model software requirements. We characterised a total of 180 participants according to each GenderMag facet. We measured their *accuracy*, *speed* and *ease* with which they accomplished their tasks by collecting measures such as precision, recall, and f-measure for assessing accuracy; the duration of those tasks for assessing speed; the visual effort (assessed with eye-tracking), the mental effort (assessed with EEG) and stress while performing them (assessed with EDA), and the participants' perceptions on their effort (measured with a NASA-TLX questionnaire), to assess ease. The combination of all these techniques gives us a multi-perspective about the way stakeholders interact with iStar 2.0 models.

Our results support the evidence that participants with a comprehensive information processing style and a more conservative attitude towards risk (both characteristics are seen more frequently in women) analyse the entire problem more thoroughly before starting the proposed task. The visual effort, attention and mental workload were also higher for these participants.

This paper extends our previous work on analysing gender differences in the interaction with social goal models [34] by adding two new tasks performed by participants, related to iStar 2.0 understanding and reviewing, and the reporting of the corresponding experiments and statistical analysis, for these two new tasks. The previous paper did not cover these tasks. Furthermore, we report on a total of 180 participants who have performed the experiments. These include the 100 presented in the previous paper and 80 additional participants.

2 Background

2.1 *i** and iStar 2.0

The *i** framework is a goal-driven modelling language [100] used to model software requirements. It provides the Strategic Dependency (SD) and the Strategic Rationale (SR) model. The SD model specifies the links and external dependencies among organisational actors. The SR model allows an analysis of goals fulfilment through several actors contributions. Actors represent stakeholders that depend on each other to accomplish their goals, perform tasks and provide resources. The *i** framework has evolved to iStar 2.0 [21], which we use in this paper.

Changes include the discontinuation of some concepts and the introduction of new ones. This new iStar version kept general actor, role, agent, goal, task and resource, but softgoal was renamed to quality. Actor's links *occupies*, *is-part-of*, *covers* and *plays*, were amalgamated into the *participates-in* link. The actor link *is-a* was preserved. The *means-end* and *task decomposition* links were combined into *refinement*, whereas contribution links were kept. Finally, two new links were added, *qualification* and *neededBy*.

To illustrate the application of some of these concepts, in Figure 1 we present an example where a meeting participant wants to use the meeting scheduler system to plan for a meeting.

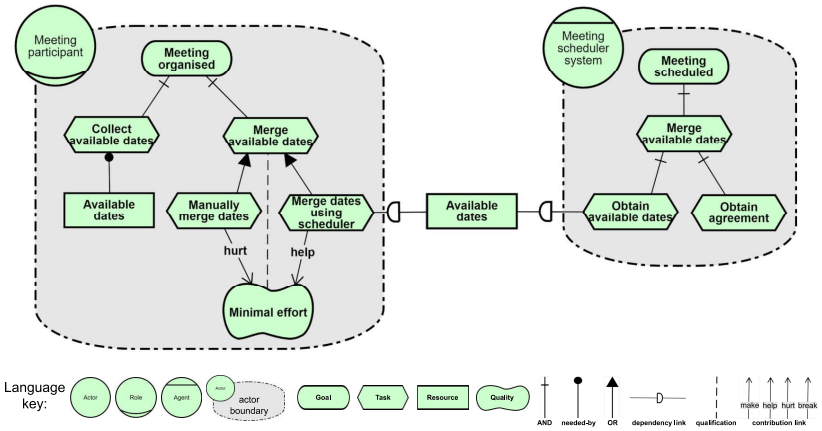


Fig. 1: Meeting scheduler, adapted from [101].

2.2 GenderMag (**Gender** Inclusiveness **Mag**nifier)

GenderMag [11] is a method for finding gender-inclusiveness issues in software features. It can be described as an analytic method for evaluating usability with a focus on gender-inclusiveness. This method has five problem-solving facets related to gender-inclusiveness, which are the ones repeatedly implicated by research from other fields, such as psychology, education and communications: *i)* motivation

for using the software, *ii)* information processing style, *iii)* computer self-efficacy, *iv)* attitude towards risk, and *v)* ways of learning new technology.

The facets come to life with 4 personas: Tim, Abby, Pat(ricia) and Pat(rick). Each persona has a value for every facet, and a specific background consistent with those facet values. Abby's facet values are more frequently seen in females, and Tim's are more frequently seen in males [11]. The Pats' (identical) facet values emphasise that differences relevant to inclusiveness lie in the facets themselves, and not in a person's gender identity. In Table 1 we summarise the facet values for each persona. A complete characterisation of the personas is available at the GenderMag Project website [10].

Table 1: Summary of the facet values for each persona in GenderMag.

	Abby	**Pats**	**Tim**
Motivation	Technology is used to accomplish tasks	Technology is used to accomplish tasks	Technology is a source of fun
Information processing	Comprehensive	Comprehensive	Selective
Self-efficacy	Low compared to peer group	Medium	High compared to peer group
Risk	Risk-averse	Risk-averse	Risk-tolerant
Learning style	Process-oriented	Tinkering (reflectively)	Tinkering (sometimes excessively)

In this paper, rather than using the personas to define how iStar 2.0 should support the different facets, we use a GenderMag questionnaire [95] to characterise stakeholders and determine their persona in each of the 5 facets. It is common to have stakeholders characterised as Abby in some of the facets, and as Tim in others. A person being a "pure" Abby or a "pure" Tim, that is, with all the 5 facets having the same persona, is rare (see, for example, [95]).

We then explore how variations in the facets impact the creation, modification, understanding, and reviewing of iStar 2.0 models.

2.3 Biometric sensors and subjective workload

In order to gain a multi-perspective on how stakeholders interact with iStar 2.0 models, we use a combination of different techniques to collect *i*) direct task performance metrics; and *ii*) indirect measures such as the effort while performing the tasks, assessed with 3 biometric devices (eye-trackers, EEG and EDA), and the participants' perceptions on their effort, measured with NASA-TLX.

Eye-tracking. Eye-tracking is a technology that measures the activity of the eyes. In human vision, eye-movements are essential to collect evidence regarding participants' cognitive processes [68,76]. Eye-tracking devices, called eye-trackers, monitor a participant's visual attention by collecting eye-movement data when (s)he looks at a stimulus, while working on a specific task. A stimulus is an object, such as text, source code, or diagram, that is necessary to perform that task [69,

24]. Beyond the analysis of visual attention and cognitive processes, eye-data can also be examined to measure the workload of a task. Furthermore, the data can be studied with respect to certain areas of the stimuli, which are called areas of interest (AOI). An AOI can either be relevant or irrelevant for a given task that is being performed by a participant.

EEG. Electroencephalography (EEG) refers to the measurement of the brain's electrical activity that arises from neuronal firing. The varying activity of neurons in the brain causes fluctuations in the voltage potential along the scalp that can be measured with an EEG scanner [1]. When analysing EEG data, the focus is generally on the spectral content of the EEG, that is, the type of neural oscillations (also known as brain waves) that can be observed. Brain waves can be divided into frequency bands, called alpha (α, 8-12 Hz), beta (β, 12-30 Hz), gamma (γ, 30-100+ Hz), delta (δ, 0-4 Hz), and theta (θ, 4-7 Hz) [39]. Although EEG scanners started by being used to diagnose epilepsy, sleep disorders, coma, encephalopathies, and brain death [90], some work has linked these specific frequency bands with mental workload, task engagement and emotions [49,50,56]. Each of the frequency bands has a specific frequency range and amplitude, exhibiting more or less activity under certain circumstances. For instance, alpha waves can typically be observed when an individual is in a relaxed state, but they either disappear or their amplitude decreases significantly as soon as the physical or mental activity increases [1].

EDA. Electrodermal activity (EDA) is a biological property of the human body that causes continuous variation in the electrical characteristics of the skin, being an output of the sweat glands on a microscopic level. Sweating is controlled by the sympathetic nervous system [52] and if the sympathetic branch of the autonomic nervous system is highly aroused, sweat gland activity increases, which in turn increases skin conductance. EDA scanners measure this electrical skin conductance, which serves as an indicator for emotional stimuli [26,37]. When an individual experiences emotional activation (such as excitement or stress), or an increased cognitive workload, or physical exertion, the brain sends signals to the skin to increase the level of sweating. One may not feel any sweat on the surface of the skin, but the electrical conductance increases in a measurably significant way as the pores begin to fill below the surface [17].

Subjetive workload. Cognitive load can be defined as a multi-dimensional construct representing the load that a task imposes on a participant [60,61]. This also refers to the level of perceived effort for learning, thinking and reasoning as an indicator of pressure on working memory during task execution [99]. This measure of mental workload represents the interaction between task processing demands and human capabilities or resources [97,38]. The NASA-Task Load Index (NASA-TLX) [41,40] is a technique for subjective workload and cognitive load assessment. It uses uses 6 (six) dimensions to assess workload and cognitive load: mental, physical, and temporal demand, performance, effort, and frustration. Twenty-step bipolar scales are used to obtain ratings for these dimensions, and a score from 0 to 100 is obtained on each scale. Then, a weighting process with a paired comparison is used: the participant chooses which dimension is more relevant to the workload for a particular task across all pairs of dimensions. The number of times each dimension is chosen is the weighted score. This is multiplied by the scale score for each dimension and then divided by 15 to get a workload score from 0 to 100.

2.4 Related work

Gender differences in problem-solving activities have been investigated in different domains. For instance, gender differences have been observed in intellectual risk-taking tasks, which require mathematical and spatial reasoning skills [13]. Some studies investigated the impact of self-efficacy on Math problem-solving success [62], as well as on strategies followed by males and females to solve problems [4, 93]. Fisher et al. [28] conducted a study to compare male and female subjects' performance on program comprehension tasks. More recently, Sharafi et al. [77] conducted an experiment with 15 males and 9 females to identify whether there is a relationship between gender and the visual effort, time and ability to memorise identifiers, namely *camelCase* and *under_score*. An eye-tracker measured the duration of the execution of each task and the visual effort. Females focused more on incorrect answers than male participants. However, this does not affect the task's duration.

Biometric sensors have been used in Software Engineering. For instance, Crosby et al. [20] used eye-tracking technology to study the differences in program comprehension and source code reading navigation strategies between experienced and less experienced software developers in Pascal. Eye-tracking has been used to assess the effort involved in software models' understanding [76]. Yusuf et al. [102] used eye-tracking to compare the visual effort involved in answering questions about UML class diagrams designed with 3 different layout strategies. Sharif et al. [79, 78] studied the effect of different layouts for design pattern roles identification in UML class diagrams. Other studies with eye-tracking focused on BPMN [64], ER [15], TROPOS [74] and *i** [73, 72].

Ikutani et al. [45] used near-infrared spectroscopy to investigate the difference in brain activity for various types of program comprehension tasks. Siegmund et al. [82] examined the active brain regions during small code comprehension tasks using functional magnetic resonance imaging (fMRI).

In terms of using multiple biometric sensors, Fritz et al. [29] and Störrle et al. [87] classify the difficulty of code or models comprehension, respectively, by using a combination of eye-tracking and EEG. Müller et al. [55] used eye-tracking, EDA and EEG to investigate developers' emotions in software change tasks and their correlation with perceived progress.

Our work differs from previous works as we use a combination of GenderMag, multiple biometric sensors (eye-tracker, EEG, and EDA scanners) and NASA-TLX questionnaire to analyse gender differences when creating, modifying, understanding, or reviewing requirements models, in particular, iStar 2.0 models.

3 Experiment planning

3.1 Goals

We describe our research goals following the GQM research goals template [3]. Our first goal (G1) is to *analyse* differences in the levels of the GenderMag facets, *for the purpose of* evaluation, *with respect to* their effects on the **creation** of iStar 2.0 models, *from the viewpoint* of researchers, *in the context of* an experiment conducted at our University and at software companies. Our second (G2), third

(G3), and fourth (G4) goals can be obtained by replacing the term *creation* with ***modification***, ***understanding***, and ***reviewing***.

We can break down each goal into three sub-goals, concerning the effect(s) of the different facets, in terms of *accuracy*, *speed*, and *ease*. The refined goals can be obtained by replacing the terms *creation* (or *modification*, *understanding*, and *reviewing*) with *accuracy to create*, *speed to create*, and *ease to create*.

3.2 Participants

This evaluation was performed by 180 participants selected by convenience and snowball sampling. We had 50 participants performing the creation task, 50 in the modification task, 40 in the understanding task, and 40 in the reviewing task. We leveraged personal contacts and participants were made aware of the study either by direct communication or by e-mail. Some of these participants actively recruited their contacts to participate, hence the snowball sampling. This technique also allowed us to have a more diversified set of participants.

We calculated the sample size needed to ensure an adequate power level, where 0.8 is considered appropriate (80% probability of correctly detecting a real effect) [47]. We chose a standardised large Cohen's effect size for $\alpha = 0.05$ (significance level). To detect a large difference between two independent sample means at $\alpha = 0.05$, at least 26 participants are required in each group [19].

Concerning participants *age* distribution, they had between 20 and 42 years old, with an average of 27 years old. With respect to *gender*, there were 125 male participants and 55 females. Participants had the option to select "other" in the gender question, but none of them did. In terms of *nationality*, 179 were Portuguese, and 1 was Brazilian. Regarding the *usage of reading devices*, 70 participants wore eyeglasses, and 31 had contact lenses.

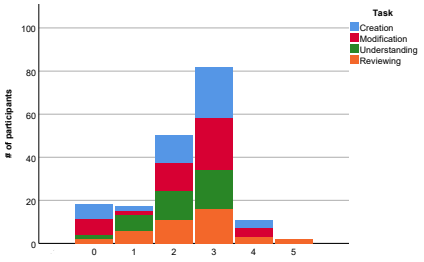
All participants had some university-level training, and their *field of studies* spanned across multiple areas. We had 114 computer scientists, 2 designers, 1 electrotechnical engineer, 26 environmental engineers, 7 historians, 18 lawyers, 2 mechanical engineers, and 10 medical doctors. For the *highest completed level of education*, 22 completed high school, 55 concluded a BSc, 101 had an MSc, and 2 a PhD degree. Concerning their *current level of education*, 1 was in the first year of the BSc degree, 12 on the second year, and 17 on the third (and final) year. As for MSc students, 14 were in the first year, and 29 were on the second (and final) year. Finally, 37 were doing a PhD, and 70 were no longer studying. The ones that were no longer studying had at least 3 years of experience. Concerning their *current occupation*, 74 of the participants were students, 36 were working students, 68 were practitioners, and 2 were researchers.

Regarding their previous *experience* with iStar 2.0, for 154 participants, it was their first contact with it. However, 24 learnt it in the context of a course, and 2 in a professional environment. In those two latter scenarios, participants *usage time* with the versions had an average of 6 months. Some participants referred to 3, 4 or 6 months. We argue that all those months correspond to a University semester, depending on how people count. As for the *last use* of the version, the vast majority of participants was no longer using it, and only had contact with iStar 2.0 in a specific University course. Lastly, in terms of *knowledge on other requirements models*, 100 participants claim to know UML in general, 9 referred

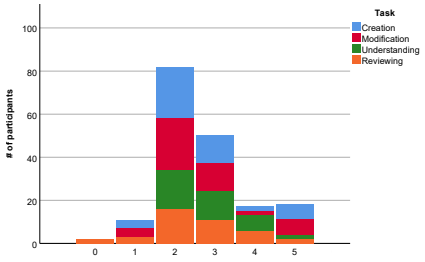
to BPMN, and 2 specifically said they work with flowcharts in particular. The remaining 69 participants did not report knowing any requirements language.

Participants spanned a reasonably wide range of values of each of the five GenderMag facets (see Figures 2a and 2b). In these Figures, the x axis has the number of facets with a given persona (Abby persona in Figure 2a) and Tim persona in Figure 2b. The Figures mirror each other. A person is classified as an Abby or as a Tim in each one of the five facets of GenderMag. If a person is an Abby on 3 facets, that person is represented in the 3 of x axis in Figure 2a. Then, that same person is a Tim on the other 2 facets, being represented in the 2 of the x axis for Figure 2b. From both Figures, we are able to understand that the majority of our participants were characterised as Abby in 3 out of the 5 GenderMag facets. The Figures also allow us to conclude that we had 2 participants characterised as a "pure" Abby (with 5 out of 5 facets as Abby), and 18 as a "pure" Tim (with 5 out of 5 facets as Tim). The other 160 participants had mixed characteristics of both Abby and Tim.

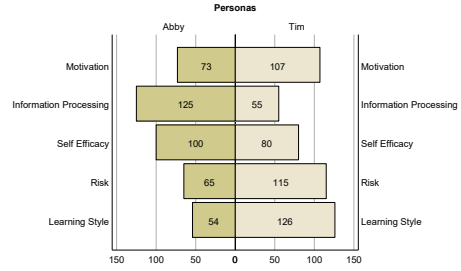
When analysing each facet (Figure 2c), the majority of the participants was identified as Tim in the *motivation*, *risk*, and *learning style* facets. For *information processing* and *self-efficacy*, on the other hand, the majority of participants was described as Abby. Taking a closer look into the relationship between the personas in each of the facets and the gender of participants (Figure 2d), the majority of female participants was characterised as Abby in all the facets, being *learning style* an exception. As for the males, the majority of participants was classified as Tim in all the facets, except for *information processing* and *self-efficacy*. These results support the literature claim [5,77] that characteristics in how people solve problems often cluster by gender.



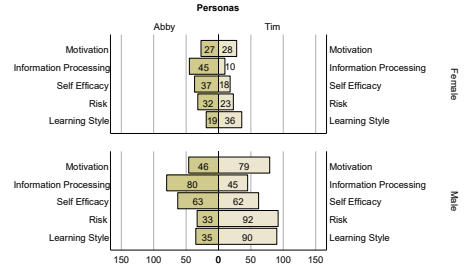
(a) Number of facets with the Abby persona.



(b) Number of facets with the Tim persona.



(c) Personas in each facet.



(d) Personas in each facet, by gender.

Fig. 2: Participants' distribution across GenderMag facets.

3.3 Experimental materials

The experimental materials for this evaluation included *i)* a participant consent form; *ii)* a video tutorial about iStar 2.0; *iii)* a task to be performed; *iv)* a NASA-TLX questionnaire; *v)* a demographic questionnaire; and *vi)* a GenderMag questionnaire.

The *participant consent form*, adapted from [71], explained that the participation was entirely voluntary, the participants could refuse to answer any question and could leave at any time, and that all the collected data would remain anonymous.

The *video of fish swimming*, with 2 minutes, served as a baseline to normalise the captured biometric data [29, 55]. It also helped participants to relax and better focus on the task at hand.

The *video tutorial*, with 3 minutes and 58 seconds, explained the elements of an iStar 2.0 model. It includes the construction of a correct model (similar to those that will be created, modified, understood, or reviewed in the experiment) about a meeting scheduler system; and an audio and textual description of both the model elements, as they are being introduced, and their role in the model under construction. The modelling elements were described using the exact phrases and explanations present in the iStar 2.0 Language Guide [21]. The participants had no control over the video, not being able to pause it or resume it, since having different viewing times and going through specific parts of the tutorial more than one time could impact the results.

The *GenderMag questionnaire* had a set of 9-point Likert questions. There are 20 questions, divided into questions related to each one of the facets. The scores for each facet are added, and each individual is compared to the grand median (median of medians) for that facet. If a participant is above the grand median on a given facet, we name him/her Tim (on that facet alone). If (s)he is below, we name him/her Abby (on that facet alone). Due to the way scores are calculated, it is common that Pats are not present in the facets [95].

The *creation* and *modification* tasks were performed in a custom web-based tool developed by the authors, available at [32]. The tool was implemented having piStar [65] as a basis. Both creation and modification tasks share a common structure, with 3 (three) Areas Of Interest (AOI): the *problem description* on the left-hand side; the *editor's toolbar* on top; and the *canvas* where participants would create or modify the models. The *understanding* and *reviewing* tasks share the same structure, with 3 (three) AOI: the *language key* on the left-hand side, the *question* the participant is supposed to answer on top, and the *iStar 2.0 model* about which the question is asked. Since the modelling tool had the editor's toolbar with all the model elements and the corresponding names, the language key was not needed in the creation and modification tasks.

The *NASA-TLX questionnaire* collected feedback on the participants' perceptions concerning their effort on the performed task. It uses six dimensions: mental, physical, and temporal demand; performance; effort; and frustration.

The *demographic questionnaire* collected the demographic information on the participants.

All the materials used in this evaluation can be found in the paper's companion site [35].

3.4 Tasks

In all the tasks, the domain was a *booking management system for a hotel*. We opted for a relatively known domain to reduce the effect of the results being related to difficulties in understanding the domain itself, and not due to the artefacts that were under study. However, we are aware that tacit knowledge may also play an important role in the performance of the participants.

Each participant completed 1 *task*. However, there were 4 types of tasks: *creation*, *modification*, *understanding* and *reviewing*. In the *creation task*, participants had to create an iStar 2.0 model given a small problem description. In the *modification task*, participants had to modify an initial iStar 2.0 model, given a problem description and a new requirement. In the *understanding task*, participants had to answer a total of 7 questions about a given iStar 2.0 model. The questions, appearing in a random order, aimed to cover the main elements of an iStar 2.0 SR model. Finally, in the *reviewing task*, participants had to identify semantic defects on a given iStar 2.0 model, but we only informed the participants that their task was to find "defects". Explicitly describing the type of defects would have introduced a bias in the participants' attention. This way, each participant was free to review the model using his best judgement as a real-world stakeholder would. Typically, requirements modelling tools should protect the user against syntactic defects, hence our choice for semantic ones.

The distribution of the tasks to the participants was random, but we balanced the number of participants performing each task.

3.5 Hypotheses, parameters, and variables

For each one of the high level goals presented in Subsection 3.1, we define the null (H_0) and alternative hypotheses (H_1). For G1, concerning *creation* tasks, we have the following hypotheses:

> $H_{0Create}$: Differences in the levels of each facet **do not** impact the *creation* of iStar 2.0 models.
> $H_{1Create}$: Differences in the levels of each facet impact iStar 2.0 the *creation* of iStar 2.0 models.

These hypotheses are further refined to cope with *accuracy*, *speed* and *ease*. For example, for accuracy:

> $H_{0CreateAcc}$: Differences in the levels of each facet **do not** impact the *accuracy to create* iStar 2.0 models.
> $H_{1CreateAcc}$: Differences in the levels of each facet impact the *accuracy to create* iStar 2.0 models.

And similar for speed and ease of *creation*. We follow the same approach to define the null and the alternative hypotheses for G2, G3 and G4, concerning *modification*, *understanding*, and *reviewing* tasks, respectively. These hypotheses are also further refined to cope with *accuracy*, *speed* and *ease*.

The **independent** variables are the levels (*Abby*, *Pats*, *Tim*) on each of the five GenderMag facets (motivation for using software, information processing style,

computer self-efficacy, attitude towards risk, and ways of learning new technology). The **dependent** variables are *accuracy*, *speed*, *ease* and *perceived effort*. For each of these variables, there is a set of metrics. From Table 2 to 7, we present an overview of these metrics. The first column shows the name of the variable, while the second one has its abbreviation. The third column presents the range, and the last column has the counting rule or formula for the metric calculation.

Assessing accuracy. In Table 2 we present the metrics for the dependent variable accuracy. Higher values of *precision*, *recall* and *f-measure*, support the claim of a better *accuracy*.

Table 2: Overview of the metrics for the *dependent* variable *accuracy*.

Name	Abbreviation	Range	Counting rule
Precision	–	$0 \leq x \leq 1$	$\frac{number\ of\ gold\ standard\ elements\ retrieved}{total\ number\ of\ retrieved\ elements}$
Recall	–	$0 \leq x \leq 1$	$\frac{number\ of\ gold\ standard\ elements\ retrieved}{total\ number\ of\ gold\ standard\ elements}$
F-measure	–	$0 \leq x \leq 1$	$\frac{2 * (precision * recall)}{(precision + recall)}$

Assessing speed. In Table 3 we present the metrics for the dependent variable speed. Lower values of these metrics correspond to better *speed*, indicating that the corresponding facet level may help in improving the speed with which the models are created, modified, understood, or reviewed. While the overall *duration* addresses the time spent in the task, *first action*, *last action*, *first detection* and *last detection* provide a detailed picture of the moment when the participant starts and ends providing valid feedback. The metrics *first action* and *first detection* are similar, but applied to different tasks. The former is used in the creation and modification tasks, while the latter is applied in the understanding and reviewing tasks. They are related with the time taken by the participant to add the *first* element (for creation and modification), or to report the *first* response element (for understanding and reviewing). The same is valid for *last action* and *last detection*, which are related with the *last* element. A higher value for *processing duration* indicates that the participant stopped working on the task, but decided to revise it before finishing.

Table 3: Overview of the metrics for the *dependent* variable *speed*.

Name	Abbreviation	Range	Counting rule
Duration	–	$0 \leq x$	$completion\ time - start\ time$
First action	FirstAct	$0 \leq x$	$first\ action\ time - start\ time$
Last action	LastAct	$0 \leq x$	$last\ action\ time - start\ time$
First detection	FirstDet	$0 \leq x$	$first\ detection\ time - start\ time$
Last detection	LastDet	$0 \leq x$	$last\ detection\ time - start\ time$
Processing duration	ProcDur	$0 \leq x$	$duration - last\ (action \vee detection)$

Assessing ease. The ease with which participants conduct their tasks is assessed by effort measures. We focus on: the *physical effort* and the *perception of*

effort reported by participants. The former is addressed with eye-tracking, EEG and EDA, while the latter is assessed through NASA-TLX.

In Table 4 we present the metrics for the dependent variable visual ease, collected with the eye-tracking device. A fixation is a stabilisation of the eye on a part of the stimulus for a period of time between 200 and 300 ms. A higher *number* and *duration* of *fixations* is associated with a higher visual attention in a given set of AOIs (in this case, relevant vs. irrelevant model elements) [66,75,76]. Regarding the *average duration of fixation*, a higher value indicates more time and attention devoted to AOIs, [76,67,15], which is correlated with cognitive processes [24,31]. A saccade is a sudden and quick eye-movement lasting between 40 to 50 ms. A higher *number of saccades* can be associated with a higher visual effort, meaning the participant may be somewhat "lost", making a more erratic navigation [85, 29,31,76]. A higher number of *saccades to the key* can also be associated with difficulties with the modelling language.

Table 4: Overview of the metrics for the *dependent* variable *visual ease*: eye-tracking

Name	Abbreviation	Range	Counting rule
Fixation rate on relevant elements	FixRel	$0 \leq x$	$\frac{number\ of\ fixations\ on\ the\ relevant\ AOI}{number\ of\ fixations\ on\ the\ AOG}$
Fixation rate on irrelevant elements	FixIrrel	$0 \leq x$	$\frac{number\ of\ fixations\ on\ the\ irrelevant\ AOI}{number\ of\ fixations\ on\ the\ AOG}$
Average duration of relevant fixations	AvgDurRelFix	$0 \leq x$	$\frac{\Sigma\ duration\ of\ fixations\ on\ the\ relevant\ AOI}{number\ of\ fixations\ on\ the\ relevant\ AOI}$
Average duration of irrelevant fixations	AvgDurIrrelFix	$0 \leq x$	$\frac{\Sigma\ duration\ of\ fixations\ on\ the\ irrelevant\ AOI}{number\ of\ fixations\ on\ the\ irrelevant\ AOI}$
Total number of saccades	TotSac	$0 \leq x$	$\Sigma\ saccades$
Total number of saccades to the key	Sac2Key	$0 \leq x$	$\Sigma\ saccades\ to\ the\ key\ AOI$

In Table 5 we present the metrics for the dependent variable mental ease, collected with the EEG scanner. The values for *average attention*, *average mental workload* and *average familiarity*, are calculated based on specific frequency bands, often referred to as alpha, beta, gamma, delta and theta. A decrease of alpha EEG activity and often an increase in theta EEG activity indicates an increase in attention demand and working memory load [86,56]. A higher *average attention* indicates the participant is engaged in the task, and a higher *average mental workload* indicates effort while performing it. For *average familiarity*, a higher value is associated with memory accessing and lower effort while performing the task.

In Table 6 we present the metrics for the **dependent variable emotional ease**, collected with the EDA scanner. A higher *average skin conductive level* is linked to a greater cognitive load, task difficulty, and stress [80,58]. For computing the *heart rate variability*, we used features that represent the difference in time between two heart beats: RMSSD (root mean square of successive differences), and NN50 (the number of pairs of successive beat-to-beat intervals that differ more than 50ms). An increase in the *heart rate*, when in a stationary state, can be related with anxiety [22,51] and mental stress [84].

Table 5: Overview of the metrics for the *dependent* variable *mental ease*: EEG

Name	Abbreviation	Range	Counting rule
Average attention	AvgAttention	$0 \leq x \leq 1$	$\frac{\Sigma\ attention\ value\ per\ ms}{total\ duration\ in\ ms}$
Average mental workload	AvgMentWL	$0 \leq x \leq 1$	$\frac{\Sigma\ mental\ workload\ value\ per\ ms}{total\ duration\ in\ ms}$
Average familiarity	AvgFam	$0 \leq x \leq 1$	$\frac{\Sigma\ familiarity\ value\ per\ ms}{total\ duration\ in\ ms}$

Table 6: Overview of the metrics for the *dependent* variable *emotional ease*: EDA

Name	Abbreviation	Range	Counting rule
Average skin conductive level	AvgSCL	$0.2 \leq x \leq 4$	$\frac{\Sigma\ SCL\ in\ \mu s}{total\ number\ of\ SCL}$
Average RMSSD	AvgRMSSD	$10 \leq x \leq 120$	$\frac{\Sigma\ RMSSD\ in\ ms}{total\ number\ of\ RMSSD}$
Average NN50	AvgNN50	$1 \leq x \leq 70$	$\frac{\Sigma\ NN50\ in\ ms}{total\ number\ of\ NN50}$

In Table 7 we present the metrics for the **dependent variable perceived effort**, assessed through the NASA-TLX questionnaire. Higher values, in all the metrics, correspond to a greater perceived effort by the participant. Each metric is weighted, in terms of its importance for the overall effort. The denominator *15* corresponds to the 15 paired comparison of all the 6 dimensions to access the perceived workload [16].

Table 7: Overview of the metrics for the *dependent* variable *perceived effort* [92].

Name	Abbreviation	Range	Counting rule
Mental demand	MD	$0 \leq x \leq 100$	$\frac{mental\ rating * mental\ weight}{15}$
Physical demand	PD	$0 \leq x \leq 100$	$\frac{physical\ rating * physical\ weight}{15}$
Temporal demand	TD	$0 \leq x \leq 100$	$\frac{temporal\ rating * temporal\ weight}{15}$
Performance	Perf	$0 \leq x \leq 100$	$\frac{performance\ rating * performance\ weight}{15}$
Effort	Eff	$0 \leq x \leq 100$	$\frac{effort\ rating * effort\ weight}{15}$
Frustration	Frust	$0 \leq x \leq 100$	$\frac{frustration\ rating * frustration\ weight}{15}$
NASA-TLX Score	–	$0 \leq x \leq 100$	$MD + PD + TP + Eff + Perf + Frust$

3.6 Experimental design

This evaluation follows a *quasi-experimental design*, since the allocation of participants to the tasks was random, but without a pre-selection process.

If a participant performed a given task, the next participant would be allocated to a different one, so that that the number of participants performing each task

would be balanced. The exception to this allocation was at the beginning of the allocation to the creation and modification tasks, which caused us to have 10 more participants than on the other tasks. This difference was caused by the way tasks were being allocated to participants: instead of changing the task from participant to participant, the participants were allocated to the same task until we had a reasonable number of participants performing that task. In particular, in the beginning of the experiments, the first 10 participants performed the creation task, and the next 10 participants performed the modification tasks. However, this could have caused a greater unbalance in the number of participants performing each task, if we were not able to find the same number of participants to perform the other tasks, and taking into account the participants was voluntary. As such, the remaining of the participants were allocated as in the previously described process: if a participant performed the modification task, the next participant would be allocated to the creation task, the next to the understanding task, and the next one to the reviewing task. In both processes, in terms of tasks distribution, we have a *between-subjects design*. This type of design is also called an independent measures design because every participant is only subjected to a single treatment, that is, only performs one of the tasks.

In a within-subjects design, we can have a smaller number of participants, as every participant performs more than one task. Furthermore, there is a reduced variability due to subject differences. However, a learning effect from one task to the next could represent a confounding factor. Even if the order of the tasks was changed, the results may still be affected by the ordering. With a between-subjects design, on the other hand, there is no learning effect, nor a side effect due to the ordering of the tasks. Nevertheless, it requires a higher number of participants and augments the variability due to subject differences. In order to reduce the latter, we have performed a random allocation of participants to tasks.

We opted for the between-subjects design for 3 main reasons: *i)* time; *ii)* fatigue; and *iii)* the learning effect. In particular, in studies with practitioners, time is a decisive factor. A quasi-experiment with multiple tasks may increase the mortality of the participants, or discourage them from participating, in the first place. Moreover, in a long experiment, the participants may become tired. This could decrease their performance on the last tasks. Alternatively, the learning effect may cause them to improve their performance over the course of the studies. Moreover, a crossover design, where every participant is subjected to more than one treatment, is complex [98] and it is discouraged based on the risk of performing an incorrect analysis [94].

4 Execution

4.1 Preparation

We carried out the data collection with a laptop connected to an external 22 inch, widescreen, full HD monitor; a The Eye Tribe eye-tracker [91]; a NeuroSky MindWave EEG headset [57]; a BioSignalsPlux Wristband [6] with BITalino [7] EDA scanner; and an external mouse and keyboard. We prepared the session on the laptop, and the participant had access to the external monitor, mouse and

keyboard. Participants sat on a chair without wheels, to avoid movements that could jeopardise the eye-tracker data.

We prepared the room setting so that all participants had similar conditions. In our University, the same meeting room was used for all the studies. For the experiments performed in software companies, the same room, in each company, was allocated to the entire day. The room was only being used for the studies, and there was only one participant in each session.

For the studies at our University, we scheduled the sessions according to participants' availability, with one hour between studies, so that the next participant would not have to wait too long, nor that the previous participant felt there were time constraints. For the studies at software companies, the participants appeared when they had a break on their normal workflow, and if a session was not being performed.

4.2 Procedure

When a participant arrived, (s)he sat on a chair in front of the external monitor, and was informed that the experiment consisted of watching a tutorial on a requirements language and performing a task. We further informed the participant that we would be recording the contents on the screen, tracking the eyes movement, and collecting information on mental effort and heart rate. These explanations where necessary so that the participant could comfortably use the biometric devices. Finally, we explained (s)he could quit at any moment, and that there was no time limit for performing the task.

The evaluation started with the participant *reading the consent form*. Next, (s)he *equipped the biometrics devices*, after a demonstration on how to correctly equip them. We then confirmed if the equipment were well placed. The goal of this procedure was to make participants more comfortable than if it was the researcher that had to touch the ear, head, or forehead of the participant. The EDA wristband was placed on the participant's non-dominant wrist, after removing any watches or bracelets. The buckle of the wristband was adjusted by the participant to a comfortable position (without it being too loose or too tight). Before putting the EEG headset, participants with earrings were asked to remove them. Special care was taken for participants with long hair so that it would not obstruct the ear clip (which acts as a ground and reference). Due to the sensibility of the forehead sensor, we helped the participant to remove any foundation (cosmetics) from the forehead. We also helped participants with hair bangs, so that nothing was obstructing the forehead sensor of the EEG headset. We helped the participant seating comfortably so that the eyes would be around 50cm away from the screen. The eye-tracker was placed below the screen, without blocking it. We adjusted the eye-tracker's angle to cope with differences among the participants' height. We then used the EyeTribe calibration application, only accepting *good* or *excellent* calibrations (top levels of a 5 points ordinal scale) to proceed to the actual data collection.

We asked the participant to *watch the video of fish swimming* while wearing the biometric sensors. This allowed us to normalise the captured biometric data.

After that, the participant *watched the video tutorial* on iStar 2.0 and then started *performing the task*. The audio was recorded since the participant needed

to give the answers to the understanding and reviewing tasks out-loud. For the creation and modification tasks, talking was not necessary, as the answer was being recorded on the screen. In all the cases, no (bio)feedback was provided to the participant during the entire evaluation, to avoid an unnecessary validity threat.

When the participant felt the task was completed, (s)he *answered the NASA-TLX questionnaire*. Finally, the participant *answered the demographic questionnaire*, with the possibility of leaving an e-mail address for receiving the aggregated results of the study, and the participant completed the *GenderMag questionnaire*.

The participant had control over the entire session, which means that, after watching the video of fish swimming, (s)he would click on a continue button and the video tutorial would start. After the video was finished, a new continue button would appear and, when clicked, the task would be presented. When the participant felt the task was completed, (s)he would click on another continue button and the NASA-TLX would appear. The procedure was the same for the demographic questionnaire and the GenderMag questionnaire.

In the end, we thanked the participant for taking the time to be part of the evaluation and answered any questions (s)he might have.

4.3 Deviations from the plan

During the modification task, there was a technical problem with the recording of the EEG data, which lost the connection with the computer twice during the collection process of 1 participant. Although the time that the collection was not made was only 11 seconds, we decided to still exclude the EEG data for that participant.

5 Analysis

5.1 Data set preparation

In each session, we recorded without pausing the video and audio. During the data collection process, we took special care not to disturb, or distract, our participants.

For the creation and modification tasks, we had two gold-standard models (one for each task), created and validated by experienced iStar 2.0 researchers. These models were iterated and changed based on the data analysis, in the case of a participant adding a particular model element that was useful and not covered by the initial gold-standard model. For the creation task, the final gold-standard model had 21 model elements (including actors' links). The final gold-standard model for the modification task had 19 model elements (also including actors' links). In the end, all the models created or modified by the participants were evaluated based on these gold-standard models. In this context, we can regard the model elements in the gold-standard model as a closed set, with a complete list of correct elements to be added to the model. When assessing each created or modified model, we counted which of the final gold-standard model elements were suitably represented in the participants' models. All elements provided by participants that were not found in the final gold-standard were counted as false positives. Although the final gold-standard model was achieved based on the data

analysis, the subjective nature of this assessment makes it possible that some of the not included elements might have been a valid extension of the gold-standard model. In the creation and modification tasks, the model creation tool collected all the elements added or modified by the participant in a CSV file. We compared the gold-standard file with the solution modelled by the participant.

Since the answers were given orally for the understanding and reviewing tasks, preparation of those data was also necessary. For the understanding tasks, we had a table with all the elements present in the model, one per column. When listening to the answers, elements that a participant described as being the correct ones were marked with 1, in a row dedicated to each participant. For the reviewing tasks, the procedure was the same, but when the answer was different from the expected, we added a column with that answer, if it was not already present. In the end, the table contained all the answers given by the participants and their frequency. These data allow us to analyse the participants' *accuracy*.

We watched the video with the audio and manually collected the times when the participant started and ended the tasks, as well as the first and last actions or detections. Since the participant had control over the session, as explained in Subsection 4.2, and the entire session was recorded in order to not disturb the participants, we needed to have the times when a participant clicked on the continue button and the task was presented, and the moment when a participant clicked on the next continue button to finish the task and go the NASA-TLX questionnaire. We also collected the timestamps for the clicks, but they were double-ckeced with the times in the video. These data allow us to analyse the participants' *speed*.

Concerning the eye-tracking data, the main areas of the stimulus and its elements were mapped into pixel coordinates to determine which regions and elements the participants were looking at, and saved in a CSV file. This enabled tagging the eye-tracking data with the elements being gazed at any given moment, which was a necessary step for computing the eye-tracking metrics. The fixations and corresponding durations were saved in a different CSV file, in order to calculate the normalised fixation durations. These data allow us to analyse the participants' *visual ease*.

Regarding the EEG and EDA scanners, the tools collecting the data save them in a CSV file. Those files have the structure needed to perform the analysis on the participant's *mental and emotional ease*, without further preparing the data. Similarly, no additional preparation is needed to analyse the participants' *perceived effort*, with NASA-TLX, nor to *characterise the participants*, with demographic data and GenderMag.

5.2 Analysis procedure

We started by collecting descriptive statistics on our variables, to get an overview of their distribution. For quantitative measurement, we collected the *mean*, *standard deviation*, *skewness* and *kurtosis*. We also used *box plots*, *Q-Q plots*, and *kernel density plots*, to help with the visual analysis of the distributions.

This was then complemented with *Welch t-tests*. A discussion on the benefits of using Welch t-test for comparing distributions to detect statistically significant differences in a robust way (as opposed to two samples t-test, or a non-parametric alternative to it, such as the Mann-Whitney U test) is in [48].

In terms of assessing accuracy, *Chi-square* would have been an adequate statistic, instead of computing the metrics precision, recall and f-measure and applying the Welch t-test. Chi-square is adequate for measuring differences in counts for a category of items, and for analysing the relationship between two nominal variables, or a nominal and an ordinal variable. In our context, this means contrasting the nominal problem-solving style (Abby *versus* Tim) with the frequency with which a given outcome is achieved. This works well if we are assessing a small number of possible outcomes. For instance, we could use two categories: the right answer *versus* the wrong answer. However, we would be losing information concerning how close participants were to getting the right answer. We could also create more categories for grouping the outcomes (e.g. divide into quartiles). In order not to lose information, we would need as many categories as possible outcomes. We computed the number of relevant and irrelevant elements created by the participants, in order to verify the results in terms of the Chi-square. Yet, given our gold-standard models, we have a great variation in terms of the number of relevant model elements (the categories) identified by the participants, spanning from 0 to 19 (for the modification task) or from 0 to 21 (for the creation tasks). This caused several categories to have less than 5 expected participants, which makes the application of the Chi-square not advisable, as the results might not be reliable. We also tested grouping the categories by quartiles and quintiles, but the distribution was not uniform and caused us to lose information with the partitions. Furthermore, the category grouping can introduce a validity threat related to the arbitrariness in the division. Given the characteristics of the dataset, using precision, recall and f-measure, and computing the Welch t-test, provides a better grasp of this relationship.

5.3 Descriptive statistics

In Table 8, we present the descriptive statistics for the metrics collected in our data analysis. For the sake of brevity, we only present the results concerning *precision*, which is related with the dependent variable *accuracy*. Due to its high number, the remainder of the data can be found in the paper's companion site [35].

For each metric, the first 10 lines refer to the creation task, the next 10 refer to the modification task, then 10 for the understanding task, and the last 10 are related with the reviewing task. In the *Facet* column, *Mot.* stands for motivation; *Inf. Proc.* for information processing; *S.E.* for self-efficacy; *Risk* for attitude towards risk; and *Learn.* for Learning style. For each facet, we divide them into personas (*Abby* and *Tim*). We further present the mean, standard deviation, skewness, kurtosis, and the p-value for the Shapiro-Wilk normality test.

The shape of the distributions suggests that, in some cases, normality is **not** a reasonable assumption ($p < 0.05$). The variance of the distributions is not similar, for several of these variables. The visual inspection of boxplot diagrams, Q-Q plots and kernel density plots (omitted for the sake of brevity) further reinforced our assessment concerning data normality.

Table 8: Descriptive statistics for precision.

	Task	Facet	Persona	Mean	S.D.	Skew.	Kurt.	S-W
Precision	Creation	Mot.	Abby	.466	.214	-.090	.368	.794
			Tim	.511	.216	.384	-.751	.051
		Inf. P.	Abby	.534	.203	.244	-.544	.106
			Tim	.351	.199	.541	1.680	.421
		S. E.	Abby	.529	.208	.195	-.620	.237
			Tim	.432	.217	.374	.535	.397
		Risk	Abby	.680	.270	-1.680	2.303	**.002**
			Tim	.422	.134	-.030	.119	.092
		Learn.	Abby	.520	.264	-.403	-.292	.689
			Tim	.486	.199	.509	-.200	.050
	Modification	Mot.	Abby	.579	.169	.101	1.608	.115
			Tim	.565	.278	-.371	-.105	.057
		Inf. P.	Abby	.626	.217	-.256	.775	**.046**
			Tim	.379	.228	-.624	-.613	.228
		S. E.	Abby	.621	.227	-.349	.641	.122
			Tim	.480	.246	-.339	.762	.206
		Risk	Abby	.766	.248	-1.180	.718	**.026**
			Tim	.495	.193	-1.259	1.887	**.000**
		Learn.	Abby	.612	.231	.016	-.532	.981
			Tim	.557	.245	-.440	.696	**.018**
	Understanding	Mot.	Abby	.705	.188	.261	-.922	.503
			Tim	.713	.260	-1.487	2.582	**.001**
		Inf. P.	Abby	.799	.160	-.324	-.403	**.020**
			Tim	.526	.269	-1.052	.751	**.026**
		S. E.	Abby	.746	.219	-1.773	5.517	**.002**
			Tim	.668	.256	-.837	1.272	.133
		Risk	Abby	.806	.212	-.692	-1.021	**.009**
			Tim	.654	.236	-1.672	3.339	**.001**
		Learn.	Abby	.540	.232	-1.238	3.262	.098
			Tim	.768	.213	-1.646	4.697	.000
	Reviewing	Mot.	Abby	.332	.248	.172	-1.462	**.021**
			Tim	.238	.163	.790	.461	.051
		Inf. P.	Abby	.350	.212	.530	-1.152	**.013**
			Tim	.231	.209	.659	-.746	.003
		S. E.	Abby	.381	.242	.098	-1.577	.107
			Tim	.241	.189	.608	-.540	.008
		Risk	Abby	.375	.249	-.266	-1.368	**.010**
			Tim	.186	.103	.267	-.774	.060
		Learn.	Abby	.296	.221	.649	-.687	**.028**
			Tim	.285	.218	.422	-1.023	.093

5.4 Hypotheses testing

We used Welch's t-test, as it is robust to deviations from the normal distribution, different sample sizes, and variance in the samples, thus following the recommendations on data analysis for Software Engineering empirical evaluations [48]. We are using $p < 0.05$ for the level of significance and thus rejecting the null hypothesis.

For the sake of brevity, we only presented the results for the hypotheses testing that are statistically significant. Due to its high number, the remainder of the data can be found in the paper's compation site [35].

RQ1: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing creation tasks on iStar 2.0 models?*

In Table 9 we summarise the Welch t-test results for the *creation* task, as well as present the mean and standard deviation for both Abby and Tim, for all the metrics that had a statistically significant difference. The only exception is the *motivation* facet, where we found no statistical evidence of differences between participants identified as Abby and the ones identified as Tim.

Table 9: Hypothesis testing for the *creation* task

Facet	**Metric**	**Mean**		**S.D.**		**Sig.**
		Abby	**Tim**	**Abby**	**Tim**	
Motivation	–	–	–	–	–	–
Information processing	Precision	.534	.351	.203	.199	.016
	LastAct	1262.692	2026.64	599.587	797.359	.011
	ProcDur	440.872	72.818	310.583	101.016	.000
	FixIrrel	5.948	1.471	4.362	1.657	.000
	AvgDurRelFix	506.559	907.655	202.686	325.390	.002
	AvgDurIrrelFix	879.526	468.981	346.001	378.193	.005
	AvgAttention	.785	.600	.163	.215	.020
	AvSCL	768.744	868.636	138.863	99.617	.014
	HRVarNN50	23.625	36.810	20.380	15.996	.035
	NASA-TLX	75.855	51.152	18.931	19.545	.002
Self-efficacy	Duration	1596.063	2136.611	551.620	683.933	.008
	LastAct	1067.531	2076.500	423.516	674.687	.000
	ProcDur	528.531	60.111	271.804	81.166	.000
	FixIrrel	6.187	2.787	4.601	2.799	.002
	AvgDurRelFix	463.663	827.932	173.218	301.122	.000
	AvgMentWL	.7563	.578	.180	.180	.002
	NASA-TLX	78.208	56.574	18.264	20.207	.001
Risk	Precision	.680	.422	.270	.134	.004
	Recall	.440	.678	.230	.216	.003
	FirstAct	343.429	189.333	253.290	133.427	.046
	NASA-TLX	82.571	65.694	16.578	21.534	.006
Learning style	Duration	2471.333	1575.711	716.492	458.800	.001
	FirstAct	487.250	152.026	172.717	96.349	.000
	LastAct	2001.000	1250.684	878.572	555.106	.015
	FixIrrel	3.020	5.577	2.840	4.574	.028
	NASA-TLX	87.750	64.947	15.762	20.268	.000

Again for the sake of brevity, we will only illustrate how to present the results of the hypotheses testing by describe the results for the *information processing* facet, as follows.

There was a statistically significant difference in variables concerning the *accuracy*, *speed* and *ease*. The *precision* achieved by participants identified as Abby in the information processing facet was higher ($M = .534$, $SD = .203$) than the one achieved by participants identified as Tim ($M = .351$, $SD = .199$, $t(1) = 7.208$, $p = .016$). The time for performing the *last action* was lower for Abby ($M = 1262.692$, $SD = 599.587$) than for Tim ($M = 2026.64$, $SD = 797.359$, $t(1) = 8.708$, $p = .011$). The *number of irrelevant fixations* was higher for Abby ($M = 5.948$, $SD = 4.362$) than for Tim ($M = 1.471$, $SD = 1.657$, $t(1) = 27.178$, $p = .000$). The *average duration of relevant fixations* was lower for Abby ($M = 506.559$, $SD = 202.686$) Tim ($M = 907.655$, $SD = 325.390$, $t(1) = 15.065$, $p = .002$).

On the other hand, The *average duration of irrelevant fixations* was higher for Abby ($M = 879.526$, $SD = 346.001$) than for Tim ($M = 468.981$, $SD = 378.193$, $t(1) = 10.487$, $p = .005$). The *average attention* was higher for Abby ($M = .785$, $SD = .1631$) than for Tim ($M = .600$, $SD = .214$, $t(1) = 7.007$, $p = .020$). The *average skin conductive level* was lower for Abby ($M = 768.744$, $SD = 138.863$) than for Tim ($M = 868.636$, $SD = 99.617$, $t(1) = 7.145$, $p = .014$). The *heart rate variability* (for NN50) was lower for Abby ($M = 23.625$, $SD = 20.380$) than for Tim ($M = 36.810$, $SD = 15.996$, $t(1) = 5.126$, $p = .035$). Finally, the *perceived effort* was higher for Abby ($M = 75.855$, $SD = 18.931$) than for Tim ($M = 51.152$, $SD = 19.545$, $t(1) = 13.895$, $p = .002$).

RQ2: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing modification tasks on iStar 2.0 models?*

In Table 10 we summarise the Welch t-test results for the *modification* task, as well as present the mean and standard deviation for both Abby and Tim, for all the metrics that had a statistically significant difference.

Table 10: Hypothesis testing for the *modification* task

Facet	Metric	Mean		S.D.		Sig.
		Abby	Tim	Abby	Tim	
Motivation	Precision	.679	.565	.169	.278	.048
Information processing	Precision	.626	.375	.217	.228	.005
	FixIrrel	5.111	.253	4.350	.356	.000
	AvgDurRelFix	379.207	835.262	273.123	397.556	.003
	AvgDurIrrelFix	704.546	309.635	454.026	341.011	.005
	AvgMentWL	.721	.564	.188	.1963	.032
	AvgAttention	.721	.473	.163	.179	.001
	HRVarRMSSD	64.168	43.794	22.029	17.993	.045
Self-efficacy	Precision	.621	.480	.226	.246	.045
	Duration	1086.686	1647.722	571.192	683.284	.006
	ProcDur	414.125	210.111	254.964	235.163	.007
	FixIrrel	5.360	1.698	4.678	2.301	.001
	AvgDurRelFix	347.516	714.247	279.480	362.099	.001
	AvgDurIrrelFix	684.150	499.472	457.580	450.531	.018
Risk	Precision	.766	.495	.248	.193	.002
	FirstAct	268.857	127.750	160.808	91.590	.007
	HRVarRMSSD	32.307	48.667	15.493	21.288	.005
Learning style	FixIrrel	4.559	2.404	4.626	2.841	.042
	AvgDurRelFix	424.178	654.849	274.552	515.699	.036
	AvgDurIrrelFix	812.657	556.089	547.578	417.058	.046

RQ3: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing understanding tasks on iStar 2.0 models?*

In Table 11 we summarise the Welch t-test results for the *understanding* task, as well as present the mean and standard deviation for both Abby and Tim, for all the metrics that had a statistically significant difference.

Table 11: Hypothesis testing for the *understanding* task

Facet	Metric	Mean		S.D.		Sig.
		Abby	Tim	Abby	Tim	
Motivation	Duration	746.692	630.259	190.005	220.225	.044
	NASA PD	26.154	16.111	23.993	18.257	.042
	NASA TD	33.077	29.259	20.263	23.234	.048
	NASA MD	70.185	59.231	9.352	12.221	.010
	NASA Frust	50.385	44.333	24.871	23.205	.032
Information processing	Precision	.800	.526	.159	.269	.004
	Recall	.802	.669	.192	.345	.032
	ProcDur	32.462	19.667	15.197	15.750	.021
	FixIrrel	5.286	.954	4.952	.442	.000
	AvgDurRelFix	153.963	194.154	57.581	32.844	.008
	AvgDurIrrelFix	557.462	436.889	98.531	172.742	.008
	AvgMentWL	.593	.385	.234	.172	.003
	AvgAttention	.607	.354	.159	.166	.000
	HRVarRMSSD	39.924	31.328	21.717	15.268	.016
	NASA PD	33.462	12.593	22.396	15.955	.014
	NASA MD	69.259	61.154	12.224	7.403	.007
	NASA Eff	67.037	59.615	15.333	11.266	.044
Self-efficacy	FirstDet	529.833	428.364	175.756	148.394	.036
	FixIrrel	4.911	2.614	4.939	3.758	.010
	AvgDurRelFix	150.693	186.986	56.832	43.693	.028
	AvgDurIrrelFix	535.958	427.080	131.080	170.496	.028
	AvgMentalWL	.591	.444	.229	.223	.048
	AvgAttention	.559	.483	.194	.204	.040
	NASA PD	23.333	16.136	21.828	19.330	.028
	NASA MD	69.546	63.056	11.742	10.310	.041
Risk	Precision	.806	.654	.212	.236	.043
	FirstDet	548.933	429.080	221.712	105.405	.045
	HRVarRMSSD	23.929	45.051	7.551	21.142	.000
	NASA PD	30.333	12.800	24.818	14.367	.022
	NASA TD	42.667	23.200	26.313	15.604	.017
	NASA Frust	58.333	43.400	26.027	20.296	.039
Learning style	Duration	736.100	645.433	167.655	227.318	.019
	FirstDet	585.400	436.900	161.670	154.124	.023
	FixIrrel	4.870	.900	4.861	.382	.000
	AvgDurRelFix	161.358	184.025	56.830	41.914	.039
	AvgDurIrrelFix	527.075	459.075	125.742	170.489	.048
	Sac2Key	62.000	59.200	8.028	9.080	.037
	NASA PD	31.000	15.500	22.828	18.539	.044

RQ4: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing reviewing tasks on iStar 2.0 models?*

In Table 12 we summarise the Welch t-test results for the *reviewing* task, as well as present the mean and standard deviation for both Abby and Tim, for all the metrics that had a statistically significant difference. The only exception is the *motivation* facet, where we found no statistical evidence of differences between participants identified as Abby and the ones identified as Tim.

Table 12: Hypothesis testing for the *reviewing* task

Facet	Metric	Mean		S.D.		Sig.
		Abby	Tim	Abby	Tim	
Motivation	–	–	–	–	–	–
Information processing	Precision	.350	.231	.212	.209	.042
	FixIrrel	6.502	3.143	4.372	1.239	.003
	AvgDurRelFix	782.606	1345.912	187.822	314.759	.000
	AvgDurIrrelFix	1234.623	881.065	340.453	409.4403	.005
	AvgMentWL	.815	.580	.1725	.174	.000
	AvgAttention	.830	.700	.153	.178	.018
	HRVarNN50	50.569	29.749	13.898	20.051	.001
Self-efficacy	Precision	.381	.241	.242	.189	.042
	Duration	1721.000	1966.923	755.369	834.556	.035
	FixIrrel	5.756	4.319	4.059	3.300	.027
	AvgDurRelFix	895.964	1154.878	342.613	378.933	.036
	AvgDurIrrelFix	1263.883	946.900	435.575	360.923	.029
	AvgMentWL	.757	.665	.183	.217	.017
	AvgAttention	.886	.700	.103	.1743	.000
	NASA TD	59.286	48.462	31.856	24.322	.028
	NASA MD	96.429	83.462	7.703	14.951	.001
	NASA Frust	90.000	65.000	15.317	31.528	.002
	NASA Eff	86.786	72.308	13.951	19.608	.011
Risk	Precision	.375	.186	.249	.103	.003
	Recall	.253	.538	.201	.190	.000
	FirstDet	778.818	490.389	193.621	81.657	.000
	FixIrrel	6.941	4.677	4.494	2.174	.048
	AvgDurRelFix	862.229	1229.554	259.491	393.113	.001
	AvgDurIrrelFix	1135.823	994.043	424.176	401.173	.029
	NASA TD	62.955	51.389	27.109	28.274	.049
	NASA Frust	85.227	59.722	14.677	36.480	.011
Learning style	FirstDet	705.750	592.300	210.544	199.112	.048

6 Discussion

6.1 Evaluation of results and implications

RQ1: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing creation tasks on iStar 2.0 models?*

We found no evidence that the *motivation* facet impacts the accuracy, speed or ease for the creation task.

Assessing accuracy. Participants identified as Abby in the *information processing* and in the *risk* facets had a higher precision when compared with those identified as Tim. However, recall for Abby in the *risk* facet was lower. Our interpretation is that Tim is able to achieve a higher recall because he is risk-tolerant, and takes a chance even when he is not sure. Yet, this causes his precision to be lower. Abby is risk-averse and only answers when she is sure. As such, when she answers, her answer tends to be correct, but incomplete (she does not add a model element if she is not absolutely confident). As such, the iStar 2.0 models created by Abby in the *risk* facet were less complex, but also less complete.

Assessing speed. We found that Abbys in the *learning style* were slower than the ones characterised as Tim, taking ≈ 10 minutes more in the overall *duration* of the task. However, Abby in the *self-efficacy* facet took less time to complete the task. Our interpretation for the latter is that, without someone to first show how tasks of this type could be performed, Abby felt she had already given her best and decided to finish the task earlier. For the former, since Tim tends to have a tinkering approach (playfully experimenting with the tool and the model elements), this may help him to be faster. Note that Tim in the *risk* and *learning style* facets makes the first action in the model really early. In fact, he starts trying to solve the task even before finishing reading the problem description. As for Abby, she only starts after some time. Finally, in the *self-efficacy* facet, the processing duration was lower for Tim. This means that, after the creation of the models, Tim submits it without performing a revision. We argue that this is due to his high confidence in his work.

Assessing ease. There was a greater visual effort for Abby in the *information processing* and *self-efficacy* facets, observable through a higher number of irrelevant fixations and average duration of irrelevant fixations. However, Abby has a lower average duration of relevant fixations. Our interpretation is that Tim, being more selective in the way he processes information, is able to focus more on the relevant fixations. As for Abby, she tends to further analyse the information provided, hence the higher number and average duration of irrelevant fixations. There was a greater mental effort for Abby in *information processing* and *self-efficacy* facets, observable through higher average attention (for information processing) and a higher average mental workload (for self-efficacy). Since Abby is more comprehensive when processing information, her level of attention indicates she is engaged in the task. Similarly, given that she as a low self-efficacy, her mental workload becomes higher, indicating effort while performing the task. There was also a greater cognitive load for Abby in the *information processing* facet, observable through a higher average skin conductive level. For the same facet, Tim's heart rate variability was higher. Our interpretation is that Tim was excited when performing the task. In all the facets (except *motivation*), there was a greater perceived effort for Abby than for Tim. Participants characterised as Abby in the *learning style* and *risk* facets had a higher perceived physical demand than the ones characterised as Tim. The perceived temporal demand was also higher for Abby in the *self-efficacy*, *learning style*, and *risk* facets. The perceived mental demand was higher for Abby in the information processing facet. Finally, the frustration was higher for Abby in all the facets. This is in line with the results obtained in terms of accuracy, speed, and biometric data, meaning the participants were well aware of their performance and effort on the task.

Summary of the results. In Table 13 we summarise the results for the *creation* task. For each persona, we indicate if the result was higher or lower, in comparison with the other persona. Higher and lower values can be interpreted differently depending on the metric, so the results considered better were highlighted in green, and the ones considered worst were highlighted in red. For example, having a higher precision is considered better, but having a lower NASA-TLX is considered better as well, since it means the participants had a lower perceived effort. We followed the same procedure for the remaining research questions.

Table 13: Summary of the results for the *creation* task

Facet	Metric	Persona	
		Abby	Tim
Motivation	–	–	–
Information processing	Precision	Higher	Lower
	LastAct	Lower	Higher
	ProcDur	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgAttention	Higher	Lower
	AvSCL	Lower	Higher
	HRVarNN50	Lower	Higher
	NASA-TLX	Higher	Lower
Self-efficacy	Duration	Lower	Higher
	LastAct	Lower	Higher
	ProcDur	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgMentWL	Higher	Lower
	NASA-TLX	Higher	Lower
Risk	Precision	Higher	Lower
	Recall	Lower	Higher
	FirstAct	Higher	Lower
	NASA-TLX	Higher	Lower
Learning style	Duration	Higher	Lower
	FirstAct	Higher	Lower
	LastAct	Higher	Lower
	FixIrrel	Lower	Higher
	NASA-TLX	Higher	Lower

RQ2: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing modification tasks on iStar 2.0 models?*

Assessing accuracy. Participants identified as Abby in the *motivation*, *self-efficacy*, *information processing*, and *risk* facets had a higher precision when compared with those identified as Tim. However, there were no differences in terms of recall. Still, some patterns emerged, as in the creation task. Abby is risk-averse and only answers when she is sure. Her answer tends to be correct, but incomplete (she does not change, add, or remove a model element if she is not absolutely confident). As such, the complexity of the iStar 2.0 SR models modified by Abby in the *risk* facet was lower, but the completeness of those models was lower as well. In fact, in the *risk* facet, Abby tended to make fewer changes than Tim.

Assessing speed. Participants identified as Abby in the *self-efficacy* facet took less time to complete the task when compared with those identified as Tim. Our interpretation is that, without someone to first show how tasks of this type could be performed, Abby felt she had already given her best and decided to finish the task earlier. Tim in the *risk* facet makes the first action in the model really early. In fact, he starts trying to solve the task even before finishing reading the problem description. As for Abby, she only starts after some time. Finally, in the *self-efficacy* facet, the processing duration was lower for Tim. This means that,

after the modification of the model, Tim submits it without performing a revision. We argue that this is due to his high confidence in his work.

Assessing ease. Participants characterised as Abby in the *information processing*, *self-efficacy*, and *learning style* facets had a greater visual effort, observable through a higher fixation rate on irrelevant elements and average duration of irrelevant fixations. However, Abby had a lower average duration of relevant fixations. Our interpretation is that Tim, being more selective in the way he processes information, is able to focus more on the relevant elements. As for Abby, she tends to further analyse the information provided, hence the focus on the irrelevant elements. There was a greater mental effort for Abby in the *information processing* facet, observable through higher average attention and average mental workload. Since Abby is more comprehensive when processing information, her level of attention indicates she is highly engaged in the task. There was also a difference in terms of cognitive load. Participants characterised as Tim had a higher heart rate variability, for RMSSD, than the ones identified as Abby. Since Tim is risk-tolerant, our interpretation is that Tim was excited when performing the task. On the other hand, participants identified as Abby in the *information processing* style had a higher heart rate variability, for RMSSD, than the ones characterised as Tim. Given that Abby is more comprehensive when processing information, we argue that the number of model elements to analyse and possibly change might have made her feel more stressed and anxious. There was no difference in terms of perceived effort for all the facets. Our interpretation is that modification tasks are perceived as easier than creation tasks.

Summary of the results. In Table 14 we summarise the results for the *modification* task.

Table 14: Summary of the results for the *modification* task

Facet	Metric	Mean	
		Abby	**Tim**
Motivation	Precision	Higher	Lower
Information processing	Precision	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgMentWL	Higher	Lower
	AvgAttention	Higher	Lower
	HRVarRMSSD	Higher	Lower
Self-efficacy	Precision	Higher	Lower
	Duration	Lower	Higher
	ProcDur	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
Risk	Precision	Higher	Lower
	FirstAct	Higher	Lower
	HRVarRMSSD	Lower	Higher
Learning style	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower

RQ3: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing understanding tasks on iStar 2.0 models?*

Assessing accuracy. Participants characterised as Abby in the *information processing* and *risk* facets have higher precision when compared with those identified as Tim. Furthermore, Abby in the *information processing* facet also had a higher recall. There were no differences in terms of *recall* for the risk facet. Our interpretation is that analysing the iStar 2.0 SR model comprehensively helped Abby to better understand it.

Assessing speed. Participants characterised as Abby in the *learning style* and *motivation* facets were slower than the ones characterised as Tim, taking $\approx$ 4 minutes more in the overall duration of the task. Since Tim tends to have a tinkering approach, we argue this may help him to be faster. Tim in the *risk*, *learning style* and *self-efficacy* facets makes the first detection in the model really early. In fact, he starts trying to solve the task even before finishing reading the question. As for Abby, she only starts after some time. Finally, in the *information processing* facet, the processing duration is higher for Abby. Our interpretation is that, since Abby is comprehensive when analysing information, she prefers to revise the model to make sure that nothing was forgotten.

Assessing ease. Participants characterised as Abby in the *information processing*, *self-efficacy*, and *learning style* facets had a greater visual effort, observable through a higher fixation rate on irrelevant elements and average duration of irrelevant fixations. However, Abby had a lower average duration of relevant fixations. Our interpretation is that Tim, being more selective in the way he processes information, is able to focus more on the relevant elements. As for Abby, she tends to further analyse the information provided, hence the focus on the irrelevant elements. Furthermore, participants characterised as Abby in the *learning style* facet had a higher total number of saccades to the key. We argue that Abby looked more to the key in order to make sure she completely understood the elements in the element, in order to be able to select the correct one. There was a greater mental effort for Abby in the *information processing* and *self-efficacy* facets, observable through a higher average attention and average mental workload. Since Abby is more comprehensive when processing information, her level of attention indicates she is highly engaged in the task. Similarly, given that she as a low *self-efficacy*, her mental workload becomes higher, indicating effort while performing the task. There was also a difference in terms of cognitive load. Participants characterised as Tim in the *risk* facet had a higher heart rate variability, for RMSSD, than the ones identified as Abby. Since Tim is risk-tolerant, our interpretation is that Tim was excited when performing the task. On the other hand, participants identified as Abby in the *information processing* style had a higher heart rate variability, for RMSSD, than the ones characterised as Tim. Given that Abby is more comprehensive when processing information, we argue that the number of model elements to analyse might have made her feel more stressed and anxious. In all the facets, there was a greater perceived effort for Abby than for Tim. Participants characterised as Abby in the *motivation*, *self-efficacy*, *learning style*, *information processing*, and *risk* facets had a higher perceived physical demand than the ones characterised as Tim. However, it was lower than 40 (out of 100) for both Abby and Tim. The perceived temporal demand was also higher for Abby in the *motivation*, and *risk* facets. The perceived mental demand was higher for Abby in the *motivation*, *self-*

efficacy, and *information processing* facets. Finally, the frustration was higher for Abby in the *motivation* and *risk* facets, while the perceived effort was higher for Abby in the *information processing* facet. This is in line with the results obtained in terms of speed, and biometric data, meaning the participants were well aware of their effort on the task.

Summary of the results. In Table 15 we summarise the results for the *understanding* task.

Table 15: Summary of the results for the *understanding* task

Facet	Metric	Mean	
		Abby	Tim
Motivation	Duration	Higher	Lower
	NASA PD	Higher	Lower
	NASA TD	Higher	Lower
	NASA MD	Higher	Lower
	NASA Frust	Higher	Lower
Information processing	Precision	Higher	Lower
	Recall	Higher	Lower
	ProcDur	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgMentWL	Higher	Lower
	AvgAttention	Higher	Lower
	HRVarRMSSD	Higher	Lower
	NASA PD	Higher	Lower
	NASA MD	Higher	Lower
	NASA Eff	Higher	Lower
Self-efficacy	FirstDet	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgMentalWL	Higher	Lower
	AvgAttention	Higher	Lower
	NASA PD	Higher	Lower
	NASA MD	Higher	Lower
Risk	Precision	Higher	Lower
	FirstDet	Higher	Lower
	HRVarRMSSD	Lower	Higher
	NASA PD	Higher	Lower
	NASA TD	Higher	Lower
	NASA Frust	Higher	Lower
Learning style	Duration	Higher	Lower
	FirstDet	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	Sac2Key	Higher	Lower
	NASA PD	Higher	Lower

RQ4: *Does a difference in the level of each facet impact the accuracy, speed and ease when performing reviewing tasks on iStar 2.0 models?*

We found no evidence that the *motivation* facet impacted the accuracy, speed, or ease for the reviewing task.

Assessing accuracy. Participants identified as Abby in the *information processing*, *self-efficacy*, and *risk* facets have higher precision when compared with those identified as Tim. However, recall for Abby in the *risk* facet was lower. Our interpretation is that Tim is able to achieve a higher recall because he is risk-tolerant, and takes a chance even when he is not sure. Yet, this causes his precision to be lower. Abby is risk-averse and only answers when she's sure. As such, when she answers, her answer tends to be correct, but incomplete.

Assessing speed. Participants identified as Abby in the *self-efficacy* facet took less time to complete the task than Tim. Our interpretation is that, without someone to first show how tasks of this type could be performed, Abby felt she had already given her best and decided to finish the task earlier. Tim in the *risk* and *learning style* facets starts answering the defects found in the model really early. As for Abby, she only starts after some time.

Assessing ease. Participants characterised as Abby in the *information processing*, *self-efficacy*, and *risk* facets had a greater visual effort, observable through a higher fixation rate on irrelevant elements and average duration of irrelevant fixation. However, Abby had a lower average duration of relevant fixations. Our interpretation is that Tim, being more selective in the way he processes information, is able to focus more on the relevant elements. As for Abby, she tends to further analyse the information provided, hence the focus on the irrelevant elements. There was a greater mental effort for Abby in the *information processing* and *self-efficacy* facets, observable through a higher average attention and average mental workload. Since Abby is more comprehensive when processing information, her level of attention indicates she is highly engaged in the task. There was also a difference in terms of cognitive load. Participants identified as Abby in the *information processing* style had a higher heart rate variability, for NN50, than the ones characterised as Tim. Given that Abby is more comprehensive when processing information, we argue that the number of model elements to analyse might have made her feel more stressed and anxious. Participants characterised as Abby in the *self-efficacy* and *risk* facets had a higher perceived temporal demand than the ones characterised as Tim. The perceived mental demand was higher for Abby in the *self-efficacy* facet. Finally, the frustration was higher for Abby in the *self-efficacy* and *risk* facets, while the perceived effort was higher for Abby in the *self-efficacy* facet. This is in line with the results obtained in terms of biometric data, meaning the participants were well aware of their effort on the task.

Summary of the results. In Table 16 we summarise the results for the *reviewing* task.

6.2 Threats to validity

For the identification of the threats to validity, we are following Wohlin et al.'s guidelines [98].

Internal validity. We used a combination of convenience and snowball sampling. This can cause a *selection* threat, since the participants tend to be more

Table 16: Summary of the results for the *reviewing* task

Facet	Metric	Mean	
		Abby	**Tim**
Motivation	–	–	–
Information processing	Precision	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgMentWL	Higher	Lower
	AvgAttention	Higher	Lower
	HRVarNN50	Higher	Lower
Self-efficacy	Precision	Higher	Lower
	Duration	Lower	Higher
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	AvgMentWL	Higher	Lower
	AvgAttention	Higher	Lower
	NASA TD	Higher	Lower
	NASA MD	Higher	Lower
	NASA Frust	Higher	Lower
	NASA Eff	Higher	Lower
Risk	Precision	Higher	Lower
	Recall	Lower	Higher
	FirstDet	Higher	Lower
	FixIrrel	Higher	Lower
	AvgDurRelFix	Lower	Higher
	AvgDurIrrelFix	Higher	Lower
	NASA TD	Higher	Lower
	NASA Frust	Higher	Lower
Learning style	FirstDet	Higher	Lower

motivated to be part of the experiments, considering that their participation is entirely voluntary. However, we found no evidence of this in the results. Furthermore, we plan to launch a replication of this experiment with participants selected through a recruitment call, and we have made available an independent replication package to colleagues from other organisations and countries.

Conclusion validity. Although we have a significantly high number of participants, higher than most sample sizes reported, in particular, in other eye-tracking experiments (see [76]), the sample size is always a risk, as results may not apply to even larger populations. We encourage replications of the quasi-experiment with a larger group. However, the distribution of participants on the GenderMag facets was not balanced. The distribution of participants to tasks did not take into account their facets, which may have influenced the results. Future replications can ask participants to first reply to the GenderMag questionnaire, and then assign the tasks in a way that the facets are evenly distributed across them.

External validity. Overall, our participants had little to no prior knowledge in i^* or iStar 2.0, and they skew young (with an average of 27 years old). Although this made them representatives of stakeholders with low requirements engineering expertise, by having participants with a greater level of experience, and with a

wider range of age groups, we could analyse the differences between these profiles. Further research is needed to assess how different facet levels in experienced *i** or iStar 2.0 users would impact the results, as well as in more mature participants. Furthermore, the iStar 2.0 models used in the understanding and reviewing tasks were relatively small, with only 2 actors and 25 elements (with 11 inside each actor, and 3 dependums). The problem description of the creation and modification tasks was also simple, in order to produce a small model as well. These models may not be representative of the ones used in industry, thus introducing an *interaction of setting and treatment* threat. In the performed quasi-experiments, we could not use larger models since we were limited by the technical specifications of the eye-tracker device, such as constraints in the external monitor dimensions and in the participant distance to the eye-tracker. The fonts and symbols used had to be big enough for easy visualisation by all participants. As such, the tested models are fragments of larger ones. Notwithstanding, presenting only model fragments to focus the attention of the stakeholders is a common technique for improving communication with them. Even so, in a future replication, it is important to vary the complexity of these models, to assess whether there is a significant variation on the success and effort on the tasks as models become more complex. Moreover, we only analysed one domain: a booking management system for a hotel. We opted for a relatively known domain in order to reduce the effect of the results being related to difficulties in understanding the domain itself, and not due to the requirements languages that were under study. We are also aware that tacit knowledge may play an important role in the performance of the participants. However, our goal was to evaluate the requirements languages, thus reducing confounding effects was considered a priority. Finally, all tasks were in English. However, our participants have Portuguese as their mother tongue. We decided to create all the materials in English so they could be used in independent replications by international researchers. However, limited English proficiency could have impacted the results. Nevertheless, all the participants were at ease with the English language and we found no impact of this decision in the results obtained.

Construct validity. We showed a video tutorial about iStar 2.0, and afterwards participants were asked to create, modify, understand or review an iStar 2.0 model, so they might have felt that they were being evaluated. This may have caused an *evaluation apprehension* threat, where participants try to look better. To mitigate this threat, we have not informed them about what exactly was being tested, that is, their accuracy, speed and ease in the performed tasks. Furthermore, we are aware that measuring precision and recall over a set of possible right answers is potentially problematic. However, given the characteristics of our dataset, explained in Subsection 5.2, we argue these were the appropriate metrics to use in our context. There is a risk that the final gold-standard model is not complete and there may be further valid extensions that were not considered. We mitigate this threat by having experienced iStar 2.0 researchers creating and validating the gold-standard models, as well as changing the models based on the data analysis.

6.3 Inferences

It's not easy to review an iStar 2.0 SR model. Our participants really struggled when reviewing the models. When compared to the other tasks, participants

achieved a much lower precision and recall in the reviewing task, with the results being lower than 40% in the majority of the facets, and independently on the persona. All the other tasks had precision and recall higher than 50%. However, this was somewhat expected. Reviewing a model can be hard, since it involves not only reasoning about what the model represents, but also about what it does not represent (and should), and what is misrepresented. In general, participants had little to no prior knowledge on iStar 2.0, although some participants had learnt it in the context of a course. We found no statistically significant difference in the performance for these two profiles. However, we are confident that, with proper training, participants would be able to achieve a higher performance. Nonetheless, the obtained results can also mean that iStar 2.0 is possibly not a good suit for communication with stakeholders not knowledgeable on the language, even through the results for the understanding task were good.

Information processing and risk have impact on accuracy. Participants identified as Abby in these facets were able to achieve an acceptable level of precision, even without much training. However, her attitude towards risk is undermining the recall. We argue that, with training, Abby would become more confident in her skills and could achieve great results for both precision and recall. As for Tim, making him aware that risking too much is possibly sabotaging his results could help with his precision.

Information processing, self-efficacy, risk and learning style have impact on speed. Participants characterised as Abby in these facets tends to take longer to act upon the model, because she's collecting the highest possible number of information. When she finishes the task at hand, Abby revises the model and reads the problem description again. As for Tim, he tends to submit the model without any further review. We argue that a lower duration is not always a desirable outcome, if it compromises the accuracy of task, which we interpret has being the Tim's case. By not revising the model, Tim may be losing an opportunity for improvement and for higher precision.

Information processing, self-efficacy and risk have impact on ease. Participants identified as Abby in these facets has a more comprehensive analysis of the problem description and the model elements available in the editor's toolbar or in the language key. The visual effort, attention and mental workload is higher due to this thorough inspection. Plus, in general, Abby is more engaged at the task she is performing. Tim, however, is able to better separate what is relevant from what is not, and he is more confident on his skills and overall performance on the tasks (even though, in some cases, the perceived performance was not in line with the accuracy results). We argue that, in this particular scenario, having a higher effort is not perceived as being harmful. Nonetheless, being able to more precisely understand what is relevant is a great advantage in terms of effort.

People diversity is key. The complementarity of results achieved by Abby and Tim suggests that, rather than targeting the requirements process to one of them, there is more to be gained in leveraging their diversity. One possible way of doing so would be to build up teams with this diversity in terms of information processing, self-efficacy and risk.

7 Conclusions and future work

We performed a quasi-experiment to report the impact of different levels in each of the five GenderMag facets, when creating, modifying, understanding, or reviewing iStar 2.0 models. We measured the accuracy, speed and ease of a total of 180 participants. We used metrics of task success, time, and effort, collected with eye-tracking, EEG and EDA sensors, and participants' feedback through a NASA-TLX questionnaire. The data collected showed that participants with a comprehensive information processing style and a conservative attitude towards risk (characteristics more frequently seen in females) took longer to start performing the tasks but had a higher accuracy. The visual and mental effort was also higher for these participants. The complementarity of results suggests there is more gain in leveraging people's diversity.

As of the writing of this paper, the *i** standardisation process is not yet concluded, and there is a need for studies about iStar 2.0 ease of use, adequacy for teaching, expressiveness, graphical notation, automated reasoning techniques, among others [21]. Our work can help in this validation process. The results showed that different problem-solving styles had an impact on the performed tasks. Since our participants had little to no prior knowledge in *i** or iStar 2.0 and they all watched the same video tutorial on the language, we argue that the different problem-solving styles should be taken into consideration when teaching iStar 2.0.

In terms of the generalisability of our results, and although they were in line with those presented in previous work regarding the usage of the GenderMag questionnaire, it is still necessary to assess how consistently our results occur with other users, problem descriptions, models, and even with other requirements tools and artefacts. However, in an initial effort for generalisation, we have applied the same techniques to use cases [33] and user stories [63], with similar results. Furthermore, we need to study the real impact of people's diversity and how different people complement each other in the context of teams.

We plan to replicate the experiment in other contexts, and apply it to bigger and more complex descriptions. Furthermore, we plan to study the real impact people's diversity and how different people complement each other in the context of teams.

Acknowledgements We thank NOVA LINCS UID/CEC/04516/2019 and FCT-MCTES SFRH/BD/108492/2015 for financial support.

References

1. Andreassi, J.L.: Psychophysiology: Human Behavior & Physiological Response. Psychology Press (2013)
2. Appel, M., Kronberger, N., Aronson, J.: Stereotype threat impairs ability building: Effects on test preparation among women in science and technology. European Journal of Social Psychology **41**(7), 904–913 (2011)
3. Basili, V.R., Rombach, H.D.: The TAME project: Towards improvement-oriented software environments. IEEE Trans. Software Eng. **14**(6), 758–773 (1988)
4. Beckwith, L., Burnett, M., Wiedenbeck, S., Cook, C., Sorte, S., Hastings, M.: Effectiveness of end-user debugging software features: Are there gender issues? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 869–878. ACM (2005)

5. Beckwith, L., Kissinger, C., Burnett, M., Wiedenbeck, S., Lawrance, J., Blackwell, A., Cook, C.: Tinkering and gender in end-user programmers' debugging. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 231–240. ACM (2006)
6. BioSignalsPlux Wristband: (2019). URL `https://biosignalsplux.com/`. (Last access: May, 2020)
7. BITalino: (2019). URL `http://bitalino.com/`. (Last access: May, 2020)
8. Burnett, M., Counts, R., Lawrence, R., Hanson, H.: Gender hci and microsoft: Highlights from a longitudinal study. In: 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 139–143. IEEE (2017)
9. Burnett, M., Fleming, S.D., Iqbal, S., Venolia, G., Rajaram, V., Farooq, U., Grigoreanu, V., Czerwinski, M.: Gender differences and programming environments: across programming populations. In: Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement, pp. 1–10 (2010)
10. Burnett, M., Horvath, A., Oleson, A.: GenderMag Personas Foundations Document. URL `http://eusesconsortium.org/gender/GenderMagPersona-FoundationDocuments/Foundations.html`. (Last access: May 2020)
11. Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., Jernigan, W.: GenderMag: A method for evaluating software's gender inclusiveness. Interacting with Computers **28**(6), 760–787 (2016)
12. Burnett, M.M., Beckwith, L., Wiedenbeck, S., Fleming, S.D., Cao, J., Park, T.H., Grigoreanu, V., Rector, K.: Gender pluralism in problem-solving software. Interacting with computers **23**(5), 450–460 (2011)
13. Byrnes, J.P., Miller, D.C., Schafer, W.D.: Gender differences in risk taking: a meta-analysis. Psychological bulletin **125**(3), 367 (1999)
14. Cafferata, P., Tybout, A.M.: Gender differences in information processing: a selectivity interpretation. Cognitive and Affective Responses to Advertising, Lexington Books (1989)
15. Cagiltay, N.E., Tokdemir, G., Kilic, O., Topalli, D.: Performing and analyzing non-formal inspections of entity relationship diagram (erd). Journal of Systems and Software **86**(8), 2184–2195 (2013)
16. Cao, A., Chintamani, K.K., Pandya, A.K., Ellis, R.D.: NASA TLX: Software for assessing subjective mental workload. Behavior research methods **41**(1), 113–117 (2009)
17. Carlson, N.R.: Physiology of Behavior, 12 edn. Pearson (2019)
18. Charness, G., Gneezy, U.: Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization **83**(1), 50–58 (2012)
19. Cohen, J.: A power primer. Psychological bulletin **112**(1), 155 (1992)
20. Crosby, M.E., Stelovsky, J.: How do we read algorithms? a case study. Computer **23**(1), 25–35 (1990)
21. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 language guide (2016). URL `https://arxiv.org/abs/1605.07767v3`
22. Dishman, R.K., Nakamura, Y., Garcia, M.E., Thompson, R.W., Dunn, A.L., Blair, S.N.: Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. International Journal of Psychophysiology **37**(2), 121–133 (2000)
23. Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G.: Individual risk attitudes: Measurement, determinants, and behavioral consequences. Journal of the European Economic Association **9**(3), 522–550 (2011)
24. Duchowski, A.: Eye tracking methodology: Theory and practice, vol. 373. Springer Science & Business Media (2007)
25. Durndell, A., Haag, Z.: 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 (2002)
26. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous system activity distinguishes among emotions. Science **221**(4616), 1208–1210 (1983). DOI 10.1126/science.6612338
27. Fisher, A., Margolis, J.: Unlocking the clubhouse: women in computing. In: S. Grissom, D. Knox, D.T. Joyce, W. Dann (eds.) Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education, 2003, p. 23. ACM (2003)
28. Fisher, M., Cox, A., Zhao, L.: Using sex differences to link spatial cognition and program comprehension. In: 2006 22nd IEEE International Conference on Software Maintenance, pp. 289–298. IEEE (2006)

29. Fritz, T., Begel, A., Müller, S.C., Yigit-Elliott, S., Züger, M.: Using psycho-physiological measures to assess task difficulty in software development. In: Proceedings of the 36th International Conference on Software Engineering, pp. 402–413. ACM (2014)
30. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: testing software for discrimination. In: Proceedings of the 11th Joint Meeting on Foundations of Software Engineering, pp. 498–510. ACM (2017)
31. Goldberg, J.H., Kotval, X.P.: Computer interface evaluation using eye movements: methods and constructs. International Journal of Industrial Ergonomics **24**(6), 631–645 (1999)
32. Gralha, C.: iStarLab Tool (2019). URL `http://microlina.github.io/Framework/tools/iStarLab2.0/`. (Last access: May, 2020)
33. Gralha, C.: Quality evaluation of requirements models: The case of goal models and scenarios. Ph.D. thesis, Universidade Nova de Lisboa, Portugal (2019)
34. Gralha, C., Goulão, M., Araújo, J.: Analysing gender differences in building social goal models: a quasi-experiment. In: Proceedings of the IEEE 27th International Requirements Engineering Conference (RE 2019), pp. 165–176. IEEE (2019)
35. Gralha, C., Goulão, M., Araújo, J.: Are there gender differences when interacting with social goal models? Supplemental Material (2019). URL `http://doi.org/10.5281/zenodo.3819208`. (Last access: May, 2020)
36. Grigoreanu, V., Burnett, M., Wiedenbeck, S., Cao, J., Rector, K., Kwan, I.: End-user debugging strategies: A sensemaking perspective. ACM Transactions on Computer-Human Interaction **19**(1), 5 (2012)
37. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First steps towards an automatic system. In: Proceedings of the Tutorial and Research Workshop on Affective Dialogue System (ASD 2004), pp. 36–48. Springer (2004). DOI 10.1007/978-3-540-24842-2_4
38. Hancock, P.A., Chignell, M.H.: Toward a theory of mental workload: Stress and adaptability in human-machine systems. IEEE Transactions on Systems, Man and Cybernetics pp. 378–383 (1986)
39. Handy, T.C.: Event-related Potentials: A Methods Handbook. The MIT press (2005)
40. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 904–908. SAGE Publications (2006). DOI 10.1177/154193120605000909
41. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. Advances in Psychology **52**, 139–183 (1988). DOI 10.1016/S0166-4115(08)62386-9
42. Hartzel, K.: How self-efficacy and gender issues affect software adoption and use. Communications of the ACM **46**, 167–171 (2003)
43. Hou, W., Kaur, M., Komlodi, A., Lutters, W.G., Boot, L., Cotten, S.R., Morrell, C., Ozok, A.A., Tufekci, Z.: "girls don't waste time": pre-adolescent attitudes toward ICT. In: G.M. Olson, R. Jeffries (eds.) Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, pp. 875–880. ACM (2006)
44. Huffman, A.H., Whetten, J., Huffman, W.H.: Using technology in higher education: The influence of gender roles on technology self-efficacy. Computers in Human Behavior **29**(4), 1779–1786 (2013)
45. Ikutani, Y., Uwano, H.: Brain activity measurement during program comprehension with nirs. In: Proceedings of the 15th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2014), pp. 1–6. IEEE (2014). DOI 10.1109/SNPD.2014.6888727
46. Jernigan, W., Horvath, A., Lee, M., Burnett, M., Cuilty, T., Kuttal, S., Peters, A., Kwan, I., Bahmani, F., Ko, A.: A principled evaluation for a principled idea garden. In: IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 235–243. IEEE (2015)
47. Kitchenham, B., Madeyski, L., Brereton, P.: Problems with statistical practice in human-centric software engineering experiments. In: Proceedings of the Evaluation and Assessment on Software Engineering, pp. 134–143 (2019)
48. Kitchenham, B., Madeyski, L., Budgen, D., Keung, J., Brereton, P., Charters, S., Gibbs, S., Pohthong, A.: Robust statistical methods for empirical software engineering. Empirical Software Engineering **22**(2), 579–630 (2017)
49. Kramer, A.F.: Physiological metrics of mental workload: A review of recent progress. In: D.L. Damos (ed.) Multiple-Task Performance, 1 edn., pp. 279–328. Taylor & Francis (1991)

50. Li, M., Lu, B.L.: Emotion classification based on gamma-band eeg. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1223–1226. IEEE (2009). DOI 10.1109/IEMBS.2009.5334139
51. Luque-Casado, A., Perales, J.C., Cárdenas, D., Sanabria, D.: Heart rate variability and cognitive processing: The autonomic response to task demands. Biological psychology **113**, 83–90 (2016)
52. Martini, F.H., Bartholomew, E.F.: Essentials of Anatomy and Physiology, 7 edn. Pearson (2016)
53. Meyers-Levy, J., Loken, B.: Revisiting gender differences: What we know and what lies ahead. Journal of Consumer Psychology **25**(1), 129–149 (2015)
54. Meyers-Levy, J., Maheswaran, D.: Exploring differences in males' and females' processing strategies. Journal of consumer research **18**(1), 63–70 (1991)
55. Müller, S.C., Fritz, T.: Stuck and frustrated or in flow and happy: sensing developers' emotions and progress. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, vol. 1, pp. 688–699. IEEE (2015)
56. Murugappan, M., Nagarajan, R., Yaacob, S.: Modified energy based time-frequency features for classifying human emotions using eeg. In: International Conference on Man-Machine Systems, pp. 1–5 (2009)
57. NeuroSky MindWave EEG headset: (2019). URL `http://neurosky.com/biosensors/eeg-sensor/biosensors/`. (Last access: May, 2020)
58. Nourbakhsh, N., Wang, Y., Chen, F., Calvo, R.A.: Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: Proceedings of the 24th Australian Computer-Human Interaction Conference, pp. 420–423. ACM (2012)
59. O'Donnell, E., Johnson, E.: Gender effects on processing effort during analytical procedures. Int. J. Auditing **5**, 91–105 (2001)
60. Paas, F.G.W.C., van Merriënboer, J.J.G.: The efficiency of instructional conditions: An approach to combine mental effort and performance measures. Human Factors: The Journal of the Human Factors and Ergonomics Society **35**(4), 737–743 (1993). DOI 10.1177/001872089303500412
61. Paas, F.G.W.C., van Merriënboer, J.J.G.: Instructional control of cognitive load in the training of complex cognitive tasks. Educational Psychology Review **6**(4), 351–371 (1994). DOI 10.1007/BF02213420
62. Pajares, F., Miller, M.D.: Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. Journal of educational psychology **86**(2), 193 (1994)
63. Pereira, R.: Avaliação da qualidade de user stories. Master's thesis, Universidade Nova de Lisboa, Portugal (2020)
64. Petrusel, R., Mendling, J.: Eye-tracking the factors of process model comprehension tasks. In: Proceedings of the 25th International Conference on Advanced Information Systems Engineering, pp. 224–239 (2013). DOI 10.1007/978-3-642-38709-8_15
65. Pimentel, J., Castro, J.: pistar tool – a pluggable online tool for goal modeling. In: Proceedings of the IEEE International Requirements Engineering Conference (RE 2018), pp. 498–499. IEEE (2018). DOI 10.1109/RE.2018.00071
66. Poole, A., Ball, L.J.: Eye tracking in HCI and usability research. Encyclopedia of human computer interaction **1**, 211–219 (2006)
67. Porras, G.C., Guéhéneuc, Y.G.: An empirical study on the efficiency of different design pattern representations in uml class diagrams. Empirical Software Engineering **15**(5), 493–522 (2010)
68. Radach, R., Hyona, J., Deubel, H.: The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, 1 edn. Elsevier (2003)
69. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin **124**(3), 372–422 (1998). DOI 10.1037/0033-2909.124.3.372
70. Rosner, D., Bean, J.: 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, pp. 419–422 (2009)
71. Runeson, P., Host, M., Rainer, A., Regnell, B.: Case study research in software engineering: Guidelines and examples. Wiley (2012)
72. Santos, M., Gralha, C., Goulão, M., Araujo, J., Moreira, A.: On the impact of semantic transparency on understanding and reviewing social goal models. In: Proceedings of the IEEE 26th International Requirements Engineering Conference (RE 2018), pp. 228–239. IEEE (2018)

73. Santos, M., Gralha, C., Goulão, M., Araújo, J., Moreira, A., Cambeiro, J.: What is the impact of bad layout in the understandability of social goal models? In: Proceedings of the IEEE 24th International Requirements Engineering Conference (RE 2016), pp. 206–215. IEEE (2016)
74. Sharafi, Z., Marchetto, A., Susi, A., Antoniol, G., Guéhéneuc, Y.G.: An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. In: Proceedings of the 21st International Conference on Program Comprehension, pp. 33–42. IEEE (2013)
75. Sharafi, Z., Shaffer, T., Sharif, B., et al.: Eye-tracking metrics in software engineering. In: 2015 Asia-Pacific Software Engineering Conference (APSEC), pp. 96–103. IEEE (2015)
76. Sharafi, Z., Soh, Z., Guéhéneuc, Y.G.: A systematic literature review on the usage of eye-tracking in software engineering. Information and Software Technology **67**, 79–107 (2015)
77. Sharafi, Z., Soh, Z., Guéhéneuc, Y.G., Antoniol, G.: Women and men – different but equal: On the impact of identifier style on source code reading. In: 20th IEEE International Conference on Program Comprehension (ICPC), pp. 27–36. IEEE (2012)
78. Sharif, B.: Empirical assessment of uml class diagram layouts based on architectural importance. In: Proceeding of the 27th International Conference on Software Maintenance, pp. 544–549. IEEE (2011). DOI 10.1109/ICSM.2011.6080828
79. Sharif, B., Maletic, J.: An eye tracking study on the effects of layout in understanding the role of design patterns. In: Proceedings of the 26th IEEE International Conference on Software Maintenance, pp. 1–10. IEEE (2010). DOI 10.1109/ICSM.2010.5609582
80. Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F.: Galvanic skin response (gsr) as an index of cognitive load. In: CHI'07 extended abstracts on Human factors in computing systems, pp. 2651–2656. ACM (2007)
81. Showkat, D., Grimm, C.: Identifying gender differences in information processing style, self-efficacy, and tinkering for robot tele-operation. In: Proceedings of the 15th International Conference on Ubiquitous Robots, pp. 443–448. IEEE (2018)
82. Siegmund, J., Kästner, C., Apel, S., Parnin, C., Bethmann, A., Leich, T., Saake, G., Brechmann, A.: Understanding understanding source code with functional magnetic resonance imaging. In: Proceedings of the 36th International Conference on Software Engineering (CAiSE 2014), pp. 378–389. ACM (2014). DOI 10.1145/2568225.2568252
83. Simon, S.J.: The impact of culture and gender on web sites: an empirical study. DATA BASE **32**(1), 18–37 (2001)
84. Sloan, R.P., Shapiro, P.A., Bagiella, E., Boni, S.M., Paik, M., Bigger Jr, J.T., Steinman, R.C., Gorman, J.M.: Effect of mental stress throughout the day on cardiac autonomic control. Biological psychology **37**(2), 89–99 (1994)
85. de Smet, B., Lempereur, L., Sharafi, Z., Guéhéneuc, Y.G., Antoniol, G., Habra, N.: Taupe: Visualizing and analyzing eye-tracking data. Science of Computer Programming **79**, 260–278 (2014)
86. Smith, M.E., Gevins, A.: Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. In: Biomonitoring for Physiological and Cognitive Performance during Military Operations, vol. 5797, pp. 116–127. International Society for Optics and Photonics (2005)
87. Störrle, H., Baltsen, N., Christoffersen, H., Maier, A.: On the impact of diagram layout: How are models actually read? In: International Conference on Model Driven Engineering Languages and Systems (MoDELS), pp. 31–35 (2014)
88. Szafir, D., Mutlu, B.: Pay attention!: designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 11–20. ACM (2012)
89. Tan, D.S., Czerwinski, M., Robertson, G.: Women go with the (optical) flow. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 209–215. ACM (2003)
90. Tatum, W.O.: Handbook of EEG Interpretation, 2 edn. Demos Medical Publishing (2014)
91. The Eye Tribe eye-tracker: (2019). URL `https://theeyetribe.com/`. (Last access: May, 2020)
92. TLX@NASA: Nasa tlx paper/pensil version. URL `https://humansystems.arc.nasa.gov/groups/TLX/tlxpaperpencil.php`. (Last access: May 2020)
93. Torkzadeh, G., Koufteros, X.: Factorial validity of a computer self-efficacy scale and the impact of computer training. Educational and psychological measurement **54**(3), 813–821 (1994)

94. Vegas, S., Apa, C., Juristo, N.: Crossover designs in software engineering experiments: Benefits and perils. IEEE Transactions on Software Engineering **42**(2), 120–135 (2016)
95. Vorvoreanu, M., Zhang, L., Huang, Y., Hilderbrand, C., Steine-Hanson, Z., Burnett, M.: From gender biases to gender-inclusive design: An empirical investigation. In: ACM SIGCHI (2019)
96. Weber, E.U., Blais, A.R., Betz, N.E.: A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. Journal of behavioral decision making **15**(4), 263–290 (2002)
97. Welford, A.T.: Mental workload as a function of demand, capacity, strategy and skill. Ergonomics **21**(3), 151–167 (1978). DOI 10.1080/00140137808931710
98. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering, 2 edn. Springer, London, UK (2012)
99. Yeh, Y.Y., Wickens, C.D.: Dissociation of performance and subjective measures of workload. Human Factors: The Journal of the Human Factors and Ergonomics Society **30**(1), 111–120 (1988). DOI 10.1177/001872088803000110
100. Yu, E.: Modelling strategic relationships for process reengineering. Ph.D. thesis, University of Toronto, Canada (1995)
101. Yu, E.: Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of ISRE'97: 3rd IEEE International Symposium on Requirements Engineering, pp. 226–235. IEEE (1997)
102. Yusuf, S., Kagdi, H., Maletic, J., et al.: Assessing the comprehension of uml class diagrams via eye tracking. In: Proceeding of the 15th International Conference on Program Comprehension, pp. 113–122. IEEE (2007)