Analysing gender differences in building social goal models: a quasi-experiment

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

Abstract*—*Context: **Recent research has shown gender differences in problem-solving, and gender biases in how software supports it. GenderMag has 5 problem-solving facets related to gender-inclusiveness: motivation for using the 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 building social goal models is largely unexplored.** ***Objectives:*** **We evaluated the impact of different levels of GenderMag facets on creating and modifying iStar 2.0 models.** ***Methods:*** **We performed a quasi-experiment. We characterised 100 participants according to each GenderMag facet. Participants performed creation and modification 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 perceived performance and speed, their accuracy was higher. We also observed some statistically significant differences in visual effort, mental effort, and stress.** ***Conclusions:*** **Participants with a comprehensive information processing style and a more conservative attitude towards risk (characteristics more frequently seen in women) solved the tasks with a lower speed but higher accuracy.**

***Index Terms*—social goal models, iStar 2.0, biometrics, gender**

I. Introduction

Research into gender differences has determined that individual characteristics in how people solve problems often cluster by gender [1], [2]. In software systems, it is common to have features that are inadvertently designed to be more supportive of problem-solving processes typically followed by males than by females [3], [4]. Awareness of these gender biases within software systems has increased in recent years [5], [6], 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, by taking this new knowledge into account. In fact, designing software systems to be more gender-inclusive can benefit all problem solvers, regardless of their gender [7], [8].

In order to help software practitioners evaluate their software system from a gender-inclusiveness perspective, GenderMag (**Gender** Inclusiveness **Mag**nifier) [9] was created. It has 5 problem-solving facets related to gender-inclusiveness, that have been extensively investigated in the literature: motivation for using the 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. GenderMag proposes personas to bring those facets to life. Although GenderMag has been used in HCI and design (e.g., [10], [11]), the role its facets may play when building social goal models is largely unexplored.

In this paper, our goal is to analyse the impact of differences in the levels of each GenderMag facet, when stakeholders perform creation and modification tasks on iStar 2.0 models [12], an evolution of *i** [13], a goal-driven modelling language used to model software requirements. We characterised 100 participants, 50 for each task, according the 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, the duration of those tasks, 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). Our results support the evidence that participants with a comprehensive information processing style and a more conservative attitude towards risk (characteristics more frequently seen in women) analyse the entire problem more thoroughly before starting the proposed task. The visual effort, attention and mental workload was also higher for these participants.

II. Background

GenderMag [9] 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 5 problem-solving facets related with gender-inclusiveness, which are the ones repeatedly implicated by research from other fields, such as psychology, education and communications: 1) motivation for using the software, 2) information processing style, 3) computer self-efficacy, 4) attitude towards risk, and 5) 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 women, and Tim's are more frequently seen in men. The Pats' (identical) facet values emphasise that differences relevant to inclusiveness lie in the facets themselves, and not in gender identity. Table I summarises the facet values for each persona.

TABLE I: Summary of the facet values for each persona.

	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 [11] to characterise stakeholders and determine their 5 facets. We then explore how variations in the facets influence iStar 2.0 models' creation and modification.

A. 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 [14]. Some studies investigated the impact of self-efficacy on Math problem-solving success [15], as well as on strategies followed by males and females to solve problems [16], [17]. Fisher et al. [18] conducted a study to compare male and female subjects' performance on program comprehension tasks. More recently, Sharafi et al. [2] 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. Yet, this does not affect the task's duration.

Biometric sensors have been used in Software Engineering. For instance, Crosby et al. [19] used the 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 [20]. Yusuf et al. [21] 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. [22], [23] studied the effect of different layouts for design pattern roles identification in UML class diagrams. Other studies with eye-tracking focused on BPMN [24], ER [25], TROPOS [26] and *i** [27], [28].

Ikutani et al. [53] used near-infrared spectroscopy to investigate the difference in brain activity for various types of program comprehension tasks. Siegmund et al. [105] 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. [30] classify the difficulty of code or models comprehension, respectively, by using a combination of eye-tracking and EEG. Müller et al. [31] 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 or modifying requirements models, in particular, iStar 2.0 models.

III. Experiment planning

A. Goals

We describe our twofold research goals following the GQM research goals template [32]. 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 organisation(s). Our second 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 **modification** of iStar 2.0 models, *from the viewpoint* of researchers, *in the context of* an experiment conducted at our organisation(s).

We can break down each goal into three sub-goals (G1.1, G1.2, G1.3, G2.1, G2.2 and G2.3), concerning the effect(s) of the different facets, in terms of *speed*, *accuracy* and *ease*. The refined goals can be obtained by replacing the terms *creation* (or *modification*) with *speed to create*, *accuracy to create*, and *ease to create* (or *speed to modify*, *accuracy to modify*, and *ease to modify*). These refined goals are further divided to be applicable for each one of the five GenderMag facets.

B. Tasks

Each participant had to complete one task. However, there were two tasks available: *creation* and *modification* of iStar 2.0 models. No participant was tested for both tasks, as a learning effect from one evaluation to the next could represent a confounding effect. 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. The distribution of the tasks to the participants was random, but we balanced the number of participants performing each task. No (bio)feedback was provided to the participant, to avoid an unnecessary validity threat.

C. Participants

This evaluation was performed by 100 participants selected by convenience sampling, where 50 performed the creation task, and 50 performed the modification task. We leveraged personal contacts and participants were made aware of the study either by direct communication or by e-mail.

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) [33]. 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 [34].

Concerning participants *age* distribution, the ones performing the creation task had between 20 and 38 years old, with an average of 25 year old. The ones performing the modification task had between 20 and 40 years old, with an average of 26 year old. With respect to *gender*, for the creation task, there were 15 female participants and 35 males. In the modification task, we had 18 females and 32 males.

For *highest completed level of education*, all participants had some university level training. For those performing the creation task, 10 completed high school, 16 had a BSc, 24 a MSc degree. For those participating in the modification task, 6 completed high school, 16 had a BSc, and 28 a MSc degree. Regarding *current occupation*, the creation task was performed by 15 practitioners, 17 working students, and 18 students. The modification task was performed by 14 practitioners, 14 working students, and 22 students. Concerning the *field of studies*, for the creation task, we had 31 computer scientists, 4 medical doctors, 9 environment engineers, 4 lawyers, and 2 mechanical engineers. For the modification task, we had 33 computer scientists, 3 medical doctors, 8 environment engineers, and 6 lawyers. For both tasks, all the practitioners or working students had over 4 years of experience.

As to *previous experience* with *i** or iStar 2.0, 3 performers of the creation task learnt *i** in the context of a course, and worked with it for four months two years ago. For the modification task, one participant learnt *i** in the context of a course, and worked with it for six months, three years ago.

Finally, in terms of the *usage of reading devices*, 18 participants performing the creation task wore eyeglasses, and 10 had contact lenses. In the modification task, 20 participants wore eyeglasses, and 13 had contact lenses.

Participants spanned a reasonably wide range of values of each of the five GenderMag facets. Only two participants were a "pure" Abby, and four were a "pure" Tim. The others had mixes of Abby and Tim facets. A complete characterisation of participants is available in the paper's companion site [35].

D. Experimental materials

A *participant consent form*, adapted from [36], explained that the participation was entirely voluntary, the participants could refuse to answer any question and leave at any time, and that all the collected data would remain anonymous.

A *video tutorial* has 3 minutes and 58 seconds, and explains the elements of an iStar 2.0 model. It includes the construction of a correct model, similar to those that will be created or modified in the experiment, 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 [12].

Both *tasks* share a common structure, with three Areas Of Interest (AOI): the *problem description* on the left side, the *editor's toolbar* on top, and the *canvas* where participants would create or modify the models. All the elements presented to participants were comfortably readable in the 22 inch monitor used to conduct the experiment.

A *NASA-TLX questionnaire* collects feedback on the participants' perceptions with respect to effort on the performed task. It uses six dimensions: mental, physical, and temporal demand, performance, effort, and frustration.

A *GenderMag questionnaire* has a set of 9-point Likert questions. There are 20 questions, divided into questions related with each one of the facets. The scores for each facet are added, and each individual is compared to the grand median for that facet. If a participant is above the 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, Pats are not present in the facets [11].

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

E. Hypotheses, parameters, and variables

For each one of the high level goals, 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 influence iStar 2.0 models *creation*.
> $H_{1Create}$: Differences in the levels of each facet influence iStar 2.0 models *creation*.

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 influence iStar 2.0 models *creation accuracy*.
> $H_{1CreateAcc}$: Differences in the levels of each facet influence iStar 2.0 models *creation accuracy*.

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

The independent variables are the levels (*Abby*, *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* and *ease*.

Assessing accuracy. The accuracy achieved by our participants is assessed using the following metrics:

- *Precision* – the fraction of model elements created or modified which are relevant.
- *Recall* – the fraction of relevant model elements created or modified by participants, over the total number of elements elements created or modified.
- *F-measure* – measure that combines precision and recall, computed as $\frac{2*(Precision*Recall)}{(Precision+Recall)}$.

Higher values of *precision*, *recall* and *f-measure*, support the claim of a better accuracy.

Assessing speed. The speed achieved by our participants is assessed by using the following metrics:

- *Duration* – the time taken to complete the task.
- *FirstAct* – First Action; the time taken to accurately add the *first* element to the model. If a participant does not correctly create at least one element, this metric will be removed from all further analysis procedure.
- *LastAct* – Last Action; the time taken to accurately add the *last* element to the model. This is dual for FirstAct.
- *ProcDur* – Processing Duration; difference between Duration and LastAct.

Lower values of these metrics indicate that the corresponding facet level may help in improving the speed with which the models are created and modified. While the overall *duration* addresses the time spent in the task, *FirstAct* and *LastAct* provide a detailed picture of the moment when the participant starts and ends providing valid feedback. A higher value for *ProcDur* indicates that the participant stopped working on the model, but decided to revise it before finishing the task.

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.

- *FixRel* – Fixation Rate on Relevant elements; the fraction of number of fixations in a given AOI over the total number of fixations on the AOG (Area Of Glance). A fixation is a stabilisation of the eye on a part of the stimulus for a period of time between 200 and 300 ms.
- *FixIrrel* – Fixation Rate on Irrelevant elements; the fraction of number of fixations in an given AOI over the total number of fixations in the AOG.
- *AvDurRelFix* – Average Duration of Relevant Fixation; the fraction of total duration of fixations for relevant AOIs over the number of elements of the relevant AOIs.
- *AvDurIrrelFix* – Average Duration of Irrelevant Fixation; the fraction of total duration of fixations for irrelevant AOIs over the number of elements of the irrelevant AOIs.
- *TotSac* – total number of saccades. A saccade is a sudden and quick eye-movement lasting between 40 to 50 ms.
- *AvAttention* – average attention time.
- *AvMentWL* – average mental workload time.
- *AvFam* – average familiarity (memory access) time.
- *AvgSCL* – average skin conductive level (tonic signal).
- *HRVar* – Hear Rate Variability; variation in the time interval between two consecutive heart beats. We used RMSSD (root mean square of successive differences), and NN50 (the number of pairs of successive beat-to-beat intervals that differ more than 50ms).
- *NASA-TLX score* – overall weighted score the from NASA-TLX questionnaire.

For the *eye-tracker*, 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) [20], [37], [38]. Regarding the *average duration of fixation*, a higher value indicates more time and attention devoted to AOIs [20], [25], [39], which is correlated with cognitive processes [40], [41]. 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 [20], [29], [41], [42]. Regarding the *EEG scanner*, the values for *attention*, *mental workload* and *familiarity*, are calculated based on specific frequency bands, often referred to as alpha, beta, gamma, delta and theta. A decrease of alpha and often an increase in theta EEG activity indicates an increase in attention demand and working memory load [43], [44]. A higher *attention* indicates the participant is engaged in the task, and a higher *mental workload* indicates effort while performing it. For *familiarity*, a higher value is associated with memory accessing and lower effort while performing the task. With respect to the *EDA scanner*, a higher *skin conductive level* have been linked to a greater cognitive load, task difficulty, and stress [45], [46]. An increase in the *heart rate*, when in a stationary state, can be related with anxiety [47], [48] and mental stress [49]. Concerning the *NASA-TLX score*, higher scores are associated with a higher perceived effort by the participant [50]. For all the metrics, lower complexity will correspond to higher *ease* in performing the tasks.

F. Design

This evaluation follows a quasi-experimental design, since the allocation of participants to the creation or the modification task was random, but without a pre-selection process. If a participant performed the creation task, the next participant would be allocated to the modification task, so that the number of participants performing each task would be balanced. In terms of tasks distribution, we have a between subjects design. This means that every participant will only perform one of the tasks, not both. However, our independent variables are the levels on each of the five GenderMag facets. Since we evaluate the differences in the levels of each facet for each participant, we have a within subjects design.

G. Procedure

We prepared the lab setting so that all participants had similar conditions. The lab was only being used for the evaluations, and there was only one participant in each evaluation session. We informed the participant that the task consisted in watching a tutorial on a requirements language, and creating or modifying a model based on a problem description. We further informed him that we would be recording the contents on the screen, tracking his eyes movement, and collecting information of his mental effort and heart rate. Finally, we explained he could quit at any moment and that there was no time limit.

After reading the consent form, the participant put the EDA wristband on the wrist of the dominant hand, after removing any watches or bracelets. The buckle of the wristband was adjusted by the participant to a comfortable position.

Before putting the EEG headset, participants with earrings were asked to remove them. A 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 bangs, so that nothing was obstructing the forehead sensor of the EEG headset.

We helped the participant seating comfortably so that his 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) to proceed to the actual data collection.

We asked participants to watch a 2 minutes video of fish swimming while wearing the biometric sensors. It allow us to record a baseline during the second minute, used to normalise the captured biometric data [29], [31].

After that, participants watched a video tutorial on iStar 2.0 and then started the task. When the participant felt the task was completed, s/he answers the NASA-TLX. Finally, each participant answered a short questionnaire about demographic information, and completed the GenderMag questionnaire.

H. Analysis procedure

We start by collecting descriptive statistics on our variables, namely the mean, standard deviation, skewness and kurtosis, to get an overview of their distribution. This was complemented with kernel density plots to help with the visual analysis of those distributions. Kernel density plots provide a more detailed picture of a distribution, when compared to box plots, and are a better fit for comparing distributions in Software Engineering experimentation. This visual analysis was then complemented with Welch *t*-tests. A discussion on the benefits of using kernel density plots vs. box plots, and using Welch *t*-test for comparing distributions in a robust way is in [51].

IV. Execution

Preparation. We carried out data collection with a laptop connected to an external 22 inch, wide screen, full HD monitor; an The Eye Tribe eye-tracker [52]; a NeuroSky MindWave EEG headset [53]; a BioSignalsPlux Wristband [54] with BITalino [55] 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. We scheduled the sessions according to participant's availability, with at least one hour between evaluations.

Deviations. 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. Although the time that the collection was not made was only 11 seconds, we decided to still exclude the EEG data for that participant. As such, for the mental effort, the total number of participants for the modification task is 49.

V. Analysis

A. Descriptive statistics

For the sake of brevity, we only present in Table II the results concerning *accuracy*, which include *precision*, *recall* and *F-measure*. 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 while the other 10 refer to the creation 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.

B. Data set preparation

We collected the times when the participant started and ended the tasks. For both tasks, we had a target model and some variations that were accepted. The model creation tool collected all the elements added by the participant in an CSV file. We compared the target model(s) with the solution modelled by the participant. Concerning the eye-data, the areas of the stimulus and its elements were mapped into pixel coordinates to determine which regions and elements the participants were looking at. This allowed tagging the eye-tracking data with elements being gazed at any given time, a necessary step for computing the eye-tracking metrics. Concerning EEG and EDA, both devices have a tool with pre-built algorithms, which help in processing the raw data.

C. Hypotheses testing

We used the 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 [51]. We are using $p < 0.05$ for the level of significance and thus rejecting the null hypothesis.

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

Table III summarises the Welch *t*-test results for the creation task, for the *motivation* facet. For all the variables, we found no statistical evidence of differences between participants identified as Abby and the ones identified as Tim.

Table IV summarises the Welch *t*-test results for the creation task, for the *information processing* facet. There was a statistically significant difference in variables concerning *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 ($M = .351$, $SD = .199$,

TABLE II: Descriptive statistics

	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	Mod.	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
Recall	Creation	Mot.	Abby	.536	.243	-.397	.204	.260
			Tim	.657	.235	-.498	-.908	.020
		Inf. P.	Abby	.580	.218	-.213	-.846	.035
			Tim	.7218	.301	-1.587	2.373	.016
		S. E.	Abby	.568	.213	-.010	-.870	.034
			Tim	.687	.279	-1.259	1.051	.012
		Risk	Abby	.440	.230	.116	.643	.731
			Tim	.678	.216	-.653	-.293	.008
		Learn.	Abby	.5383	.271	-.424	-.105	.730
			Tim	.634	.233	-.385	-.764	.057
	Modification	Mod.	Abby	.631	.264	-.355	-.593	.256
			Tim	.684	.310	-.966	.131	.001
		Inf. P.	Abby	.683	.257	-.610	-.236	.007
			Tim	.595	.399	-.584	-1.263	.033
		S. E.	Abby	.673	.261	-.528	-.221	.014
			Tim	.648	.348	-.866	-.510	.008
		Risk	Abby	.652	.283	-.439	-.536	.423
			Tim	.669	.299	-.858	-.010	.001
		Learn.	Abby	.684	.288	-.710	-.155	.303
			Tim	.657	.297	-.766	-.146	.003
F-Measure	Creation	Mot.	Abby	.454	.171	-.907	1.332	.230
			Tim	.505	.127	-.065	.761	.426
		Inf. P.	Abby	.500	.136	-.154	-.023	.685
			Tim	.435	.173	-1.624	3.768	.030
		S. E.	Abby	.494	.137	.012	.066	.903
			Tim	.479	.162	-1.466	3.126	.035
		Risk	Abby	.486	.198	-.920	2.030	.306
			Tim	.485	.123	-.320	-.427	.555
		Learn.	Abby	.483	.203	-.950	2.520	.267
			Tim	.486	.126	-.334	-.546	.402
	Modification	Mod.	Abby	.566	.166	-.769	1.106	.399
			Tim	.581	.247	-.930	1.368	.006
		Inf. P.	Abby	.613	.185	-.539	2.378	.088
			Tim	.443	.278	-.699	-.814	.053
		S. E.	Abby	.609	.197	-.478	2.084	.165
			Tim	.516	.244	-1.185	.597	.010
		Risk	Abby	.665	.237	-.535	.637	.618
			Tim	.541	.203	-1.473	2.279	.000
		Learn.	Abby	.610	.214	-.555	1.784	.716
			Tim	.564	.582	-.996	1.769	.005

$t(1) = 7.208$, $p = .016$). 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 ($M = 907.655$, $SD = 325.390$, $t(1) = 15.065$, $p = .002$). The *average duration of irrelevant fixations* was higher for Abby ($M = 879.526$, $SD = 346.001$) than for Tim ($M = 468.981$, $SD = 378.193$, $t(1) = 10.487$,

TABLE III: Welch *t*-test: *creation* task, *motivation* facet

Metric	Statistic	df1	df2	Sig.
Precision	.509	1	38.526	.480
Recall	3.023	1	37.131	.090
F-Measure	1.229	1	30.117	.276
Duration	1.751	1	28.616	.196
FirstAct	.390	1	42.009	.536
LastAct	2.321	1	24.742	.140
ProcDur	.733	1	31.040	.398
FixRel	1.462	1	39.399	.234
FixIrrel	1.293	1	42.429	.262
AvDurRelFix	.444	1	28.088	.511
AvDurIrrelFix	.995	1	34.582	.325
TotSac	.207	1	38.757	.652
AvAttention	2.962	1	42.334	.093
AvMentWL	.861	1	34.936	.360
AvFam	.126	1	38.957	.725
AvSCL	2.078	1	43.689	.157
HRVarRMSSD	1.256	1	37.181	.270
HRVarNN50	.947	1	34.672	.337
NASA-TLX	.417	1	47.854	.521

$p = .005$). 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 ($M = 36.810$, $SD = 15.996$, $t(1) = 5.126$, $p = .035$). The *perceived effort* was higher for Abby ($M = 75.855$, $SD = 18.931$) than for Tim ($M = 51.152$, $SD = 19.545$, $t(1) = 13.895$, $p = .002$).

TABLE IV: Welch *t*-test: *creation* task, *information proc.* facet

Metric	Statistic	df1	df2	Sig.
Precision	7.208	1	16.328	**.016**
Recall	2.121	1	13.103	.169
F-Measure	1.335	1	13.710	.268
Duration	2.278	1	13.004	.155
FirstAct	.274	1	15.344	.608
LastAct	8.708	1	13.355	**.011**
ProcDur	39.830	1	46.822	**.000**
FixRel	.511	1	15.744	.485
FixIrrel	27.178	1	43.530	**.000**
AvDurRelFix	15.065	1	12.270	**.002**
AvDurIrrelFix	10.487	1	15.058	**.005**
TotSac	.137	1	17.912	.715
AvAttention	7.007	1	13.434	**.020**
AvMentWL	4.262	1	15.078	.057
AvFam	.230	1	13.129	.639
AvSCL	7.145	1	22.210	**.014**
HRVarRMSSD	.050	1	16.432	.827
HRVarNN50	5.126	1	20.142	**.035**
NASA-TLX	13.895	1	15.703	**.002**

Table V summarises the Welch *t*-test results for the creation task, for the *self-efficacy* facet. There was a statistically significant difference in several of the variables, concerning *speed* and *ease*. The *duration* of participants identified as Abby in the self-efficacy facet was lower ($M = 1596.063$, $SD = 551,620$) than the one of participants identified as Tim ($M = 2136,611$, $SD = 683,933$, $t(1) = 8.232$,

$p = .008$). The time for performing the *last action* was lower for Abby ($M = 1067,531$, $SD = 423,516$) than for Tim ($M = 2076,500$, $SD = 674,687$, $t(1) = 32.952$, $p = .000$). The *processing duration* was higher for Abby ($M = 528.531$, $SD = 271.804$) than for Tim ($M = 60.111$, $SD = 81.166$, $t(1) = 82.036$, $p = .000$). The *number of irrelevant fixations* was higher for Abby ($M = 6.187$, $SD = 4.601$) than for Tim ($M = 2.787$, $SD = 2.799$, $t(1) = 10.540$, $p = .002$). The *average duration of relevant fixations* was lower for Abby ($M = 463.663$, $SD = 173.218$) than for Tim ($M = 827.932$, $SD = 301.122$, $t(1) = 22.207$, $p = .000$). The *average mental workload* was higher for Abby ($M = .756$, $SD = .179$) than Tim ($M = .578$, $SD = .180$, $t(1) = 11.344$, $p = .002$). The *perceived effort* was higher for Abby ($M = 78.208$, $SD = 18.264$) than for Tim ($M = 56.574$, $SD = 20.207$, $t(1) = 14.137$, $p = .001$).

TABLE V: Welch *t*-test: *creation* task, *self-efficacy* facet

Metric	Statistic	df1	df2	Sig.
Precision	2.348	1	34.147	.135
Recall	2.453	1	28.232	.128
F-Measure	.289	1	30.784	.595
Duration	8.232	1	29.548	**.008**
FirstAct	1.001	1	38.747	.323
LastAct	32.952	1	24.706	**.000**
ProcDur	82.036	1	39.785	**.000**
FixRel	.353	1	35.953	.556
FixIrrel	10.540	1	47.621	**.002**
AvDurRelFix	22.207	1	23.472	**.000**
AvDurIrrelFix	2.359	1	26.210	.137
TotSac	.147	1	37.149	.703
AvAttention	3.619	1	27.486	.068
AvMentWL	11.344	1	35.274	**.002**
AvFam	.516	1	30.307	.478
AvSCL	.598	1	40.586	.444
HRVarRMSSD	.083	1	34.168	.775
HRVarNN50	2.018	1	40.182	.163
NASA-TLX	14.137	1	32.456	**.001**

Table VI summarises the Welch *t*-test results for the creation task, for the *risk* facet. There was a statistically significant difference in variables concerning *accuracy*, *speed* and *ease*. The *precision* achieved by participants identified as Abby in the risk facet ($M = .680$, $SD = .270$) was higher than the one of participants identified as Tim ($M = .422$, $SD = .134$, $t(1) = 11.698$, $p = .004$). The *recall* was lower for Abby ($M = .440$, $SD = .230$) than for Tim ($M = .678$, $SD = .216$, $t(1) = 11.109$, $p = .003$). The time for performing the *first action* was higher for Abby ($M = 343.429$, $SD = 253.290$) than for Tim ($M = 189.333$, $SD = 133.427$, $t(1) = 4.677$, $p = .046$). The *perceived effort* was higher for Abby ($M = 82.571$, $SD = 16.578$) than for Tim ($M = 65.694$, $SD = 21.534$, $t(1) = 8.761$, $p = .006$).

Table VII summarises the Welch *t*-test results for the creation task, for the *learning style* facet. There was a statistically significant difference in variables concerning *speed* and *ease*. The *duration* of participants identified as Abby in the learning style facet was higher ($M = 2471.333$, $SD = 716.492$) than the one of participants identified as Tim ($M = 1575.711$, $SD = 458.800$, $t(1) = 16.601$, $p = .001$). The time for performing the *first action* was higher for Abby ($M = 487.250$, $SD = 172.717$) than for Tim ($M = 152.026$, $SD = 96.349$, $t(1) = 41.160$, $p = .000$). The time for performing the *last action* was higher for Abby ($M = 2001.000$, $SD = 878.572$) than for Tim ($M = 1250.684$, $SD = 555.105$, $t(1) = 7.772$, $p = .015$). The *number of irrelevant fixations* was lower for Abby ($M = 3.020$, $SD = 2.840$) than for Tim ($M = 5.577$, $SD = 4.574$, $t(1) = 5.346$, $p = .028$). The *perceived effort* was higher for Abby ($M = 87.750$, $SD = 15.762$) than for Tim ($M = 64.947$, $SD = 20.268$, $t(1) = 16.500$, $p = .000$).

TABLE VI: Welch *t*-test: *creation* task, *risk* facet

Metric	Statistic	df1	df2	Sig.
Precision	11.698	1	15.567	**.004**
Recall	11.109	1	22.418	**.003**
F-Measure	.000	1	17.024	.984
Duration	3.118	1	16.319	.096
FirstAct	4.677	1	15.888	**.046**
LastAct	2.994	1	16.559	.102
ProcDur	.088	1	24.745	.769
FixRel	.301	1	23.117	.588
FixIrrel	.577	1	19.105	.457
AvDurRelFix	1.598	1	17.938	.222
AvDurIrrelFix	.415	1	29.412	.524
TotSac	1.110	1	27.051	.302
AvAttention	.841	1	27.685	.367
AvMentWL	.014	1	18.523	.906
AvFam	1.973	1	23.199	.173
AvSCL	1.179	1	29.974	.286
HRVarRMSSD	.034	1	18.159	.855
HRVarNN50	2.526	1	27.259	.124
NASA-TLX	8.761	1	30.741	**.006**

TABLE VII: Welch *t*-test: *creation* task, *learning style* facet

Metric	Statistic	df1	df2	Sig.
Precision	.171	1	15.180	.685
Recall	1.219	1	16.444	.286
F-Measure	.004	1	13.758	.950
Duration	16.601	1	13.964	**.001**
FirstAct	41.160	1	13.230	**.000**
LastAct	7.772	1	13.883	**.015**
ProcDur	1.645	1	16.351	.218
FixRel	.055	1	16.770	.817
FixIrrel	5.346	1	30.348	**.028**
AvDurRelFix	2.335	1	13.805	.149
AvDurIrrelFix	1.588	1	14.675	.227
TotSac	.180	1	18.103	.676
AvAttention	.481	1	20.951	.495
AvMentWL	.581	1	16.073	.457
AvFam	1.655	1	14.873	.218
AvSCL	.043	1	22.084	.837
HRVarRMSSD	.001	1	13.570	.973
HRVarNN50	.001	1	17.681	.979
NASA-TLX	16.500	1	23.576	**.000**

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

Table VIII summarises the Welch *t*-test results for the modification task, for the *motivation* facet. For all the vari-

ables, we found no statistical evidence of differences between participants identified as Abby and the ones identified as Tim.

TABLE VIII: Welch *t*-test: *modification* task, *motivation* facet

Metric	Statistic	df1	df2	Sig.
Precision	.046	1	47.99	.830
Recall	.419	1	42.94	.521
F-Measure	.066	1	47.521	.798
Duration	2.384	1	29.925	.133
FirstAct	.099	1	36.548	.755
LastAct	3.269	1	23.764	.083
ProcDur	.913	1	38.223	.345
FixRel	.330	1	38.315	.569
FixIrrel	.305	1	42.911	.583
AvDurRelFix	.002	1	37.104	.963
AvDurIrrelFix	1.114	1	38.005	.298
TotSac	.232	1	39.986	.632
AvAttention	.709	1	40.371	.405
AvMentWL	.009	1	37.221	.925
AvFam	.917	1	35.388	.345
AvSCL	.050	1	35.649	.825
HRVarRMSSD	.496	1	39.747	.485
HRVarNN50	.424	1	40.045	.519
NASA-TLX	.114	1	46.987	.737

Table IX summarises the Welch *t*-test results for the modification task, for the *information processing* facet. There was a statistically significant difference in variables concerning *accuracy*, *speed* and *ease*. The *precision* achieved by participants identified as Abby in the information processing facet was higher ($M = .626$, $SD = .217$) than the one achieved by participants identified as Tim ($M = .379$, $SD = .228$, $t(1) = 10.600$, $p = .005$). The time for performing the *last action* was lower for Abby ($M = 772.744$, $SD = 473.777$) than for Tim ($M = 1569.273$, $SD = 864.023$, $t(1) = 8.618$, $p = .013$). The *processing duration* was higher for Abby ($M = 415.436$, $SD = 246.111$) than for Tim ($M = 75.636$, $SD = 119.354$, $t(1) = 40.540$, $p = .000$). The *number of irrelevant fixations* was higher for Abby ($M = 5.110$, $SD = 4.350$) than for Tim ($M = .253$, $SD = .356$, $t(1) = 47.490$, $p = .000$). The *average duration of relevant fixations* was lower for Abby ($M = 379.207$, $SD = 273.123$) than for Tim ($M = 835.262$, $SD = 97.556$, $t(1) = 12.775$, $p = .003$). The *average duration of irrelevant fixations* was higher for Abby ($M = 704.546$, $SD = 454.026$) than for Tim ($M = 309.635$, $SD = 341.011$, $t(1) = 9.835$, $p = .005$). The *average attention* was higher for Abby ($M = .721$, $SD = .163$) than for Tim ($M = .473$, $SD = .179$, $t(1) = 17.042$, $p = .001$).

Table X summarises the Welch *t*-test results for the modification task, for the *self-efficacy* facet. There was a statistically significant difference in variables concerning *speed* and *ease*. The *duration* of participants identified as Abby in the self-efficacy facet was lower ($M = 1086.688$, $SD = 571.192$) then the one of participants identified as Tim ($M = 1647.722$, $SD = 683.284$, $t(1) = 8.711$, $p = .006$). The time for performing the *last action* was lower for Abby ($M = 672.563$, $SD = 406.856$) than for Tim ($M = 1437.611$, $SD = 750.079$, $t(1) = 16.067$, $p = .001$). The *processing duration* was higher for Abby ($M = 414.125$, $SD = 254.964$) than for Tim ($M = 210.111$, $SD = 235.163$, $t(1) = 8.155$, $p = .007$). The *number of irrelevant fixations* was higher for Abby ($M = 1.611$, $SD = .590$) than for Tim ($M = 1.564$, $SD = .749$, $t(1) = 13.709$, $p = .001$). The *average duration of relevant fixations* was lower for Abby ($M = 347.512$, $SD = 279.480$) than for Tim ($M = 714.247$, $SD = 362.099$, $t(1) = 13.829$, $p = .001$). The *average attention* was higher for Abby ($M = .741$, $SD = .162$) than for Tim ($M = .533$, $SD = .178$, $t(1) = 16.600$, $p = .000$). The *average mental workload* was higher for Abby ($M = .756$, $SD = .168$) than for Tim ($M = .561$, $SD = .191$, $t(1) = 13.036$, $p = .001$).

TABLE IX: Welch *t*-test: *modification* task, *info. proc.* facet

Metric	Statistic	df1	df2	Sig.
Precision	10.600	1	15.483	**.005**
Recall	.480	1	12.434	.501
F-Measure	3.654	1	12.607	.079
Duration	2.975	1	13.035	.108
FirstAct	1.114	1	12.955	.311
LastAct	8.618	1	11.746	**.013**
ProcDur	40.540	1	35.088	**.000**
FixRel	2.822	1	16.234	.112
FixIrrel	47.490	1	39.742	**.000**
AvDurRelFix	12.775	1	12.780	**.003**
AvDurIrrelFix	9.835	1	21.111	**.005**
TotSac	1.639	1	17.132	.218
AvAttention	17.042	1	14.954	**.001**
AvMentWL	5.580	1	15.570	.032
AvFam	1.913	1	16.031	.186
AvSCL	.403	1	14.111	.536
HRVarRMSSD	.003	1	19.333	.955
HRVarNN50	.186	1	18.985	.671
NASA-TLX	1.789	1	16.696	.199

TABLE X: Welch *t*-test: *modification* task, *self-efficacy* facet

Metric	Statistic	df1	df2	Sig.
Precision	3.988	1	32.935	.054
Recall	.066	1	27.942	.799
F-Measure	1.904	1	29.586	.178
Duration	8.711	1	30.414	**.006**
FirstAct	.016	1	33.088	.899
LastAct	16.067	1	22.751	**.001**
ProcDur	8.155	1	37.841	**.007**
FixRel	.053	1	28.997	.820
FixIrrel	13.709	1	47.409	**.001**
AvDurRelFix	13.829	1	28.544	**.001**
AvDurIrrelFix	1.914	1	35.835	.175
TotSac	4.030	1	42.441	.051
AvAttention	16.600	1	32.672	**.000**
AvMentWL	13.036	1	31.723	**.001**
AvFam	.051	1	32.781	.823
AvSCL	.054	1	28.751	.818
HRVarRMSSD	.450	1	39.179	.506
HRVarNN50	.090	1	37.169	.766
NASA-TLX	1.739	1	40.880	.195

Table XI summarises the Welch *t*-test results for the modification task, for the *risk* facet. There was a statistically significant difference in variables concerning *accuracy*, *speed* and *ease*. The *precision* achieved by participants identified as Abby in the risk facet was higher ($M = .766$, $SD =$

.248) than the one achieved by participants identified as Tim ($M = .495$, $SD = .193$, $t(1) = 13.523$, $p = .002$). The time for performing the *first action* was higher for Abby ($M = 268.857$, $SD = 160.808$) than for Tim ($M = 127.750$, $SD = 91.590$, $t(1) = 9.572$, $p = .007$). The *heart rate variability* (for RMSSD) was lower for Abby ($M = 32.307$, $SD = 15.493$) than for Tim ($M = 48.667$, $SD = 21.288$, $t(1) = 9.001$, $p = .005$).

TABLE XI: Welch *t*-test: *modification* task, *risk* facet

Metric	Statistic	df1	df2	Sig.
Precision	13.523	1	19.433	**.002**
Recall	.033	1	25.036	.858
F-Measure	3.025	1	20.854	.097
Duration	3.213	1	16.014	.092
FirstAct	9.572	1	16.390	**.007**
LastAct	3.691	1	15.181	.074
ProcDur	.306	1	21.597	.586
FixRel	.001	1	22.126	.974
FixIrrel	.108	1	19.359	.746
AvDurRelFix	2.251	1	16.632	.152
AvDurIrrelFix	.119	1	28.408	.733
TotSac	.004	1	25.342	.948
AvAttention	.069	1	20.806	.795
AvMentWL	.112	1	26.203	.740
AvFam	.900	1	25.734	.352
AvSCL	.011	1	21.705	.916
HRVarRMSSD	9.001	1	32.575	**.005**
HRVarNN50	3.728	1	29.102	.063
NASA-TLX	3.963	1	17.214	.063

Table XII summarises the Welch *t*-test results for the modification task, for the *learning style* facet. There was a statistically significant difference in variables concerning *accuracy* and *speed*. The *duration* of the task performed by participants identified as Abby in the learning style facet ($M = 1988.000$, $SD = 758.318$) was higher than the one of participants identified as Tim ($M = 1067.816$, $SD = 454.088$, $t(1) = 15.872$, $p = .001$). The time for performing the *first action* was higher for Abby ($M = 383.583$, $SD = 71.299$) than for Tim ($M = 98.947$, $SD = 31.052$, $t(1) = 180.439$, $p = .000$). The time for performing the *last action* was higher for Abby ($M = 1565.000$, $SD = 860.680$) than for Tim ($M = 753.132$, $SD = 443.540$, $t(1) = 9.851$, $p = .008$).

TABLE XII: Welch *t*-test: *modification* task, *learning* facet

Metric	Statistic	df1	df2	Sig.
Precision	.494	1	19.466	.491
Recall	.079	1	19.007	.781
F-Measure	.417	1	18.975	.526
Duration	15.872	1	13.580	**.001**
FirstAct	180.439	1	12.344	**.000**
LastAct	9.851	1	12.895	**.008**
ProcDur	.850	1	13.125	.373
FixRel	.079	1	19.808	.781
FixIrrel	3.757	1	30.720	.062
AvDurRelFix	2.204	1	13.026	.161
AvDurIrrelFix	2.227	1	15.247	.156
TotSac	.174	1	16.447	.682
AvAttention	.335	1	14.673	.571
AvMentWL	.430	1	26.202	.518
AvFam	1.479	1	20.673	.238
AvSCL	.236	1	17.234	.633
HRVarRMSSD	3.828	1	18.422	.066
HRVarNN50	3.492	1	19.698	.077
NASA-TLX	.759	1	15.389	.397

VI. Discussion

A. Evaluation of results and implications

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

We found no evidence that the *motivation* facet influences 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's 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).

Assessing speed. We found that Abbys in the *learning style* were slower than Tims. However, Abbys 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, 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 on 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 thought a 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 thought 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*), we found that there was a greater perceived effort for Abby than for Tim, which is in line with biometrics data.

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

The results found for the modification task are similar to those of the creation task (RQ1), with some exceptions:

Assessing accuracy. Although differences in terms of precision were the same as the ones in the creation task, there was no difference in terms of recall in the modification task.

Assessing speed. The *learning style* facet had no influence in the duration of the task.

Assessing ease. 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.

B. Threats to validity

Conclusion validity. Although we have a significant high number of participants, higher than most sample sizes reported, in particular, in other eye-tracking experiments (see [20]), sample size is always a risk, as results may not apply to even larger populations. We plan to extend this study by performing replicas, and we facilitate independent replications, by sharing the materials used in this work.

Internal validity. We used convenience sampling, where the actual participants tend to be more motivated to be part of the experiments, since their participation is entirely voluntary, and can bias the results. We plan to launch a replication of this experiment with participants selected through a recruitment call, and make an independent replication package available to colleagues from other organisations and countries.

External validity. Overall, our participants had little to no prior knowledge in *i** or iStar 2.0. By having participants with a greater level of experience with the language we could analyse the differences between these two profiles. Further research is needed to assess how different facet levels in experienced *i** or iStar 2.0 users would impact the results.

Construct validity. We have showed a video tutorial about iStar 2.0, and afterwards participants were asked to create or modify iStar 2.0, 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, we have not informed them about what was being tested.

C. Inferences

Information processing and risk have impact on accuracy. Abby in these facets is 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. 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. Abby in these facets has a more comprehensive analysis of the problem description and the model elements available in the editor's toolbar. 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's 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 thought 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.

Diversity is key. The complementarity of results achieved by Tim and Abby according to these facets 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.

VII. Conclusions and future work

We performed a quasi-experiment to analyse the impact of different levels in each of the five GenderMag facets, when creating or modifying iStar 2.0 models. We measured the accuracy, speed and ease of a total of 100 participants (50 for each task). 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 participants with a comprehensive information processing style and a more conservative attitude towards risk (characteristics more frequently seen in women) took longer to start performing the tasks but had a higher accuracy. The visual effort, attention and mental workload was also higher for these participants. Finally, motivation is the only facet that is not related with accuracy, speed or ease.

It is necessary to assess how consistently our results occur with other users, problem descriptions, and models. We plan to replicate the experiment in other contexts, and apply it to bigger and more complex descriptions.

Acknowledgment

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

REFERENCES

[1] L. Beckwith, C. Kissinger, M. Burnett, S. Wiedenbeck, J. Lawrance, A. Blackwell, and C. Cook, "Tinkering and gender in end-user programmers' debugging," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2006, pp. 231–240.

[2] Z. Sharafi, Z. Soh, Y.-G. Guéhéneuc, and G. Antoniol, "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)*. IEEE, 2012, pp. 27–36.

[3] V. Grigoreanu, M. Burnett, S. Wiedenbeck, J. Cao, K. Rector, and I. Kwan, "End-user debugging strategies: A sensemaking perspective," *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 1, p. 5, 2012.

[4] D. Szafir and B. Mutlu, "Pay attention!: designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM, 2012, pp. 11–20.

[5] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*. ACM, 2017, pp. 498–510.

[6] D. Showkat and C. Grimm, "Identifying gender differences in information processing style, self-efficacy, and tinkering for robot teleoperation," in *Proceedings of the 15th International Conference on Ubiquitous Robots*. IEEE, 2018, pp. 443–448.

[7] D. S. Tan, M. Czerwinski, and G. Robertson, "Women go with the (optical) flow," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2003, pp. 209–215.

[8] W. Jernigan, A. Horvath, M. Lee, M. Burnett, T. Cuilty, S. Kuttal, A. Peters, I. Kwan, F. Bahmani, and A. Ko, "A principled evaluation for a principled idea garden," in *IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 2015, pp. 235–243.

[9] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, "GenderMag: A method for evaluating software's gender inclusiveness," *Interacting with Computers*, vol. 28, no. 6, pp. 760–787, 2016.

[10] M. Burnett, R. Counts, R. Lawrence, and H. Hanson, "Gender hcl and microsoft: Highlights from a longitudinal study," in *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 2017, pp. 139–143.

[11] M. Vorvoreanu, L. Zhang, Y. Huang, C. Hilderbrand, Z. Steine-Hanson, and M. Burnett, "From gender biases to gender-inclusive design: An empirical investigation," in *ACM SIGCHI*, 2019.

[12] F. Dalpiaz, X. Franch, and J. Horkoff. (2016) iStar 2.0 language guide. [Online]. Available: https://arxiv.org/abs/1605.07767v3

[13] E. Yu, "Modelling strategic relationships for process reengineering," Ph.D. dissertation, University of Toronto, Canada, 1995.

[14] J. P. Byrnes, D. C. Miller, and W. D. Schafer, "Gender differences in risk taking: a meta-analysis." *Psychological bulletin*, vol. 125, no. 3, p. 367, 1999.

[15] F. Pajares and M. D. Miller, "Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis." *Journal of educational psychology*, vol. 86, no. 2, p. 193, 1994.

[16] L. Beckwith, M. Burnett, S. Wiedenbeck, C. Cook, S. Sorte, and M. Hastings, "Effectiveness of end-user debugging software features: Are there gender issues?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2005, pp. 869–878.

[17] G. Torkzadeh and X. Koufteros, "Factorial validity of a computer self-efficacy scale and the impact of computer training," *Educational and psychological measurement*, vol. 54, no. 3, pp. 813–821, 1994.

[18] M. Fisher, A. Cox, and L. Zhao, "Using sex differences to link spatial cognition and program comprehension," in *2006 22nd IEEE International Conference on Software Maintenance*. IEEE, 2006, pp. 289–298.

[19] M. E. Crosby and J. Stelovsky, "How do we read algorithms? a case study," *Computer*, vol. 23, no. 1, pp. 25–35, 1990.

[20] Z. Sharafi, Z. Soh, and Y.-G. Guéhéneuc, "A systematic literature review on the usage of eye-tracking in software engineering," *Information and Software Technology*, vol. 67, pp. 79–107, 2015.

[21] S. Yusuf, H. Kagdi, J. Maletic *et al.*, "Assessing the comprehension of uml class diagrams via eye tracking," in *Proceeding of the 15th International Conference on Program Comprehension*. IEEE, 2007, pp. 113–122.

[22] B. Sharif and J. Maletic, "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*. IEEE, 2010, pp. 1–10.

[23] B. Sharif, "Empirical assessment of uml class diagram layouts based on architectural importance," in *Proceeding of the 27th International Conference on Software Maintenance*. IEEE, 2011, pp. 544–549.

[24] R. Petrusel and J. Mendling, "Eye-tracking the factors of process model comprehension tasks," in *Proceedings of the 25th International Conference on Advanced Information Systems Engineering*, 2013, pp. 224–239.

[25] N. E. Cagiltay, G. Tokdemir, O. Kilic, and D. Topalli, "Performing and analyzing non-formal inspections of entity relationship diagram (erd)," *Journal of Systems and Software*, vol. 86, no. 8, pp. 2184–2195, 2013.

[26] Z. Sharafi, A. Marchetto, A. Susi, G. Antoniol, and Y.-G. Guéhéneuc, "An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension," in *Proceedings of the 21st International Conference on Program Comprehension*. IEEE, 2013, pp. 33–42.

[27] M. Santos, C. Gralha, M. Goulao, J. Araújo, A. Moreira, and J. Cambeiro, "What is the impact of bad layout in the understandability of social goal models?" in *2016 IEEE 24th International Requirements Engineering Conference (RE)*. IEEE, 2016, pp. 206–215.

[28] M. Santos, C. Gralha, M. Goulão, J. Araujo, and A. Moreira, "On the impact of semantic transparency on understanding and reviewing social goal models," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*. IEEE, 2018, pp. 228–239.

[29] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014, pp. 402–413.

[30] H. Störrle, N. Baltsen, H. Christoffersen, and A. Maier, "On the impact of diagram layout: How are models actually read?" in *International Conference on Model Driven Engineering Languages and Systems (MoDELS)*, 2014, pp. 31–35.

[31] S. C. Müller and T. Fritz, "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. IEEE, 2015, pp. 688–699.

[32] V. R. Basili and H. D. Rombach, "The TAME project: Towards improvement-oriented software environments," *IEEE Trans. Software Eng.*, vol. 14, no. 6, pp. 758–773, 1988.

[33] B. Kitchenham, L. Madeyski, and P. Brereton, "Problems with statistical practice in human-centric software engineering experiments," in *Proceedings of the Evaluation and Assessment on Software Engineering*, 2019, pp. 134–143.

[34] J. Cohen, "A power primer," *Psychological bulletin*, vol. 112, no. 1, p. 155, 1992.

[35] Gender differences in building social goal models. (2019) Supplemental material. (access: April, 2019). [Online]. Available: https://sites.google.com/view/gendermag-istar/

[36] P. Runeson, M. Host, A. Rainer, and B. Regnell, *Case study research in software engineering: Guidelines and examples*. Wiley, 2012.

[37] A. Poole and L. J. Ball, "Eye tracking in HCI and usability research," *Encyclopedia of human computer interaction*, vol. 1, pp. 211–219, 2006.

[38] Z. Sharafi, T. Shaffer, B. Sharif *et al.*, "Eye-tracking metrics in software engineering," in *2015 Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2015, pp. 96–103.

[39] G. C. Porras and Y.-G. Guéhéneuc, "An empirical study on the efficiency of different design pattern representations in uml class diagrams," *Empirical Software Engineering*, vol. 15, no. 5, pp. 493–522, 2010.

[40] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer Science & Business Media, 2007, vol. 373.

[41] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, 1999.

[42] B. de Smet, L. Lempereur, Z. Sharafi, Y.-G. Guéhéneuc, G. Antoniol, and N. Habra, "Taupe: Visualizing and analyzing eye-tracking data," *Science of Computer Programming*, vol. 79, pp. 260–278, 2014.

[43] M. E. Smith and A. Gevins, "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. International Society for Optics and Photonics, 2005, pp. 116–127.

[44] M. Murugappan, R. Nagarajan, and S. Yaacob, "Modified energy based time-frequency features for classifying human emotions using eeg," in *International Conference on Man-Machine Systems*, 2009, pp. 1–5.

[45] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (gsr) as an index of cognitive load," in *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 2007, pp. 2651–2656.

[46] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*. ACM, 2012, pp. 420–423.

[47] R. K. Dishman, Y. Nakamura, M. E. Garcia, R. W. Thompson, A. L. Dunn, and S. N. Blair, "Heart rate variability, trait anxiety, and perceived stress among physically fit men and women," *International Journal of Psychophysiology*, vol. 37, no. 2, pp. 121–133, 2000.

[48] A. Luque-Casado, J. C. Perales, D. Cárdenas, and D. Sanabria, "Heart rate variability and cognitive processing: The autonomic response to task demands," *Biological psychology*, vol. 113, pp. 83–90, 2016.

[49] R. P. Sloan, P. A. Shapiro, E. Bagiella, S. M. Boni, M. Paik, J. T. Bigger Jr, R. C. Steinman, and J. M. Gorman, "Effect of mental stress throughout the day on cardiac autonomic control," *Biological psychology*, vol. 37, no. 2, pp. 89–99, 1994.

[50] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis, "NASA TLX: Software for assessing subjective mental workload," *Behavior research methods*, vol. 41, no. 1, pp. 113–117, 2009.

[51] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust statistical methods for empirical software engineering," *Empirical Software Engineering*, vol. 22, no. 2, pp. 579–630, 2017.

[52] The Eye Tribe eye-tracker. (2019) (access: April, 2019). [Online]. Available: https://theeyetribe.com/

[53] NeuroSky MindWave EEG headset. (2019) (access: April, 2019). [Online]. Available: http://neurosky.com/biosensors/eeg-sensor/biosensors/

[54] BioSignalsPlux Wristband. (2019) (access: April, 2019). [Online]. Available: https://biosignalsplux.com/

[55] BITalino. (2019) (access: April, 2019). [Online]. Available: http://bitalino.com/