Are Requirements Elicitation Sessions Influenced by Participants' Gender? An Empirical Experiment

Eduardo Díaz[1], José Ignacio Panach[1], Silvia Rueda[1], Marcela Ruiz[2], Oscar Pator[3]

[1] Escola Tècnica Superior d'Enginyeria, Departament d'Informàtica, Universitat de València, Avenida de la Universidad s/n, 46100 Burjassot, València, España

[2] Zurich University of Applied Sciences, ZHAW School of Engineering InIT, Technikumstrasse 9, 8401 Winterthur, Switzerland

[3] Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain

[1]diazsua@alumni.uv.es, [1]joigpana@uv.es, [1]silvia.rueda@uv.es, [2]ruiz@zhaw.ch, [3]opastor@pros.upv.es

Abstract

Context: Requirements elicitation is a crucial phase in the software development life cycle. During requirements elicitation sessions, requirements engineers capture software requirements, and motivate stakeholders to express needs and expected software functionalities. In this context, there is a lack of extensive empirical research reporting the extent to which elicitation sessions can be influenced by participants' gender. **Objective:** This paper presents our research endeavour to investigate requirements engineers' *effort* and elicited requirements' *accuracy* based on participants' gender. **Method:** We conducted an experiment in two rounds with a total of 59 students who played the role of requirements engineers. In the first experimental task, the participant watched two videos where men and women stakeholders expressed software requirements. Later on, the participants specified software requirements in the shape of Business Process Model and Notation (BPMN) and next they generated Graphical User Interfaces (GUIs) from those models. **Results:** We observed two significant differences. One between men and women requirements engineers in terms of dedicated effort during requirements specification: men took less effort. Other between stakeholders' gender in terms of accuracy resulted of BPMN models: models built from men stakeholders yield more accuracy. On the contrary, accuracy of resulted GUIs models did not show significant differences regarding requirements engineers or stakeholders' gender. **Conclusions:** Analysing descriptive data, women spent more time both as stakeholders and as requirements engineers but their accuracy is better.

Keywords: Requirements Elicitation, BPMN model, Graphical User Interfaces, Empirical Experiment, Gender.

1. Introduction

Requirements elicitation is an important and critical activity within the software development life cycle [1]. A defect in the requirements elicitation may involve a system failure during the implementation stage. Requirements elicitation consists of learning, extracting, or discovering needs from users or other potential stakeholders [2]. Currently, requirements engineers specify requirements in models, as for instance in the works of Mkpojiogu et al. [3], Zareen et al. [4] and Embely et al [5]. Some of these models are: UML use cases [6]; UML sequence diagrams [7]; user stories [8]; Business Process Model Notation (BPMN) [9] (which captures requirements of business process to provide a common language for describing process behavior [10] [11]). Once requirements are elicited, requirements engineers design graphical user interfaces (GUIs) compliant with such requirements [12]. In practice, elicited requirements' accuracy is measured in both: requirements models and designed GUIs.

Research results have shown that requirements engineers and stakeholders' gender influence the software development life cycle. Examples of these works are Trauth et al. [13], who described two recommendations about the role of the gender in information systems research. These recommendations are: 1) Analyze possible differences between genders driving analysis between women and men, e.g. women behaved this way, men behaved that way. 2) Gender research is not limited to gender differences research,

sometimes an important finding in gender research is that no significant gender differences were found. The work of Ridley et al. [14] shows an under-representation of women in the Australian Information Technology workforce. Stoislescu et al. [15], conclude that women students are more interested in the use of computers rather than programming, whereas men students prefer programming. Marques et al. [16], state that women have characteristics that make mixed gender teams more effective and coordinated. Ioannis et al. [17] present a study of gender differences in Computer Science education. Results show that men students preferred courses related to hardware and software engineering, whereas women students selected courses related to theoretical Computers Science, humanities and social sciences. Beyer et al. [18] conducted a study to analyze why women are underrepresented in Computer Science. Results show gender differences in computer self-efficacy; men students had high interest in the Computer Science course. All these previous studies provide evidence regarding women and men differences that may influence over the software development life cycle. Nevertheless, there is insufficient knowledge that reflect the role of gender in specific phases of the software development life cycle.

Our main research goal focuses on investigating the influence of gender in the requirements elicitation process both from the point of view of the stakeholder (who describes the requirements) and the requirements engineer (who elicits and reports the requirements).

This paper presents our research work that investigates requirements engineers' effort and elicited requirements' accuracy considering the possible differences between stakeholders and requirements engineers' genders in the requirements elicitation process. This is helpful to know how to combine genders in a requirements elicitation team to optimize the results. The comparison focuses on assessing both the gender of the speaker that describes the requirements (the stakeholder) and the gender of the people that elicit requirements (requirements engineer). The experiment consists of two problems where participants play the role of requirements engineers. The gender of the stakeholder in each problem varies, we work with two stakeholders, one man and one woman. Both stakeholders are researchers that describe the requirements of each problem through a video. Both stakeholders are experimenters experts at model-driven development and software engineering. As problem domain we chose two well-known contexts among participants: sending an academic work, and an online purchase. For each problem, experimental participants have to watch a video and elicit the requirements described in it. Note that the use of videos to elicit requirements is not as in the real life, where requirements are elicited through interviews. However, the use of videos ensures that all participants have exactly the same information for the elicitation process. Elicited requirements are described in BPMN models. Finally, participants have to draw a GUI supported with the BPMN model and the requirements. We analyze the whole requirements elicitation process through three response variables: requirements engineers' effort, BPMN model accuracy, and GUI accuracy. From existing models to specify requirements, we chose BPMN models since it allows to focus on the processes and is widely adopted in industrial settings [19].

Our experiment is conducted in two rounds of 28 participants (22 men and 6 women) and 31 participants (26 men and 5 women) respectively. All the participants that play the role of requirements engineers are undergraduate computer science students of course 2018/2019 and 2019/2020 of University of Valencia (Spain) (one course per round), and results can be summarized in next points:

- Regarding the variable effort, there are significant differences with a small effect between requirements engineers' gender. Men requirements engineers spent less effort to elicit requirements. Even though there are not significant differences between stakeholders' gender, men stakeholders involved less effort.
- Regarding the variable BPMN model accuracy, there are not significant differences between requirements engineers' gender. However, in the descriptive data we identify that women requirements engineers have more accuracy in BPMN models than men requirements engineers. There is a significant difference regarding stakeholders' gender.

Requirements engineers have more accuracy in BPMN models during elicitation sessions with men stakeholders.

- Regarding the variable GUIs accuracy, there are not significant differences between requirements engineers' gender nor stakeholders' gender. Analyzing descriptive data, we can see that women requirements engineers got slight better accuracy drawing GUIs, while a stakeholder-man yields better results for GUIs accuracy.

The paper is structured as follows. Section 2 discusses related experiments in the area of requirements elicitation. Section 3 describes the requirements elicitation method used in our experiment. Section 4 introduces the design of the experiment. Section 5 discusses threats to validity of the results. Section 6 shows the statistical results. Section 7 discusses the interpretation of the results. Finally, Section 8 presents the conclusions and future works.

2. Related Work

In this section, we review related work of requirements elicitation by running a Targeted Literature Review (TLR), a non-systematic, in-depth and informative literature review aimed at keeping only the significant references maximizing rigorousness while minimizing selection bias. The search string used is: ("elicit" OR "requirement") AND "gender" AND ("experiment" OR "BPMN" OR "user interface" OR "software"). Inclusion criteria are: (1) comparison of genders in requirements elicitation; (2) experiments that compare genders in requirements elicitation; (3) analysis of genders in BPMN modelling. Exclusion criteria are: (1) papers with no analysis of genders; (2) a topic not related with requirements elicitation. The first search returns 100 papers. After applying inclusion and exclusion criteria, we reduce the sample to 9. Next we describe accepted papers in the search.

In first place, we have the work of Fernandez-Sanz et al. [20], that applies a simple evaluation experience named Teamwork Benefits Awareness (TBA) to a group of students. TBA allows to measure individual and team performance during a requirements analysis session based on real projects. The percentage of women participants was low in general (21.2%) in all degrees used in the evaluation. Results concluded not significant differences between genders. Another work is Huang et al. [21], whose experiment was conducted to investigate the user needs through mental patterns. This evaluation focuses on studying gender differences with 36 undergraduate students (18 men and 18 women). Results show that men prefer a more animated visual design, such as flash animation, and tend to read more text-based information than women. The women pay attention to the home page and Web page's visual design more than men did. Gonzales et al. [22] present four studies to evaluate the effectiveness eliciting user requirements in different circumstances. Results show gender differences with Appreciative Inquiry (AI) [23]. Another work is Sagnier et al. [24], who propose a study for the effect of gender in virtual reality (VR). They conducted an experiment with 52 students, asking them to perform an aeronautical assembly task in VR. Results show that men students tend to give a better evaluation of their performance in VR. Men felt it was easier to perform the task than women students. The work of Schina et al. [25] analyses participants' perceptions (45 men and 39 women) regarding their learning with First Lego League (FLL). The experiment was based on a questionnaire that shows that women tend to be more engaged, enthusiastic, creative, eager to experiment and more likely to adopt collaborative strategies rather than men. Another work is Nguyen-Duc et al. [26], who conduct an investigation with men and women students in terms of software engineering activities and team dynamics in a software project course that involves a real customer. Results show that women students are more active with project management and requirement engineering than men students. Another work is Tiwari et al. [27], who show an exploratory pilot study about Case-Based Learning (CBL) methodology (which aims to facilitate the learning of several Requirements Engineering (RE) concepts). This experiment was conducted by 112 postgraduate students (76 men and 36 women). Results conclude that the difference in gender has no effect on the CBL outcome and it helps both men and women students equally in achieving various learning objectives.

Another work is Sobieraj et al. [28], that presents an experiment of 148 participants to examine the interaction between technological complexity and user' gender. Also, there are works that study the influence of stakeholders' gender, such as the work of Othmani et al. [29]. This work applied the Perceptual Evaluation of Speech Quality (PESQ) [30] in an experiment carried out to encode 350 speech files. Results show that there is a significant difference in perceptual quality score between female and male speech signals.

As conclusion of the related work, we highlight that there are differences between men and women regarding how they perceive the system to develop. Men are more focused on animated images, textual information while women are more focused on the design and are more creative (according to [21], [24], [25], [26], [27]). This paper aims to study whether these differences also appear in the first stages of the software requirements elicitation process. As notation to represent requirements in the experiment, we have opted for BPMN, since there are many previous works that have also analysed the requirements elicitation process with such models.

3. Applying BPMN and GUIs for Requirements Elicitation

This section shows how BPMN and GUIs are applied in requirements elicitation processes.

3.1. BPMN model

Requirements are translated into processes according to the BPMN notation (pool, lane, task of several types (user type, service type, message type [31]), exclusive gateway, parallel gateway, among others [31]). Next, we summarize most relevant BPMN primitives and when to use them [9].

- When the requirement involves a person or area, a lane is specified.
- When the requirement involves providing data, a User Type Task is specified.
- When the requirement involves verifying data, a Service Type Task is specified.
- When the requirement involves sending an email, a Message Type Task is specified.
- When the requirement involves at least two possible paths to continue and only one must be chosen, an Exclusive Gateway is specified.
- When the requirement involves at least two possible paths to continue, a Parallel Gateway is specified.

3.2. GUIs

Once we have the BPMN model, next we design GUIs that support such model and the requirements. Even though there is not a standard to generate interfaces from BPMN models, we show some generation rules extracted from previous works [32] [33]:

- A user type task to fill in data is represented through a form.
- A service type task to verify data is represented through a report or a datagrid.
- Three or more user type tasks that are in the same lane are represented through a wizard, or a tabbed dialog box, or a group box.
- The text that appears in an Exclusive gateway is represented through a label.
- An Exclusive gateway is represented through a radio button.
- A Parallel gateway is represented through a wizard or a tabbed dialog box, or a group box.
- From the fields of the list of requirements, we can extract the fields of the form, whose widget depends on the data type: (1) a text box, for any string; (2) a list box or a combo box for any enumeration with simple choice, (3) a radio button or a check box for any boolean.

Note that how to build BPMN models from requirements and how to transform BPMN models into GUIs is beyond the scope of this paper. There is a proposal for these generations in [32] [33], and there are works with similar approaches in [19] [34] [35]. Participants participated in the experiment without knowing the rules, they acted

intuitively. This is because we do not aim to check the suitability of the rules, but the skill at requirements elicitation. Transformation rules are described to clarify alternative ways to generate GUIs from BPMN models, they were not distributed through participants. Any of these rules is used subconsciously by the participants, and all of them are suitable. If we measure the accuracy of the BPMN model and the accuracy of the GUI, we are measuring the accuracy of the elicited requirements by transitivity.

Next, we show an illustrative example of how to design BPMN and to draw GUIs with some rules of this method. We have a few requirements (only functional requirements of the system) that represent the process Invoice payment: Requirement 1: The employee has to record the invoice data introducing: Invoice date, Invoice number, Due date; and choosing an option of Supplier (between National or International). Requirement 2: The employee has to approve the invoice choosing between "Yes" or "No"). If the invoice is approved, then employee records the invoice payment date. If the invoice is not approved, then employee records comments on the cancelled invoice.

Fig. 1 shows the BPMN model; Requirement 1 involves providing data, so we represent it through a User Type Task. Requirement 2 involves at least two possible paths to continue (Yes / No) and we must only choose one. So, we use an Exclusive Gateway and a User Type Task for each condition in the BPMN model.

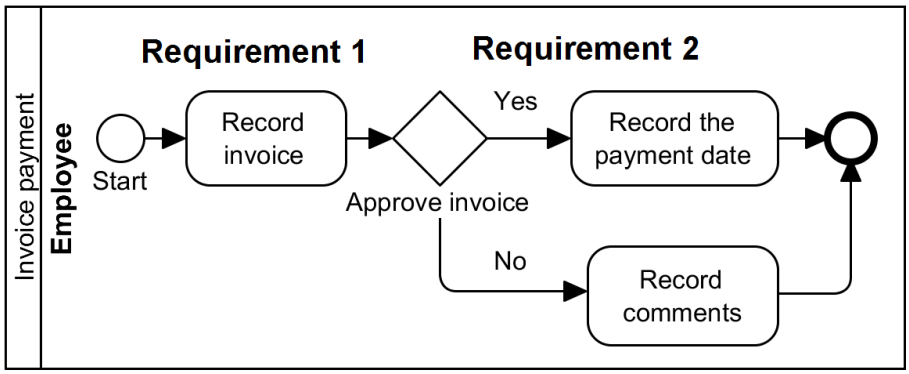

Fig. 1. BPMN model of the example.

Fig. 2 shows the GUIs generated form the BPMN model of Fig. 1. Requirement 1 is a User Type Task from which we generate a Form with the fields Invoice date, Invoice number, Due date and Supplier. Since the three first fields can receive any string, they are represented through a text box. Supplier is represented through a combo box since the user can only select some possible values. Requirement 2 is an Exclusive Decision that is represented through radio button with two options (Yes / No). Moreover, the text that appears in the gateway is represented through a label. Next, both User Type Tasks are represented through text boxes in a Form.

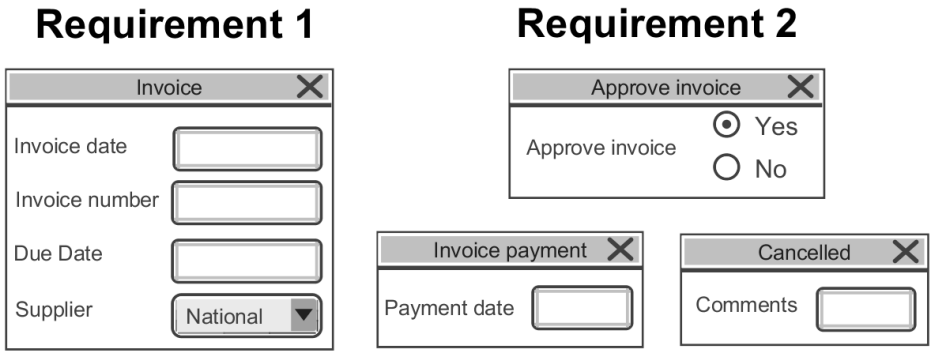

Fig. 2. GUIs of the example.

4. Experiment

This section describes the experiment conducted to empirically study the differences between men and women during the requirements elicitation process. The experiment is

conducted in two rounds with the same design, so data analysis has been done aggregating the data of both rounds. Characteristics of both rounds are the same in order to facilitate the aggregation of both rounds. The following is a description of the experiments setting according to Wohlin [36]:

4.1. Goal

The goal of this experiment is to identify differences in requirements engineers' effort and elicited requirements' considering the gender of both stakeholders and requirements engineers. The focus is placed on the possible differences in how each gender tackle the process of requirements elicitation. Of all the existing differences, we focus on differences on how a requirements model is built and how GUIs are designed. The experiment is conducted from the perspective of researchers and practitioners interested in investigating how to optimize the requirements elicitation process taking into account the possible differences between genders.

4.2. Experimental Participants

This experiment is composed of two rounds with a population of undergraduate students in computer science, that play the role of requirements engineers. First round had Twenty-eight N=28 participants (22 men vs 6 women, 16 aged in the range 17-20, 6 in 21-25, 5 in 26-30 and 1 in 31-35, M = 24.60, SD = 4.80), and second round had thirty-one N=31 participants (26 men vs 5 women, 8 aged in the range 17-20, 21 in 21-25, and 2 in 26-30, M= 22.97, SD = 2.53). All of them were recruited from two groups of University of Valencia (Spain) in Course 2018/2019 and Course 2019/2020 respectively, who have previously taken Software Engineering course, with knowledge in GUIs. We have recruited two courses to increase the number of participants to improve the statistical power. We used these two courses because these students work with requirements elicitation, so we ensure that they know how to work with BPMN models. Table 1 shows the previous knowledge (M=Men, W= Women) of BPMN models and GUIs. Most of the women requirements engineers in both courses have no knowledge or low knowledge of BPMN models, but all women requirements engineers have medium or high knowledge of GUIs. Most of men requirements engineers have medium or high knowledge of BPMN models. Moreover, most men requirements engineers have medium or high knowledge of GUIs. In order to ensure that participants had enough knowledge, they were trained before the experiment in BPMN modeling and GUI generation. All participants participated voluntarily in the experiment and we certified that they had enough knowledge of BPMN and GUI after the training through a test. The recruitment of non-professional participants has been used in other works in the requirements elicitation field, such as [37] which validate tools of requirements engineers with students. Other example is the work [38], which validates a method of software development where the roles of requirements engineers were a group of students.

Table 1 Knowledge of BPMN models and GUIs.

Knowledge	**% Participants of Course 2018/2019**				**% Participants of Course 2019/2020**			
	BPMN		**GUIs**		**BPMN**		**GUIs**	
	M	W	M	W	M	W	M	W
None	5%	50%	0	0	38%	0	0	0
Low	45%	17%	9%	0	42%	60%	19%	0
Medium	36%	33%	55%	17%	16%	40%	77%	40%
High	14%	0	36%	83%	4%	0	4%	60%
Total of participants	28				31			

4.3. Research Question, Hypotheses, Response Variables and Metrics

The experiment deals with two perspectives, the gender of the stakeholder that describes the requirements and the gender of the requirements engineer who gathers them. The experiment is based on next research questions:

RQ1: When specifying BPMN and GUI models in a requirements elicitation session, is stakeholders and requirements engineers' effort affected by their gender? Effort is defined as the number of labour units required to complete a schedule activity or work breakdown structure component, usually expressed as person-hours, person-days or person-weeks [39]. We measure effort as the time in minutes spent in the whole process, from watching the video with the requirements to drawing the GUIs. The hypothesis to test is H_{01}: The participants' effort to elicit requirements is independent of the gender of the stakeholder and the gender of the requirements engineer. The alternative hypothesis is H_{11}: The participants' effort to elicit requirements is affected by the gender of the stakeholder and the gender of the requirements engineer.

RQ2: Is BPMN model accuracy affected by the gender of the stakeholder or the gender of the requirements engineer? Accuracy is defined in ISO 9126-1 [40] as the capability of the software product to provide the right results. For each experimental problem, we have defined a list of requirements that compose the "experimenters' solution", i.e., the requirements that must be elicited by participants. This experimenters' solution of the BPMN model was elaborated by one experimenter and checked by other two experimenters to ensure that this solution involved all requirements. We measure BPMN model accuracy as the percentage of requirements expressed in the BPMN model developed by the participants that are compliant with the experimenters' solution. The hypothesis to test is H_{02}: The participants' accuracy when building BPMN models to elicit requirements is independent of the gender of the stakeholder and the gender of the requirements engineer. The alternative hypothesis is H_{12}: The participants' accuracy when building BPMN models to elicit requirements is affected by the gender of the stakeholder and the gender of the requirements engineer.

RQ3: Is GUIs accuracy affected by the gender of the requirements engineer or the gender of the stakeholder? We measure GUIs accuracy as the percentage of requirements in the experimenters' solution that are compliant with the GUIs designed by each participant. Experimenter' solution of the GUI was created and validated in the same way as the experimenters' solution of the BPMN model. The hypothesis to test is H_{03}: The participants' accuracy when designing GUIs from BPMN models is independent of the gender of the stakeholder and the gender of the requirements engineer. The alternative hypothesis is H_{13}: The participants' accuracy when designing GUIs from BPMN models is affected by the gender of the stakeholder and the gender of the requirements engineer.

Response variables are the effects studied in the experiment, in our experiment we have three response variables (RQ).

RQ1: Effort is measured as the time in minutes since watching the video with the requirements until drawing the GUIs.

RQ2: BPMN model accuracy is measured as the percentage of requirements expressed in the BPMN model developed by the participants that are compliant with the experimenters' solution (Eq. 1). For example, if we have three requirements in the experimenters' solution and only two of them are compliant with the BPMN model built by the participant, BPMN model accuracy is: 2/3 x 100% = 66,6%. NRBS (Number of requirements expressed in the BPMN developed by the participant that are compliant with experimenters' solution). NRS (Number of requirements in the experimenters' solution).

$$\% \text{ BPMN model Accuracy} = NRBS / NRS \times 100\% \quad (1)$$

RQ3: GUI accuracy is measured as the percentage of requirements identified in the GUI by the participants compliant with the experimenters' solution (Eq. 2). For example, if we have four requirements in the experimenters' solution and all of them appear in the GUIs drawn by the participants, GUIs accuracy is 4/4 x 100% = 100%. NRGS (Number requirements expressed in the GUI developed by the participants that are compliant with experimenters' solution). NRS (Number of requirements in the experimenters' solution).

$$\% \text{ GUIs Accuracy} = NRGS / NRS \times 100\% \quad (2)$$

4.4. Factors and Treatments

Factors are variables whose effect on the response variable we want to understand [41]. Our experiment studies two factors: requirements engineer gender and stakeholder gender. Each factor has two treatments: men and women. With the aim of generalizing the results as much as possible, we have worked with two problems even though identifying differences between them is not the target of our study. So these problems are considered as a block variable (we analyze this variable just in case it affects the results even though we are not interested in this variable specifically).

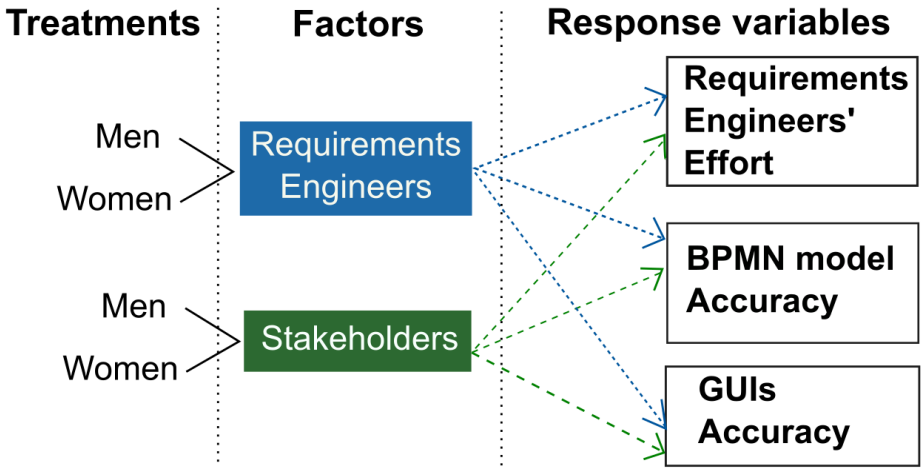

Fig. 3. Combination of factors, treatments and response variables.

Fig. 3 shows the combination of factors, treatments and response variables. Both factors (requirements engineers and stakeholders) have the same possible treatments (men and women). Both for requirements engineers and for stakeholders we analyze requirements engineers' effort, BPMN model accuracy and GUIs accuracy.

4.5. Experimental Design

This section describes the design of the experiment, which is a crossover design [42]. Stakeholder's gender and the block variable are designed as within-participants; both levels are applied to all participants. Requirements engineers' gender is designed as between-participants and only one level can be applied to each participant. We use two different problems in our design in order to avoid that results depend on a specific problem. This way, results are more generalizable than working with a single problem. Depending on the combination of treatments and the block variable (Experimental Problem), we have four different profiles for participants. Note that this combination of profiles aim to mitigate the carryover that appears in a within-subjects design. We assigned a profile to each participant randomly but balanced. Next, we describe the characteristics of each profile. Table 2 shows the characteristics of the different profiles, indicating the order of the gender of the stakeholder and the order of the applied experimental problem. One stakeholder-man and one stakeholder-woman described both experimental problems, both of them are experimenters with knowledge in software engineering, elicitation of requirements and model-driven development.

In all profiles, for each stakeholders' gender, the participants had to elicit requirements from the stakeholder through a BPMN model. Next, these models were the input for designing a GUI. Note that details of this process are described in Section 3.

Table 2 Design used in the experiment.

	Stakeholder-man	**Stakeholder-woman**
Profile 1	Experimental Problem 1	Experimental Problem 2
Profile 2	Experimental Problem 2	Experimental Problem 1
	Stakeholder-woman	**Stakeholder-man**

Profile 3	Experimental Problem 1	Experimental Problem 2
Profile 4	Experimental Problem 2	Experimental Problem 1

4.6. Experimental Problems

The experiment was conducted with two experimental problems of similar complexity, both experimental problems have the same number of tasks (4 tasks in both problems) and gateways (1 and 2 respectively), also similar number of GUIs (4 interfaces in both problems). These experimental problems were small to avoid the fatigue of the requirements engineers. Both experimental problems were described through a video with audio where a stakeholder described the requirements [43]. The stakeholders described each problem as they considered, without scripts. The only artefact that both genders share was a set of slides to support the description of the requirements. Slides contain very summarized ideas to ensure that both stakeholders talk about all requirements, but each requirement was described by each person with her/his own words. Even though both videos share the same schema, there are slight differences between them. The man's videos were shorter: 1:30 minutes in Problem 1 and 2:48 in Problem 2. The woman spent 2:32 in Problem 1 and 4:00 in Problem 2. So, each step described by the stakeholder-woman had more details than the same step described by the man. The use of the video to describe the requirements ensures that all participants have the same information to conduct the experiment, which avoids the threat of different contexts for several participants. These two problems were selected for the following reasons: participants are familiar with the context of use, familiarity exposure is comparable, easy enough to be completed in no more than one hour. These two experimental problems were used in both rounds. Next, we describe the requirements of each experimental problem.

Problem 1: Sending an academic work. This problem aims at designing a GUI where students can send an academic work through a virtual classroom. This problem has the following requirements:

Requirement 1: the student logs in to enter in the virtual classroom through user and password.

Requirement 2: the student submits an academic work. The information to fill in is: Title of the work, path of the File, and Comments.

Requirement 3: the teacher reviews the academic work, if the academic work is approved, the teacher will record the mark of the student. If the academic work results in a fail, the teacher will record the feedback and corrections for the student.

Fig. 4 shows the experimenters' solution of the BPMN model. We see that both Requirement 1 and Requirement 2 require to get data from the student, so we use two User type tasks. Requirement 3 requires at least two possible paths to continue (Approved / Failed) where only one must be chosen. So, we use an Exclusive gateway and a User type task for each condition.

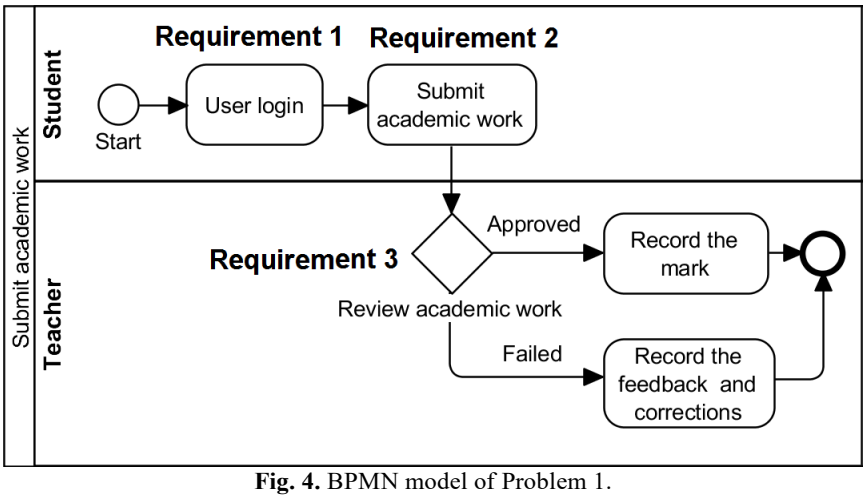

Fig. 4. BPMN model of Problem 1.

Fig. 5 shows the experimenters' solution of the GUIs of Problem 1 (derived from the BPMN model drawn in Fig. 4). Requirement 1 is represented through a User Type Task, so we use a form with the required fields: User and Password. Since they can get any string, both are represented through text boxes. Requirement 2 is represented through a User Type Task, that generates a form with the fields: Title, File and Comments. Since all these fields can get any string, they are represented through text boxes. Requirement 3 is represented through an Exclusive Decision Gateway that generates a radio button with two options Approved / Failed. Moreover, the text that appears in the gateway is represented through a label. Next, both User Type Tasks with fields Mark and Corrections are represented through two forms and two text boxes.

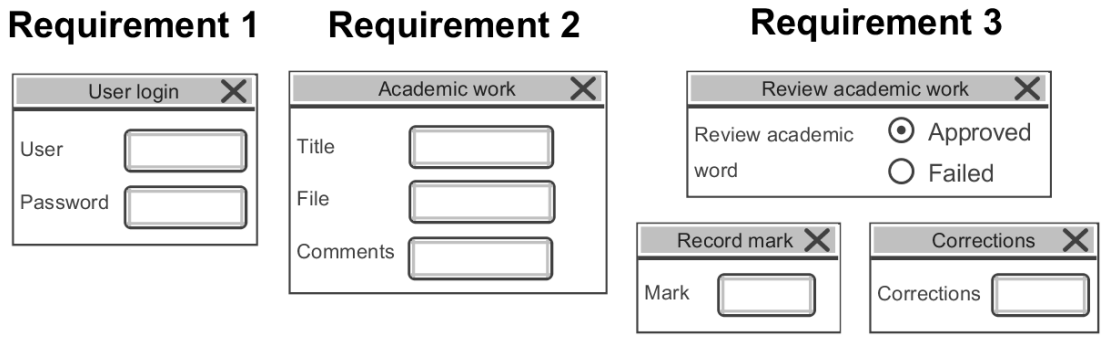

Fig. 5. GUIs of Problem 1.

Problem 2: Buying an on-line product. This problem aims at designing a system where a user can buy an on-line product. This problem has the following requirements:

Requirement 1: the user fills in the following information: Email, First Name, Last Name, Product (only one), and Quantity.

Requirement 2: the user has to perform the subsequent tasks in any order: (1) Specify the shipping address and fill in the following information: Address, City, Country, Postal Code and Mobile phone. (2) Select shipping option (among: one week, three days, and now). (3) Select payment method and fill in next data: Card type (Visa, MasterCard, and American Express), Card Number, and CVV. The amount to be paid is shown at the end of the process. Fig. 6 shows the experimenters' solution of the BPMN model. We see that Requirement 1 requires to get data, a User Type Task is specified. Requirement 2 involves at least three possible paths to continue, so we use a Parallel Gateway with three User Type Tasks (record shipping address, choose a shipping option (among one week, three days, and now), and choose a payment method (among Visa, MasterCard, and American Express)).

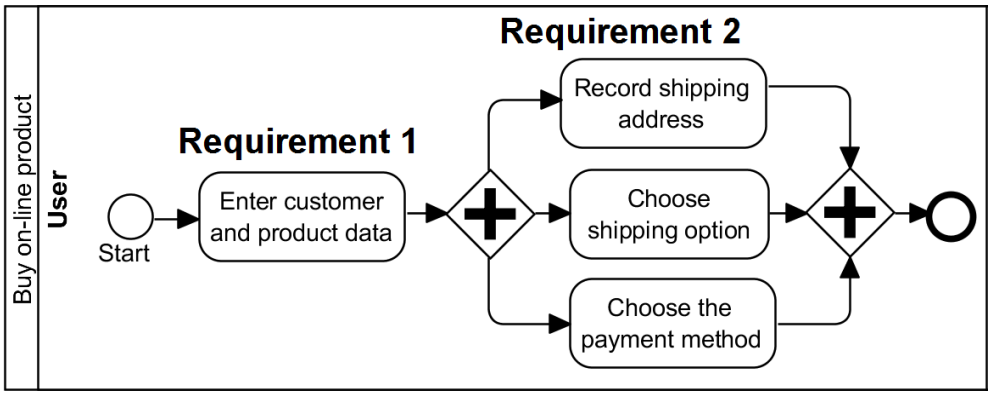

Fig. 6. BPMN model solution of Problem 2.

Fig. 7 shows the experimenters' solution of the GUIs for Problem 2. Requirement 1 is a User Type Task represented through a form with next fields: Email, First name, Last name, Quantity and Product. The first four fields can get any string, so they are

represented through text boxes. Since Product can only get some possible values, it is represented through a combo box. Requirement 2 is represented through a Parallel Gateway, which generates a wizard with three forms (one form from each User Type Task with its fields).

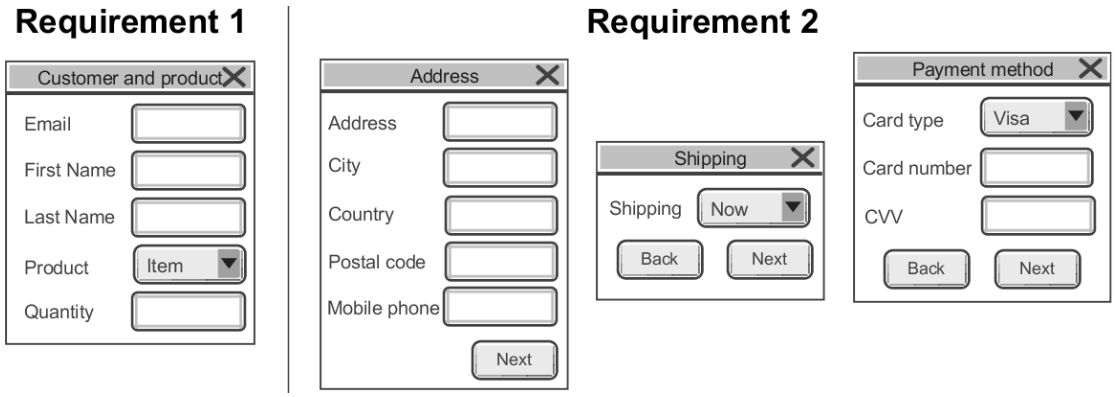

Fig. 7. GUIs solution of Problem 2.

4.7. Procedure

The procedure for this study was structured in one session of one hour:

- **Introduction to the method,** a week before the experiment, the participants had to follow a tutorial about how to build a BPMN model and how to generate GUIs from them (explained in Section 3). Participants had to train with a toy problem. The trainer was the same in both rounds of the experiment.
- **Filling a previous test**, participants were evaluated through a test before conducting the experiment. The test consisted of ten questions of BPMN models, each question had a value of one score (higher value score is ten). Each question had four alternatives and only one possible correct answer. The correct answer was computed as one point, so possible points were between 0 (no correct answers) and 10 (all answers are correct). We considered that participants that got more than five points were capable of participating in the experiment. Results of Table 3 show that all participants of the Course 2018/2019 and 2019/2020 obtained a high mark in the test (there is no participant with less than 6 points). This means that all participants had enough knowledge of BPMN models and GUIs to participate in the experiment.

Table 3 Results of the entrance test.

Scores	#Participants for Course 2018/2019		# Participants for Course 2019/2020	
	M	**W**	**M**	**W**
6	0	0	1	0
7	2	0	1	0
8	4	1	2	0
9	7	2	6	3
10	9	3	16	2
TOTAL	22	6	26	5
	28		**31**	

- **Filling the demographic questionnaire**, participants filled in a demographic questionnaire (first name, last name, age, knowledge about BPMN models and GUIs) to identify their background. Each participant signed a consent form.
- **Solving the experimental problem**, participants had to solve two experimental problems described through videos (the order of the problems depend on the participants' profiles in Table 2). One where the stakeholder was a men and other where the stakeholder was a women (or vice versa). For each problem, the

participants had to build a BPMN model to represent the elicited requirements and to draw GUIs compliant with such requirements (as described in Section 3).

4.8. Data Analysis

We used a Mixed Model as statistical method since we are working with a within-participants factor (stakeholder gender and the block variable) and a between-participants factor (requirements engineer gender). The assumption for applying the mixed model is normality of residuals. The normality of residuals can be tested with Saphiro-Wilk test applied to the residuals automatically calculated during the application of the mixed model test. We checked the assumption of normality of residuals for all replications. The p-value of Mixed Model test shows whether or not there is a significant difference between treatments of each factor. If the p-value is less than 0.05, we assume that there are significant differences between treatments. Since the test does not indicate the magnitude of that difference, we use the non-parametric test called Cohen's d [44] to calculate the effect size. Cohen's d is defined as the difference between two means divided by a standard deviation of the data. According to Cohen [44], the meaning of the effect size is as follows: more than 0.8 is a large effect; from 0.79 to 0.5 is a moderate effect; from 0.49 to 0.2 is a small effect. We use this technique for the three response variables: Effort, BPMN model accuracy and GUIs accuracy. The power of any statistical test is defined as the probability of rejecting a false null hypothesis. Statistical power is inversely related to beta or the probability of making a Type II error. In short, power = $1 - \beta$. Power in software engineering experiments tends to be low, e.g. Dyba et al. [45] reports values of 0.39 for medium power and 0.63 for large power. Low values of power mean that non-significant results may involve accepting null hypotheses when they are false. The Mixed Model does not allow to calculate the statistical power for each response variable. Through GPower [46] we have estimated the minimum sample size that we need to avoid accepting false null hypothesis. According to GPower, to get a power of 0.96 with an effect size of 0.5 (moderate), we need a sample size of 16 men and 16 women. Since in our experiment we have 48 men and 11 women after the aggregation of both rounds, we can assume that we have high power in our statistical analysis for men, however we have low power statistical for women, this is a threat that could affect the results.

5. Threats to Validity

This section discusses the threats to validity of our experiment. In the following we describe the threats according to Wohlin's classification [36]. For each group of threats, we made a distinction between threats that we were unable to address, threats whose effect we managed to minimize, and threats that we solved. We classify the threats into four groups:

Conclusion validity: This type of threat deals with the ability to draw the correct conclusion about relations between the treatment and outcome. The experiment may suffer the following threats of this type: *Low statistical power,* which means that the number of participants is not enough to reveal a true pattern in the data. In order to minimize this threat, we applied the GPower to estimate the number of participants to avoid accepting false null hypotheses. We calculated that a power of 0.95 can be reached with 16 sample units. We have 48 men but 11 women, so we suffer this threat for the woman treatment. Another threat that appears is *Subjects of random heterogeneity,* which means that there is always heterogeneity in a study group. There is a risk that variation due to individual differences is larger than due to the treatment. In order to minimize this threat, we recruited undergraduate students of computer science with similar profiles. Moreover, we used a demographic questionnaire to know participants' background to properly interpret and draw final results. Furthermore, the number of women requirements engineers is lower than the number of men. The number of women may involve a low power compared with the power of men. Note importantly that this threat is very common in previous works that deal with gender analysis, where women are frequently in less number than men [20] [25] [27].

Internal validity: This type of threat deals with influences that can affect the factor concerning causality. The experiment may suffer the following threats of this type: *History,* which means that differences may arise when treatments are applied at different times. In

order to minimize this threat, we conducted the experiment in a unique session of one hour, even though each round took place in a different academic period. Another threat that appears is *Selection,* which means that how the participants are selected from a larger group may affect the results. In order to minimize this threat, we have recruited all the participants voluntarily and each participant signed a consent form. Another threat that appears is *Ecological validity* [47], which means that the context of use in which the experiment was conducted may affect the results. In order to minimize this threat, participants conducted the experiment in a controlled environment (i.e., a requirements laboratory) and not in a real environment, such as a corporate environments where real analysts and designers work collaboratively. Another threat that suffer the experiment is *External Factors*, which means that factors do not considered in the study may affect he results. This threat is common in gender studies according to Trauth et al. [13]. This threat appears mainly in factor stakeholders' gender, where we have only 1 participant per gender.

Construct validity: This type of threat concerns generalizing the result of the experiment to the concept or theory behind the experiment. Our experiment may suffer the following threats of this type: *Evaluation apprehension,* which means that some people are afraid of being evaluated. This threat may appear since participants were not anonymous. We needed to know who participates in order to get marks for the lesson where the experiment was conducted. In order to minimize this threat, we communicate to the participants that these experimental problems will allow to learn teaching objectives of the course.

External validity: This type of threats concerns with conditions that limit our ability to generalize the results of our experiments to industrial practice. Our experiment may suffer the following threats of this type. *Interaction of selection and treatment*, which means the effect of having a participant population not representative of the population we want to generalize. We cannot ensure that results of the experiment are valid for participants with a different profile of ours. Another threat that may appear is *Interaction of setting and treatment,* which means the effect of not having the experimental setting or material representative of industrial practice. The context of our experiment is just an academic environment and results can only be generalized in such context.

6. Results

This section reports the quantitative results of our experiment in order to address the research questions. All analyses have been performed using SPSS v 20. The raw data is the addition of both rounds through a moderator variable named "Course" with two possible values 2018/2019 and 2019/2020. We are not interested in studying possible differences between courses but we would like to check whether the course may affect the results.

6.1. Effort

Effort is measured as the time in minutes spent for the whole process (the less time spent in the development, the best effort). Fig. 8(a) shows the box-and-whisker plot comparing effort (in minutes) of men requirements engineers and women requirements engineers. The first quartile is similar for both men and women, but medians and third quartiles are different. The median of effort for men requirements engineers is lower than for women requirements engineers, so men requirements engineers are faster. Fig. 8(b) shows Box-and-whisker for effort combining requirements engineer gender and stakeholder gender. We see that when the stakeholder is a woman, the median for men requirements engineers is less than for women requirements engineers. When the stakeholder is a man, requirements engineers take less effort. In general, the median for men requirements engineers are slightly better than for women requirements engineers independently of the stakeholder gender. We can also appreciate that the time needed for participants that watched the video with the stakeholder-woman spent more time than the participants that worked with the stakeholder-man. This makes sense since woman's videos were slightly longer. The aim of our experimental investigation is to identify whether or not effort depends on the gender (H_{01}). The p-value for requirements engineer

gender is 0.018 with an effect size of 0.37 (small), showing differences between genders (Men requirements engineers yield less effort than women requirements engineers). This result means that requirements engineers need more time to process the requirements when the stakeholder is a woman. The p-value for stakeholder gender is 0.409, so in this case there are not significant differences between genders. Requirements engineer gender*problem interaction is significant with a p-value of 0.001. This means that when men requirements engineers solve Problem 2 we obtain the least effort. Stakeholder gender*problem interaction is not significant with a p-value of 0.557. We identify also that there are not significant differences in the moderator variable Course with a p-value of 0.058, even though the Course 2018/2019 yields better results than Course 2019/2020 according to the descriptive data. We conclude that ***H01*** for requirements engineers' gender is rejected, men took less effort. However, we cannot reject this null hypothesis for stakeholders' gender.

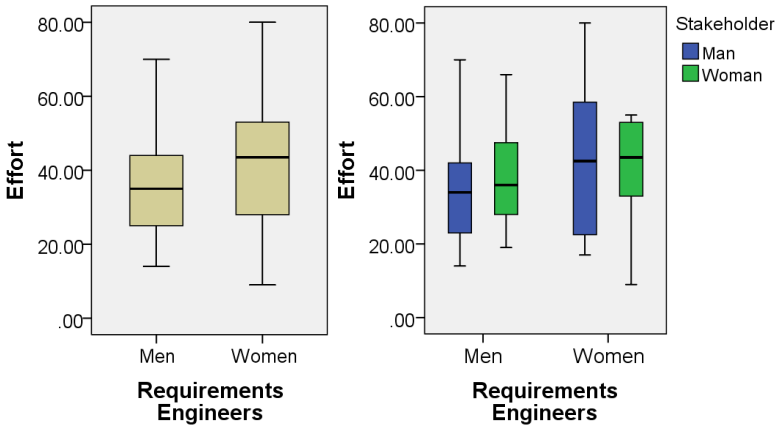

Fig. 8. (a) Box-and-whisker plot for effort comparing requirements engineers men versus requirements engineers women. (b) Box-and-whisker plot for effort considering stakeholder gender.

Table 4 shows the output of mixed model test for Effort.

Table 4 Output of mixed model test for Effort.

	p-value	Effect size	Average men	Average women	Average 2018-2019	Average 2019-2020
Requirement engineer gender	0.018	0.037	34.819	43.035	-	-
Stakeholder gender	0.409	-	35.288	39.827	-	-
Requirement engineer gender * Problem	0.001	-	-	-	-	-
Stakeholder gender * Problem	0.557	-	-	-	-	-
Course	0.058	-	-	-	33.940	41.176

6.2. BPMN model Accuracy

BPMN model accuracy was measured as the percentage of requirements expressed in the BPMN model developed by the participants that are compliant with the experimenters' solution (the higher percentage, the best BPMN model accuracy). Fig. 9(a) shows the box-and-whisker plot comparing BPMN model accuracy (percentage) of men requirements engineers and women requirements engineers. The median and third quartile is the same for both genders. The first quartile shows that women requirements engineers obtain better results for BPMN model accuracy.

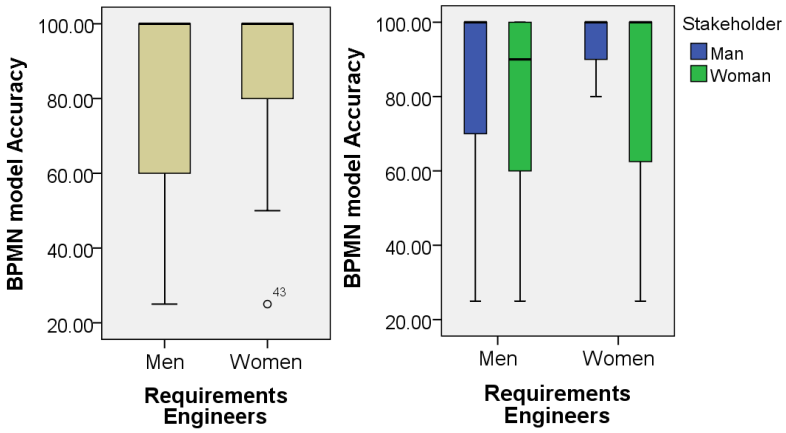

Fig. 9. (a) Box-and-whisker plot for BPMN model accuracy comparing men requirements engineers versus women requirements engineers. (b) Box-and-whisker plot for BPMN model accuracy considering stakeholder's gender.

Fig. 9(b) shows Box-and-whisker plot for the accuracy of the BPMN models for requirements engineers' gender and stakeholder's gender. We see that when the stakeholder is a woman, the median for women requirements engineers is higher than for men requirements engineers. When the stakeholder is a man, the median for women requirements engineers is the same as the median of men requirements engineers, but the results of men requirements engineers are more dispersed. In general, the median for women requirements engineers are better than for men requirements engineers independently of the stakeholder gender. This means that BPMN models built by women requirements engineers elicit more requirements than those built by men requirements engineers. We can also appreciate that the participants that watched the video with a men stakeholder developed better BPMN models than with women stakeholder. The aim of our experimental investigation is to identify whether or not BPMN model accuracy is independent of requirements engineers' gender and stakeholders' gender (*H02*). The p-value for requirements engineers' gender is 0.144, so there are not significant differences. The p-value for stakeholders' gender is 0.007 with an effect size of 0.23 (small), so there are significant differences showing that a stakeholder-man involves better accuracy in the construction of BPMN models. The requirements engineer gender*problem and stakeholder gender*problem interactions are not significant with a p-value of 0.05 and 0.744 respectively.

We identify also that there are significant differences in the Course with a p-value of 0.048 with an effect size of 0.58 (moderate) showing that the Course 2019/2020 involves better accuracy in the construction of BPMN models. We conclude that ***H02*** cannot be rejected both for requirements engineer gender. However, we can reject this null hypothesis for stakeholders' gender. Table 5 shows the output of mixed model test for BPMN model Accuracy.

Table 5 Output of mixed model test for BPMN model Accuracy.

	p-value	Effect size	Average men	Average women	Average 2018-2019	Average 2019-2020

6.3. GUIs Accuracy

Requirement engineer gender	0.144	-.907	.788	-	-	
Stakeholder gender	0.007	0.023	.880	-	-	
Requirement engineer gender * Problem	0.050	-	-	-	-	
Stakeholder gender * Problem	.744	-	-	-	-	
Course	0.048	0.058	-	-	.782	.8739

GUI accuracy was measured as the percentage of proposed requirements that are expressed in the GUIs developed by the participants that are the same with the experimenters' solution (the higher percentage, the best GUIs accuracy). Fig. 10(a) shows the box-and-whisker plot comparing GUIs accuracy for men and women requirements engineers. The third quartile is the same for both requirements engineers' genders. However, the median and first quartile shows slightly better accuracy for women requirements engineers. Fig. 10(b) shows Box-and-whisker plot for the accuracy of the GUIs considering both requirements engineers' gender and stakeholders' gender. We observe that for women stakeholders, the median for men requirements engineers is higher than for women requirements engineers. For a stakeholder-man, the median for women requirements engineers is higher than for men requirements engineers. In general, the median for women requirements engineers is slightly better than for men requirements engineers independently of stakeholders gender. This means that GUIs built by women requirements engineers elicit more requirements than those built by men requirements engineers. We can also appreciate that the participants that watched the video with a stakeholder-man developed better GUIs than the participants that watched the video with a stakeholder-woman.

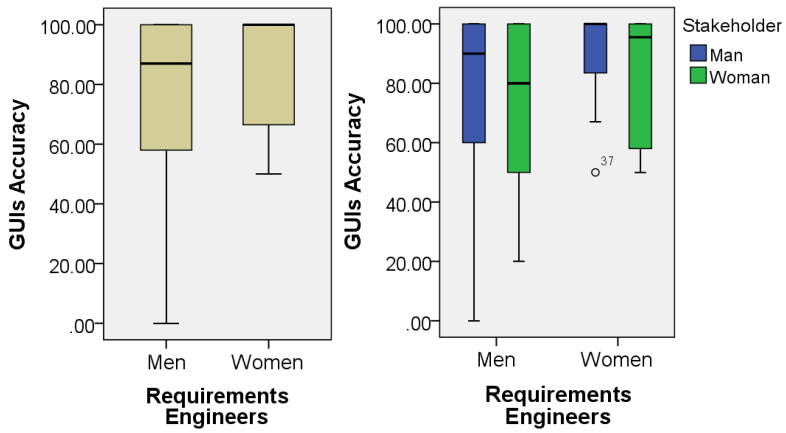

Fig. 10. (a) Box-and-whisker plot for GUIs accuracy comparing man requirements engineers versus women

requirements engineers. (b) Box-and-whisker plot for GUIs accuracy considering stakeholder gender.

The aim of our experimental investigation is to identify whether or not accuracy in the design of GUIs is independent of requirements engineer's gender and stakeholder's gender (*H03*). The p-value for requirements engineer gender is 0.132 and 0.741 for stakeholders' gender, so there are not significant differences. Requirements engineers' gender*problem and stakeholders' gender*problem interactions are not significant with a p-value of 0.094 and 0.377 respectively. Also the p-value of the course is of 0.29, which means that there are not significant differences.

We conclude that we cannot reject ***H03*** both for requirements engineers' gender and stakeholder's gender. So, there are not differences in gender for the accuracy in the design of GUIs. Table 6 shows the output of mixed model test for GUIs accuracy.

Table 6 Output of mixed model test for GUIs Accuracy.

	p-value	Effect size	Average men	Average women	Average 2018-2019	Average 2019-2020
Requirement engineer gender	0.132	-	74.743	93.957	-	-
Stakeholder gender	0.741	-	81.275	81.201	-	-
Requirement engineer gender * Problem	0.094	-	-	-	-	-
Stakeholder gender * Problem	0.377	-	-	-	-	-
Course	0.29	-	-	-	75.100	87.196

7. Discussion

Results related to **effort** show that there are significant differences between requirements engineers' gender; men requirements engineers took less effort than women requirements engineers. This means that most women requirement engineers in our experiment spent more time specifying the models and designing GUIs, and watched the video with requirements description more times than men requirements engineers. At first sight, it is possible that the different levels of experience of the participants may influence these results. In order to check this idea, we have repeated the Mixed Model including previous experience in BPMN modelling and in GUI design (Table 1) as block variables. We have analysed these block variables independently and their interaction with the design factors (requirements engineers' gender and stakeholders' gender). In all cases results yield no significant differences, so we can state that the previous experience is not affecting the results.

The result that shows a better effort in the case of men requirements engineers is similar to the work reported by Nguyen-Duc et al. [26]. According to Nguyen-Duc et al, in general, women are more active in requirements engineering and try to be more precise when they elicit requirements. This would justify that women watched the videos several times just to be more precise when specifying requirements. It is important to highlight that our demographic questionnaire revealed that men requirements engineers have more experience with BPMN models, which reduces the effort in building BPMN models. Even though there are not significant differences, men stakeholders yield better values of effort for requirements engineers according to the descriptive data in Fig. 8(b). This could be because the stakeholder-man had a shorter video. Schina et al. [25] described that women stakeholders are more emotional and enthusiastic, while men are more focused on a specific target. Given the fact that our research involves videos of man and stakeholders-woman that present equivalent information for specifying requirements, it is relevant to investigate to what extent men and women stakeholders can provide the same information but with different levels of abstraction and granularity. Significant differences in requirements engineers' gender*problem may be due to the fact that

Problem 1 has more requirements (three) than Problem 2 (two). Nevertheless, the two problems have the same complexity.

Regarding **BPMN model accuracy,** results show that requirements engineers' gender does not causes significant differences even though descriptive data evidences that women requirements engineers obtain the best results. Even though women requirements engineers invested more effort than men, BPMN models accuracy was better. These results are aligned with the work of Nguyen-Duc et al. [26], which states that women are more active with requirements engineering.

Analysing in detail the results for each problem, we see that in Problem 1, the combination of women requirements engineers with a men stakeholder yields the best accuracy, mainly in Requirement 1 and Requirement 2. The same combination of women requirements engineers with a stakeholder-man causes low accuracy levels for BPMN models when specifying Requirement 3. A possible cause of this phenomena could be explained by the need of specifying BPMN *Exclusive Gateway* elements, which are not frequent elements in comparison to BPMN *Lanes* and *Tasks*. We see that in Problem 2, the combination of men requirements engineer with a stakeholder-woman yields low levels of BPMN model accuracy. Analyzing this combination for each requirement individually, we see that Requirement 1 obtains the best BPMN accuracy while Requirement 2 yields the worst result, maybe because it involves using *Parallel gateway*, and requirements engineers opted for other BPMN notation such as *Exclusive gateway*.

Results of BPMN model accuracy for stakeholders gender show significant differences; a stakeholder-man involves better model accuracy. This result is also related with the result for the variable effort. Previously, we talked about the possibility that maybe a stakeholder-man took less effort because he had described the requirements more precisely. This precision may motivate requirements engineers to build more accurate BPMN models to specify such requirements. There are some studies such as the work of Sagnier et al. [24] that have identified a better performance in men, which is related with the idea of getting more precision when a man describes requirements.

Results from measuring **GUIs accuracy** do not provide significant differences among requirements engineer gender. Collected data evidences that women requirements engineers specified GUI models with high level of accuracy. This is aligned with the results of the BPMN model accuracy. A more accurate BPMN model to elicit requirements leads to a more accurate GUI design. These results are aligned with the work of Huang et al. [21], stating that women work better in visual design (like the GUI design).

Analyzing in detail the results of each problem, we see that in Problem 1, women requirements engineers obtain the best accuracy independently of stakeholders' gender. In general, men requirements engineers get the lowest accuracy in Requirement 3, since they used others widgets different from the experimenters' solution. We see that in Problem 2, the combination of men requirements engineers with a stakeholder-man gets better GUI accuracy than for women requirements engineers. This could be explained by the low number of requirements, which facilitates its interpretation.

Results of GUIs accuracy for stakeholders' gender show no significant differences. Men stakeholder yields better accuracy in the design of GUIs according to the descriptive data in Fig. 10(b), which is compliant with the result for accuracy in the BPMN model. Models with more accuracy involve designs of GUIs with more accuracy as well.

It is important to take into consideration the percentage of women requirements engineers that participated in our study (18.64%). This can be explained by the fact that the number of women students is considerable lower than men in computer science degrees [20]. This issue also appears in most of the related works described in this paper, such as the work of Ridley et al. [14] Fernandez-Sanz et al. [20], Schina et al. [25] or Tiwari et al. [27]. There are also works whose target is to analyse the low number of women in the field of computer science, such as the work of Beyer et al. [18]. Ioannis et al. [17] state that men prefer courses related to hardware and software engineering, whereas women opt for courses related to theory in computer science, humanities and social sciences. These unbalanced samples involve that small variations in the data of

women have more impact on the results than in the case of men. According to the results, we can state that high levels of accuracy and effort can be appreciated when requirements are described by men and the requirements are elicited by women. Even though women requirements engineers take more effort to build BPMN model and to design GUIs than men requirements engineers, their models are more accurate. This leads to propose mixed teams, such a way both genders can collaborate in the same target. This idea is aligned with some related works described previously (see section 1), such as the work of Marques et al. [16], which proposes mixed models to get more effective and coordinated teams.

Our research is aligned with the work of Trauth et al. [13] since we have considered both recommendations as they are described in the introduction of this paper. According to the first recommendation, we have divided the analysis into women and men to look for differences in the behaviour of both genders. The second recommendation states that differences between genders can be caused by other factors different from genders. This means that other factors apart from the gender not studied in the experiment may affect the results. For example, other factors such as pronunciation, or how much precise is the speaker may affect the results. We have considered these extra factors as a threat to validity.

Results of our experiment also contradict some ideas of the related work discussed in Section 2. Some of those works did not identified any difference between genders, such as the work of Fernandez-Sanz et al. [20] and Tiwari et al. [27], while we report significant and non-significant differences between men and women. Other previous works state that women obtain better performance with more theoretical tasks of computer science. This also contradicts our results, since both women and men requirements engineers tackle a practical problem, yielding women a better accuracy in the BPMN modelling. Other works such as Stoislescu et al. [15] conclude that women are not much interested in programming. Our results yield that even though there are not significant differences between the accuracy of the GUI, women requirements engineers tend to obtain better results.

Note that even though we identified significant differences between genders, these differences have a low effect. This means that differences are not so evident. Maybe the poor statistical power in the case of the recruited women may affect this low effect size. This effect also means that could be more factors beyond the gender do not considered in our study, which may affect the significant results. Both limitations, low power for women and the possibility of more factors, do not mean that results are not valid. Our experiment shows that even with not so large statistical power, results yield differences between genders. This means that more research must be done in this field.

8. Conclusion and Future Work

This paper analyses how gender may affect the requirements elicitation process. In particular, we focus on studying requirements engineers and stakeholders' effort, accuracy in BPMN models, and accuracy in GUIs. For this aim, we conducted an experiment where 59 participants played the role of requirements engineers (men and women). The experimental participants were asked to specify requirements regarding two given problems. The problems where described through videos with a man and a woman stakeholders who described different requirements. The participants were asked to watch the videos and build corresponding BPMN models that would represent the requirements stated by stakeholders. Next, these models were transformed into GUIs. Our experiment analyzes two factors: requirements engineers' gender and stakeholders' gender.

Results provided significant differences between requirements engineers' gender regarding invested effort; men requirements engineers are faster. There are also significant differences between stakeholders' gender for BPMN model accuracy, where a stakeholder-man gets better results. Analyzing descriptive data, we can also conclude that women requirements engineers yield better accuracy both in BPMN models and GUIs. Other conclusion extracted from descriptive data reveals that a stakeholder-man involves less effort and more accuracy in GUIs. How these results agree or disagree with the related works are also analyzed in the

discussion section. As conclusion of our experiment, we can recommend that mixed teams can lead to optimize both effort and accuracy.

The experiment suffers from the following limitations: the participants had few knowledge to build BPMN models before the experiment, but they had a good experience in the design of GUIs. We mitigated this gap by providing participants with a BPMN tutorial and a training problem. We only recruited one man and one woman as stakeholders. Other factors not related with the gender might also affect the results. The problems used during the experimental tasks were necessarily simple, since they had to be completed from scratch in one hour maximum according to experiment constraints.

Results must be interpreted within the context in which the experiment has been run: (i) Participants were students of undergraduate computer science. (ii) the sample of participants is low (even though the statistical power is enough). (iii) Experimental problems are not very complex and are based on the functional behavior of the system. The influence that the experimental problem caused on the dependent variables is part of our future research endeavors.

As future work, we plan to replicate this experiment changing some elements of the design. First, we plan to elaborate a video of requirements with several stakeholders. Second, we aim to recruit participants that are working in practical settings as requirements engineers. Third, we plan to conduct replications such a way we could aggregate experimental data and enhance the statistical power.

Acknowledgements The first author has the support of the Ministry of Education of Peru with the National Scholarship Program PRONABEC – Republic President. This project also has the support of Spanish Ministry of Science and Innovation through project DATAME (ref: TIN2016-80811-P). We would like to thank the participants of the experiments.

References

[1] F. P. Brooks and N. S. Bullet, "Essence and accidents of software engineering," *IEEE computer,* vol. 20, pp. 10-19, 1987.

[2] A. M. Hickey and A. M. Davis, "A unified model of requirements elicitation," *Journal of Management Information Systems,* vol. 20, pp. 65-84, 2004.

[3] E. O. Mkpojiogu, G. E. Akusu, A. Hussain, and W. Hashim, "Eliciting and modeling the requirements for anoOnline data archival management system," *International Journal of Advanced Science and Technology,* vol. 29, pp. 296-306, 2020.

[4] S. Zareen, A. Akram, and S. Ahmad Khan, "Security Requirements Engineering Framework with BPMN 2.0. 2 Extension Model for Development of Information Systems," *Applied Sciences,* vol. 10, p. 4981, 2020.

[5] D. W. Embley, S. Liddle, and Ó. Pastor, "Conceptual-Model Programming: A Manifesto," in *Handbook of Conceptual Modeling*, ed: Springer, 2011, pp. 3-16.

[6] D. Rosenberg and K. Scott, *Use case driven object modeling with UML*: Springer, 1999.

[7] X. Li, Z. Liu, and H. Jifeng, "A formal semantics of UML sequence diagram," in *2004 Australian Software Engineering Conference. Proceedings.*, 2004, pp. 168-177.

[8] G. Lucassen, F. Dalpiaz, J. M. E. van der Werf, and S. Brinkkemper, "The use and effectiveness of user stories in practice," in *International working conference on requirements engineering: Foundation for software quality*, 2016, pp. 205-222.

[9] BPMN. (2013). *Business Process Modeling Notation*. Available: http://www.bpmn.org

[10] K. Decreus and G. Poels, "A goal-oriented requirements engineering method for business processes," in *International Conference on Advanced Information Systems Engineering*, 2010, pp. 29-43.

[11] M. Brambilla, P. Fraternali, and C. Vaca, "BPMN and design patterns for engineering social BPM solutions," in *International Conference on Business Process Management*, 2011, pp. 219-230.

[12] T. R. Silva, M. Winckler, and H. Trætteberg, "Ensuring the Consistency Between User Requirements and GUI Prototypes: A Behavior-Based Automated Approach," in *IFIP Conference on Human-Computer Interaction*, 2019, pp. 644-665.

[13] E. M. Trauth, "The role of theory in gender and information systems research," *Information and Organization,* vol. 23, pp. 277-293, 2013.

[14] G. Ridley and J. Young, "Theoretical approaches to gender and IT: examining some Australian evidence," *Information Systems Journal,* vol. 22, pp. 355-373, 2012.

[15] D. Stoilescu and G. Egodawatte, "Gender differences in the use of computers, programming, and peer interactions in computer science classrooms," *Computer Science Education,* vol. 20, pp. 283-300, 2010.
[16] M. Marques, "Software engineering education—Does gender matter in project results?—A Chilean case study," in *2015 IEEE Frontiers in Education Conference (FIE)*, 2015, pp. 1-8.
[17] B. Ioannis and K. Maria, "Gender and student course preferences and course performance in Computer Science departments: A case study," *Education and Information Technologies,* vol. 24, pp. 1269-1291, 2019.
[18] S. Beyer, "Why are women underrepresented in Computer Science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades," *Computer Science Education,* vol. 24, pp. 153-192, 2014.
[19] M. Brambilla, S. Butti, and P. Fraternali, "Webratio bpm: a tool for designing and deploying business processes on the web," in *International Conference on Web Engineering*, 2010, pp. 415-429.
[20] L. Fernandez-Sanz and S. Misra, "Analysis of cultural and gender influences on teamwork performance for software requirements analysis in multinational environments," *IET software,* vol. 6, pp. 167-175, 2012.
[21] F.-H. Huang, "An experimental study of home page design on green electronic products web site," in *International Conference on Human Centered Design*, 2011, pp. 509-518.
[22] C. K. Gonzales and G. Leroy, "Eliciting user requirements using appreciative inquiry," *Empirical Software Engineering,* vol. 16, pp. 733-772, 2011.
[23] D. Cooperrider, D. D. Whitney, J. M. Stavros, and J. Stavros, *The appreciative inquiry handbook: For leaders of change*: Berrett-Koehler Publishers, 2008.
[24] C. Sagnier, E. Loup-Escande, and G. Valléry, "Effects of Gender and Prior Experience in Immersive User Experience with Virtual Reality," in *International Conference on Applied Human Factors and Ergonomics*, 2019, pp. 305-314.
[25] D. Schina, M. Usart, and V. Esteve-Gonzalez, "Participants' Perceptions About Their Learning with FIRST LEGO® League Competition–a Gender Study," in *International Conference on Robotics and Education RiE 2017*, 2019, pp. 313-324.
[26] A. Nguyen-Duc, L. Jaccheri, and P. Abrahamsson, "An empirical study on female participation in software project courses," in *Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings*, 2019, pp. 240-241.
[27] S. Tiwari, D. Ameta, P. Singh, and A. Sureka, "Teaching requirements engineering concepts using case-based learning," in *2018 IEEE/ACM International Workshop on Software Engineering Education for Millennials (SEEM)*, 2018, pp. 8-15.
[28] S. Sobieraj and N. C. Krämer, "Similarities and differences between genders in the usage of computer with different levels of technological complexity," *Computers in Human Behavior,* vol. 104, p. 106145, 2020.
[29] A. Z. Al-Othmani, A. A. Manaf, A. M. Zeki, Q. Almaatouk, A. Aborujilah, and M. T. Al-Rashdan, "Correlation Between Speaker Gender and Perceptual Quality of Mobile Speech Signal," in *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2020, pp. 1-6.
[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, 2001, pp. 749-752.
[31] S. A. White, *BPMN Modeling and reference guide*, First Edition ed., 2008.
[32] E. Diaz, J. I. Panach, S. Rueda, and O. Pastor, "Towards a method to generate GUI prototypes from BPMN," in *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, 2018, pp. 1-12.
[33] E. Diaz, J. I. Panach, S. Rueda, and O. Pastor, "Generación de Interfaces de Usuario a partir de Modelos BPMN con Estereotipos," presented at the Jornada de la Sociedad de Ingeniería de Software y Tecnologías de Desarrollo de Software (SISTEDES), 2018.
[34] L. Han, W. Zhao, and J. Yang, "An approach towards user interface derivation from business process model," *Communications in Computer and Information Science,* vol. 602, pp. 19-28, 2016.
[35] K. Sousa, H. Mendonça, and J. Vanderdonckt, "User Interface Derivation from Business Processes: A Model-Driven Approach for Organizational Engineering," in *TAMODIA*, Toulouse (France), 2007, pp. 112-125.
[36] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction*: Springer, 2012.
[37] J. P. Winkler, J. Grönberg, and A. Vogelsang, "Optimizing for recall in automatic

requirements classification: An empirical study," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019, pp. 40-50.
[38] N. Ali and R. Lai, "A method of requirements elicitation and analysis for Global Software Development," *Journal of Software: Evolution and Process,* vol. 29, p. e1830, 2017.
[39] IEEE, "Systems and software engineering -- Vocabulary," ISO/IEC/IEEE 24765:2010(E), Ed., ed, 2010, pp. 1-418,.
[40] ISO/IEC, "ISO/IEC 9126-1, Software engineering - Product quality - 1: Quality model," ed, 2001.
[41] N. Juristo and A. Moreno, *Basics of Software Engineering Experimentation*: Springer, 2001.
[42] Byron Wm. Brown, Jr., "The Crossover Experiment for Clinical Trials," *Biometrics,* vol. 36, pp. 69-79, 1980.
[43] M. Jirotka and P. Luff, "Supporting requirements with video-based analysis," *IEEE software,* vol. 23, pp. 42-44, 2006.
[44] L. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd. Edition ed.: Lawrence Earlbaum Associates, 1988.
[45] T. Dybå, V. B. Kampenes, and D. I. K. Sjøberg, "A systematic review of statistical power in software engineering experiments," *Information and Software Technology,* vol. 48, pp. 745-755, 2006.
[46] E. Erdfelder, F. Faul, and A. Buchner, "GPOWER: A general power analysis program," *Behavior Research Methods, Instruments, & Computers,* vol. 28, pp. 1-11, 1996/03/01 1996.
[47] S. Kieffer, "ECOVAL: Ecological validity of cues and representative design in user experience evaluations," *AIS Transactions on Human-Computer Interaction,* vol. 9, pp. 149-172, 2017.