Improving the Requirement Elicitation Process using Empathy Maps and Personas: A Quasi-Experiment

Ezequiel Kahan[1], Emilio Insfran[2], Marcela Genero[3], Alejandro Oliveros[1]

[1] Universidad Nacional de 3 de Febrero, Buenos Aires, Argentina
[2] IUMTI - Universitat Politècnica de València, Valencia, Spain
[3] University of Castilla-La Mancha, Ciudad Real, Spain

ekahan@untref.edu.ar, einsfran@dsic.upv.es,
marcela.genero@uclm.es, aoliveros@untref.edu.ar

Abstract. There is a growing interest in the use of Design Thinking (DT) to enrich requirements elicitation processes. This fact motivated us to explore the use of user-centered and empathy-oriented techniques taken from the DT process, in combination with the Brainstorming technique, usually used alone, for requirements elicitation. Specifically, we carried out a quasi-experiment to compare the *Effectiveness* of Brainstorming sessions in terms of the number of ideas of requirements generated, and the *Identified Stakeholders*, complementing the Brainstorming sessions with two of the most widely used DT techniques: Empathy Maps and Personas. Therefore, we consider three treatments: Personas + Brainstorming, Empathy Maps + Brainstorming, and Brainstorming alone (control group). The quasi-experiment was carried out with 74 students enrolled in the Bachelor of Computer Engineering course at the Universitat Politècnica de València in 2021. The results reveal a statistically significant effect on *Identified Stakeholders* when using Empathy Maps. Descriptive analysis also shows an increase in *Identified Stakeholders* when using Personas, and in the *Effectiveness* of Brainstorming sessions when used together with Empathy Maps or Personas. There is also a variation in the type of ideas, with the number of functional ideas being higher, and that of non-functional ideas being lower when Empathy Maps are used. These results seem to indicate that Brainstorming sessions are improved when complemented with Empathy Maps or personas techniques. However, we still do not have enough evidence to recommend either Personas or Empathy Maps. Therefore, further experimentations will be needed to obtain more conclusive results considering some improvements that are discussed in the paper.

Keywords: Requirement Engineering, Requirements elicitation process, Design Thinking, Empathy Map, Personas, Quasi-experiment.

1 Introduction

The first work that linked Design Thinking (DT) to Requirements Engineering (RE) appeared more than nine years ago [1] and, since then, the interest in this topic has been increasing. Nowadays, there are several studies demonstrating the potential of applying DT in synergy with RE, in particular with the requirements elicitation phase [2–5]. Although these studies show that it would be possible to improve the RE process by the

Proceedings of the 26th Workshop on Requirements Engineering (WER23), August 15-17, 2023, Porto Alegre, RS, Brazil.
DOI: 10.29327/1298356.26-2

application of DT [6], there is insufficient evidence as to which DT techniques are more appropriate or yield better results. As Brainstorming is often used as an individual elicitation technique [7], previous research would suggest that it is possible to improve the *Effectiveness* of the ideas of requirements generated during a Brainstorming session, complementing it with other user-centered and empathy-driven ideas [5]. As already stated in previous work, empathy is a key feature of the requirements elicitation process and also of DT in general [5, 8]. Empathy is a concept that includes both the *involuntary act* of feeling sympathy for someone else and the *cognitive act* of placing oneself in another's position and adopting their perspective [9]. It is the attempt to reconstruct the specific perspective of another person and how they perceive the situation. Although empathy development occurs in all stages of DT, it is particularly relevant during the "Empathize" stage, which is usually the first stage of the process.

In an earlier quasi-experiment, we proposed a requirements elicitation process that included an empathy stage [5]. The independent variable of interest was the elicitation technique used with two treatments 1) Empathy Maps + Personas + Brainstorming (EM+P+B), and 2) Personas + Brainstorming (P+B). The dependent variable was *Effectiveness*, measured as the quantity of ideas of requirements (QIR) generated. Several interesting insights emerged from this study, such as an increase in the number of functional ideas of requirements, and an improvement in the perception of usefulness of the Brainstorming technique when using Empathy Maps. Some of the learnings and limitations of this quasi-experiment were: 1) not having defined a specific treatment to evaluate the Empathy Maps technique alone prevented from discovering if it contributed more or less to the Brainstorming session; 2) also, the subject´s perception revealed that the Perceived Utility of Personas was lower in the Empathy Maps and Personas treatment (first treatment) - this could be interpreted as a certain level of overlapping between Empathy Maps and Personas; and 3) after analysing the ideas of requirements generated by the participants, several inconsistencies, contradictions, or reiterations were found. This showed the need to provide the participants with a template for specifying the ideas of requirements during the generation phase of the Brainstorming session.

Considering the above limitations, and the need for more evidence about the usefulness of DT techniques for the requirements elicitation process, we proposed a new quasi-experiment, in which the following changes were introduced: 1) Empathy Maps and Personas techniques, were separated into two different treatments to compare the contribution of these techniques when used separately. In addition, a third treatment was added as a control group, in which no other techniques were applied prior to the Brainstorming session; 2) An analysis of the identified stakeholders was included as well. This information is valuable when analysing the quality of the ideas of requirements generated, because more accurate stakeholders identification allows to get a more complete understanding of the needs of the system. The quantity of stakeholders was consider as additional dependent variable; 3) given the popularity of User Stories (US) in the agile software development [9-11], it was decided in this new quasi-experiment to provide US as the reference template for specifying the ideas of requirements. The rest of the experiment conditions, including instructions and supporting materials were the same of our previous quasi-experiment.

This paper presents the results of this new quasi-experiment carried out to evaluate the effect in the *Effectiveness* and in the *Identified Stakeholders* of Brainstorming sessions when using them together with Empathy Maps or Personas. *Effectiveness* was measured in terms of the quantity of ideas of requirements obtained in the Brainstorming session. We also measured the quantity of stakeholders identified by the students and the distribution of ideas among them. The quasi experiment was conducted with a group of 74 students enrolled on a Bachelor's degree course at the Universitat Politècnica de València in October 2021.

The remainder of this paper is structured as follows. Section 2 provides an overview of the related work. Section 3 introduces the main characteristics of the quasi-experiment; Section 4 presents the data analysis and interpretation of the data collected in the quasi-experiment; and Section 5 discusses the threats to validity. Finally, Section 6 presents the conclusions and outlines suggestions for future work.

2 Related work

The interest in employing DT techniques for requirements elicitation has grown in the Information System field in recent years, as evidenced by the fact that several secondary studies have appeared on the subject [12, 13]. However, the empirical studies that evaluate the contribution of usual techniques in DT in RE, such as Empathy Map, or Persona, are still very scarce. Searching Scopus for the following search string "("experiment*" OR "empiric*" OR "survey" OR "case study" OR "action research") AND ("Empathy Map" OR "persona") AND Requirement", only 3 studies were found: Teixeira et al. In [14] the use of the Lean Persona technique with 21 software professionals is investigated. They carried out a comparison to see whether the startup professionals use the technique in a different way from the established company professionals. Results revealed that the professionals used the technique for similar purposes and wrote up UX-related requirements in different levels of abstraction. Costa et al. [15] carried out an exploratory case study with 17 undergraduate Computer Science students with the aim of discovering: "What are the perceptions of students regarding learning DT?". Projects using individual techniques (Personas, Empathy Maps) and team techniques (Brainstorming and co-creation workshop) were then employed for the development of the authors' mobile application. The students considered techniques very useful but stated that more training time was required to carry out the case study. Ferreira et al. [16] presents a controlled experiment carried out with 37 Computer Science undergraduates in order to compare two Personas-related techniques: traditional Personas and PATHY. The authors analysed the efficiency of the techniques and the participants' perceptions of their use. PATHY generated more relevant characteristics for the application design than did the technique that followed the traditional description. It was also more efficient as regards creating Personas.

The existing evidence is therefore isolated empirical studies on different DT techniques, which differ from the objective pursued in the current quasi-experiment (see Section 3.1.), which is part of a long-term investigation whose first results, as mentioned in Section 1, were presented in [5].

3 Quasi-experiment description

The main characteristics of the quasi-experiment are described in the following subsections. This quasi-experiment was designed and reported by following the recommendations provided in [18]. Due to space constraints, the experimental material, guidelines to perform experimental tasks, and examples of the results of the experimental tasks performed by the subjects have been published online as an appendix [17].

3.1 Goal, variables and hypotheses

Following the GQM template [19], the goal of this quasi-experiment was **to analyse** *Elicitation Techniques* **for the purpose of** comparing them **with respect to** their *Effectiveness* and *Identified Stakeholders* **from the point of view** of requirements analysts **in the context** of students enrolled on a Bachelor's degree course in Computer Engineering.

The independent variable is the *Elicitation Technique* used, taking into consideration three treatments: Personas + Brainstorming, Empathy Map + Brainstorming, and Brainstorming alone (P+B, EM+B and B, respectively).

The dependent variables were *Effectiveness* and *Identified Stakholders*. The authors of this paper consider that a greater number of ideas of requirements generated during a Brainstorming session implies a greater *Effectiveness* for it. Also, the identification of a greater number of stakeholders could implies a greater degree of completeness in the requirements elicitation process [20], since they represent the holders of the needs and goals of the problem to be solved. Therefore, the following hypotheses was formulated:

- H1-0: There is no significant difference between the subjects' *Effectiveness* when using P+B or EM+B or B / H1-a: ≠ H1-0.
- H2-0: There is no significant difference between the subjects' *Identified stakeholders* when using P+B or EM+B or B / H2-a: ≠ H2-0.

Effectiveness and *Identified stakeholders* were measured as being the quantity of ideas of requirements generated by the students (QIR), and the quantity of different stakeholders identified by the students (QS), respectively. To define the measure QIR, since the ideas of requirements generated by the students were very different, the ideas of requirements were clustered into two categories, as is usual in RE processes:

- Functional ideas / Business-oriented (QIR-F). This category included all the ideas of requirements that describe or propose functionalities for the software system for an Animal Adoption Centre (the problem domain chosen for this quasi-experiment, which is introduced in Section 3.4).
- Non-functional ideas (QIR-NF). This category included all the ideas of requirements that describe or propose restrictions or constraints for the software system to be developed [20]. This category was sub-divided into two sub-categories: Technology-oriented ideas (QIR-NF-T), which refer to ideas of requirements that establish

technological needs, and People-centered needs (QIR-NF-P), which refer to ideas of requirements where people that will use the application are the central target.

When classifying the ideas of requirements, it was necessary to define another category, called *Others*, to deal with those proposed ideas of requirements not directly related to the software system to be developed, e.g., "*Creation of tutorials on how to properly take care for animals"*. Therefore, these ideas were not considered to measure the *Effectiveness*, which was calculated using the following formula: QIR = QIR-F + QIR-NF. Regarding the *Identified stakeholders*, a baseline with the stakeholders of interest was defined, according to the problem description. This baseline was used to compare with the stakeholders identified by the students (QS) and determine how complete and correct were the points of views considered by the students when proposing ideas of requirements. Although the *quality* of the ideas of requirements was not directly evaluated, we understand that the total quantity of different ideas of requirements, which are directly related to the problem domain addressed, together with the correctly identified stakeholders, may be considered an as indicator of the quality of the ideas of requirements identified in terms of coverage.

3.2 Selection of the subjects

We took a convenience sample of undergraduate students enrolled on a Bachelor's degree course in Computer Engineering the Universitat Politècnica de València. The students attended a theorical-practical course on RE during the academic year 2021-2022. The practical part of the course was divided into 3 class time shifts. This course included an introduction to and examples of the use of the techniques employed in the quasi-experiment, i.e., Personas, Empathy Maps, and Brainstorming. The students had no prior experience in the use of any of these three techniques. Finally, considering that the main purpose of the quasi-experiment was to study the improvement of a Brainstorming session when using Empathy Maps or Personas, and that Brainstorming is a group-based technique, we set up several working groups with which to run the quasi-experiment. For this reason, each time shift was divided in groups, between 4 to 7 students, were randomly assigned by the course instructor. Three types of groups were defined: Groups A, which used Personas together with Brainstorming (27 students organized in 5 groups); Groups B, which used Empathy Map together with Brainstorming (33 students organized in 6 groups), and finally, Groups C, which used only Brainstorming (14 students organized in 2 groups). It was decided that the time shift with the fewest students would be assigned to group C (control group).

3.3 Experimental object, tasks, and design

The experimental object of the quasi-experiment describes the characteristics and principal needs of an Animal Adoption Centre, called "MODEPRAN". This description provided the subjects with an overview of and context in which to begin identifying the main stakeholders, and the scope in which to propose the ideas of requirements for the software system during the Brainstorming sessions. This domain was chosen because

the participants may be familiar with the problem to address, and also because it does not have a strong technical component. The authors of this paper consider that a very unknown or highly-technical problem could influence negatively the objective of the experiment. In addition, this case has a moderate length, that can be addressed in one lab session without the need for an intensive training or explanation of the concepts to be managed. Both, the case and the support material, are the same used in [5], so it has been tested and validated in terms of clarity of the requested instructions. The experimental task included the generation of ideas of requirements by means of a Brainstorming session using only Personas, in the case of the Groups A (P+B treatment), using Empathy Maps, in the case of the Groups B (EM+B treatment), or running the Brainstorming session without any previous technique Groups C (B treatment). Both Personas and Empathy Maps were created by the group of students themselves using the provided material. Multiple documents were defined as instrumentation [17], including: i) the problem statement; ii) guidelines on how to run a Brainstorming session and the template required to report the ideas of requirements generated; iii) an introduction to the Personas technique with examples; iv) an introduction to the Empathy Maps technique with examples. A between-subject design was used, meaning that the subjects (i.e., working groups) in the quasi-experiment were assigned to different treatments, with each working group experiencing only one of the treatments.

3.4 Execution

The students were not aware that they were participating in a quasi-experiment. For them, this activity was just another exercise in the context of the RE course on which they were enrolled. Since the RE course is a weekly course of three hours per week, the training and the experiment were performed in two sessions over two weeks. The first week was the training session, whose purpose was to introduce the concepts, examples, and short exercises concerning the techniques that would then be applied in the quasi-experiment: Personas, Empathy Map, and Brainstorming. The quasi-experiment took place in the second week. During the execution, students were assigned to one of three groups A, B or C, and organised in smaller sub-groups composed of between four to seven students. The quasi-experiment was controlled, meaning that no interactions took place between the working groups. The training and experimental sessions lasted approximately three hours each. Once the quasi-experiment had finished, two of the authors of this paper classified the ideas of requirements obtained by each of the working groups in accordance with the classification introduced in Section 3.1 (i.e., functional, non-functional, others). The authors of this paper then analysed and classified each idea of requirement into one or more of the categories defined in Section 2.2, reaching a consensus when necessary. Examples of the results of the experimental tasks performed by the groups can be found in [17].

4 Data Analysis and Interpretation

In this section the data analysis and interpretation of the results obtained in the quasi-experiment is presented.

4.1 Analysis of Effectiveness

Table 1 classifies the descriptive statistics of the ideas of requirements. Groups A and B shows bigger mean values than Groups C (the control group), which might indicate a contribution of Empathy Maps and Personas on the generation of ideas of requirements.

Table 1. Descriptive statistics of QIR

	Groups A *(Treatment P+B)*				**Groups B *(Treatment EM+B)***				**Groups C *(Treatment B)***			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
QIR	**37.4**	5.68	30	46	**40.67**	5.20	34	47	**35.5**	0.71	35	36
QIR- F	**31.20**	8.17	22	44	**36.17**	3.87	32	43	**30.5**	2.12	29	32
QIR-NF	**6.2**	5.17	2	15	**4.5**	3.94	1	12	**5**	2.83	3	7
QIR-NF-T	**5.2**	3.70	2	11	**2.67**	2.87	0	8	**3**	1.41	2	4
QIR-NF-P	**1**	1.73	0	4	**1.83**	2.14	0	5	**2**	1.41	1	3

QIR-F is 15.9% higher and QIR-NF is 37.77% lower in Groups B than A, showing that there is also a difference in the number of functional and non-functional ideas generated between groups. Similar to previous quasi-experiment [5], there is a reduction in QIR-NF-T, showing that in the Empathy Maps treatment, fewer technological ideas of requirements were generated and more functional ones. Table 2 compares Number of ideas of requirements by category (QIR, QIR-F, QIR-NF, QIR-NF-T and QIR-NF-P) as between the previous quasi-experiment and this current one. The number of ideas generated in this quasi-experiment was on average 21.5% higher than in the previous one [5].

Table 2. Number of ideas of requirements (QIR) by category - related measures in the previous and current quasi-experiment

Experiment	**Treatment**	**QIR-F**	**QIR-NF**		**QIR**	**Mean**
			QIR-NF-T	**QIR-NF-P**		
Previous experiment (2019) [5]	*P+B*	147	42	31	220	**31.43**
	E+P+B	157	28	31	216	**30.86**
Current experiment (2021)	*P+B*	156	26	5	187	**37.4**
	E+B	217	16	11	244	**40.67**
	B	61	6	4	71	**35.5**

Although there were differences in the groups between both experiments, the instructions and support material were similar in both, so these differences can probably be explained by the use of the template of US to support the specification of the generated ideas of requirements. To test the hypothesis formulated, we analysed the effect of every treatment (P, EM, and B) on the measures considered (QIR, QIR-F, QIR-NF, QIR-NF-T; QIR-NF-P) using the non-parametric Kruskal Wallis test. All these values

were calculated using a standard configuration of SPSS. Is was also carried out the non-parametric Mann-Whitney U test, taking Groups A and C, and Groups B and C separately. Results obtained do not allow to reject H1-0, i.e., the techniques had no effect on the QIR. Similar results were obtained after repeating the test for each individual variable (QIR-F, QIR-NF, QIR-NF-T and QIR-NF-U), i.e., it was not possible to reject H1-0 in any of the cases. In the case of the comparison between Groups B and C, the values for QIR-F were close to the rejection condition (p-value = 0.062), with a moderate Observed Power (OP) = 0.362, which indicates a slight correlation. As in previous quasi-experiment [5], the total number of ideas does not differ significantly between the three techniques. The number of functional ideas was higher, and that of non-functional ideas was lower in Groups B, to which the EM+B techniques were applied. This supports the idea that Empathy Maps enabled subjects to become more aware (or emphatic) of functional requirements than non-functional ones. However, this result was not significant enough to confirm the hypothesis - this may be because the number of groups involved in the quasi-experiment was not large enough.

When comparing the ideas generated in this quasi-experiment against the previous one, there is an increase in the total number of ideas of requirements generated, as well as in the number of ideas of functional requirements. A possible explanation is that the use of US helped the participants to better focus and conceptualise an idea of requirement that is relevant to the problem. For both groups, A (P + B treatment) and B (EM + B treatment), the number of ideas and the number of stakeholders is higher than in the control treatment. A significant aspect seems to be the fact that the number of ideas that arise in the current quasi-experiment applying the Personas technique (Groups A) or the Empathy Map (Groups B) is greater than the number of ideas generated by applying both techniques together (Empathy Maps and Personas) to run the Brainstorming session, as in the previous quasi-experiment. It can be observed that there is no positive effect in terms of the number of ideas when using both techniques (Personas and Empathy Maps) together. We believe that it can be valuable to use both techniques in cases where it is necessary to define stakeholders in a more "formal" way, something that the structure of the Personas technique could do more effectively. Empathy Maps provide a more general view of stakeholders emphasizing their feelings and thoughts, meanwhile Personas provides a more descriptive view of the stakeholders. Moreover, there is no statistically significant difference between these techniques that would allow us to suggest or recommend one over the other.

4.2 Analysis of Identified stakeholders

To assess the *Identified Stakholders* of the ideas of requirement, we built a baseline against which to compare them. This baseline was built from the explicit description of stakeholders in the material provided to the students and includes the following 6 stakeholders: adopter, donor, employee, partner, sponsor, and volunteer. QSB variable was created to measure the number of stakeholders that the participants identified, and which coincided with the baseline established by the authors. From the analysis of the descriptive statistics shown in Table 3, it was observed that Personas and Empathy

Maps techniques contribute to a greater identification of stakeholders, with Groups A identifying 40%, and Groups B 50%, more than Groups C (control group).

Table 3. Descriptive statistics of QS

	Groups A *(Treatment P+B)*				**Groups B *(Treatment EM+B)***				***Groups C (Treatment B)***			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
QS	7	1.22	5	8	7.5	1.38	5	9	5	1.41	4	6
QSB	5.4	0.89	4	6	5.5	0.84	4	6	3.5	0.71	3	4

To test the hypothesis related to *Identified Stakeholders* (H2-0), the effect of each one of the treatments (P + B, EM + B, B) was analysed on the measures considered (QS, QSB) using the non-parametric Kruskal Wallis test. It was also carried out the non-parametric Mann-Whitney U test in pairs taking Groups A and B, A and C, and B and C separately. The results obtained allow to reject H2-0: for Groups B / C, variable QSB, given that the p-value is 0.049, which is lower than 0.05, i.e., the Empathy Maps influenced QS. For Groups A / C, even when p-value was higher than 0.05 for all the variables, in the case of QSB with a p-value = 0.068, and OP of 0.604, the result and the observed power allow us to make a slightly correlation between the treatment and the result achieved. From the above, it can be said that both Personas and the Empathy Maps helped identify the essential stakeholders, which evidences the empowerment of using these techniques in combination with Brainstorming. In the case of Groups C, not only was the number of stakeholders significantly lower, but also the percentage of essential stakeholders identified. Additionally, upon analysing the stakeholders found by the subjects, it seemed valuable to establish a categorisation of stakeholders in a way that could reflect some of the expected contribution of the techniques. Authors proposed grouping the stakeholders into 3 categories that reflect the relation of the interested party with the Animal Adoption Centre. The categories proposed were:

- Ideas related to Internal stakeholders (IRI): includes the ideas related to the stakeholders who are staff members of the Animal Adoption Centre. Examples of this are: Veterinarian, Employee, Manager, etc.
- Ideas related to External stakeholders (IRE): includes the ideas related to the stakeholders who are not staff member of the Animal Adoption Center. Examples of this are: Sponsor, Volunteer, etc.
- Ideas related to Internal and External stakeholders (IRIE): includes the ideas related to the stakeholders that may either be employees or people who are not part of the Animal Adoption Center. Included in this are generic stakeholders, or ideas of requirements that mention more than one stakeholder for a single idea, and where those stakeholders belong to both the categories.

It was evaluated how many of the stakeholders identified in the different groups belonged to the IRI, IRE and IRIE categories. Table 4 shows descriptive distribution of ideas for the different groups according to the proposed categories.

Table 4. Descriptive statistics of IRI, IRE and IRIE

	Groups A *(Treatment P+B)*				**Groups B *(Treatment EM+B)***				***Groups C (Treatment B)***			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
IRI	18.20	3.11	13	21	17.33	3.50	11	21	9.50	0.71	9	10
IRE	16.00	3.81	11	21	17.83	7.30	4	24	13.50	3.54	11	16
IRIE	3.20	2.95	0	6	5.50	5.24	1	15	12.50	2.12	11	14

In Groups A, the IRI category is the majority, with almost 49% of the ideas. In Groups B, although the percentage is slightly lower than the IRE category (43% vs 44%), it remains at a high value compared to Groups C. In this, the IRI category is only 27%. This seems to indicate that the techniques contribute to the identification of ideas related to internal stakeholders. In the case of Groups C, where no techniques were applied, most of the ideas are related to external stakeholders. In addition, we see that the number of ideas that are simultaneously attributable to both categories is, in the case of Groups C, 3.8 to 1 time greater than Groups A, and 2.5 to 1 times greater than Groups B. In terms of the quantity of stakeholders, this would indicate that in the case of Groups C there is a greater proportion of "generic" ideas than is attributable to either type of stakeholder, which makes us suppose that the techniques help to generate more “precise” ideas, attributable to a specific type of stakeholder.

Once again, the effect of every one of the treatments (P+B, EM+B and B) was analysed on the measures considered (IRE, IRI and IRIE), using the non-parametric Kruskal Wallis test with the three groups A, B and C, and the non-parametric Mann-Whitney U test taking pairs of groups A and B, A and C, and B and C. Table 5 shows the results obtained for each measure employed in the Kruskal-Wallis and in the Mann-Whitney U tests.

Table 5. Kruskal-Wallis and Mann-Whitney U test results for IRI, IRE and IRIE

Group	**Test**	**Variable**	**P-value**	**OP**	**ES** ($Mean_1 - Mean_2$/ Standard error)	**R**
A / B / C	Kruskal-Wallis	IRI	0.081	< 0.744	0.418	NO
		IRE	0.283	< 0.103	0.210	NO
		IRIE	0.225	< 0.335	0.298	NO
A / B	Mann-Whitney U	IRI	0.579	0.067	-0.167	NO
		IRE	0.271	0.074	-0.332	NO
		IRIE	0.437	0.05	-0.259	NO
A / C	Mann-Whitney U	IRI	**0.053**	0.839	-0.732	**NO***
		IRE	0.430	0.1	-0.298	NO
		IRIE	**0.076**	0.966	-0.795	**NO***
B / C	Mann-Whitney U	IRI	**0.044**	0.703	-0.711	**YES**
		IRE	0.182	0.101	-0.471	NO
		IRIE	0.182	0.318	-0.471	NO

The results obtained allow to reject Groups B / C, variable IRI, given that the p-value is 0.044, which is lower than 0.05., i.e., the Empathy Maps had no effect on IRI. For Groups A / C, even when the p-value was higher than 0.05 for all the variable, in the case of IRI with a p-value = 0.053, and OP of 0.839 and IRIE, with a p-value of = 0.076 and op of 0.966, the result and the Observed Power allows to evaluate a correlation between the treatment and the result achieved.

Summarizing, statistical analysis shows the influence of the Empathy Maps technique on the number of identified baseline stakeholders, and a slight influence of Personas. So, the contribution of the techniques to increase the identified stakeholders of the ideas generated, is positive. Likewise, comparing these values with the number of ideas generated, it could be concluded, with some caution, that the number of identified stakeholders influence positively in the number and content of the ideas generated.

In Groups A (P+B treatment) and B (EM+B treatment), the number of ideas around internal stakeholders was equal to or greater than the external ones, which could be explained as a contribution of the technique to empathise with these stakeholders. Therefore, the Empathy Maps and Personas techniques would seem to facilitate empathy with stakeholders in domains in which they are not experts, such as veterinarian or employee, which would then be reflected in a greater flow of ideas. This is an important result because it confirms the contribution of the techniques to help subjects to empathise. In the case of Groups C (Brainstorming treatment), there are more ideas around external stakeholders. This was an expected result, as usually people tend to find it easier to put themselves in the place of these types of stakeholders (member, adopter, or volunteer) than the internal ones (veterinarian, employee, manager, etc.). On the other hand, the number of stakeholders that simultaneously belonged to both stakeholders (IRIE variable) were, in the case of Groups B, 3.9 times greater than Groups A, and 2.27 times greater than Groups C, which shows that the ideas generated by applying Brainstorming only are, in terms of stakeholders, much more generic.

5 Threats to validity

Certain issues which may have threatened the validity of the quasi-experiment must be considered [21]:

- External validity may be threatened when experiments are performed with students, as doubts have been raised regarding the representativeness of the subjects with respect to software professionals. Despite this, the tasks to be performed did not require real world experience, and we believe, therefore, that this quasi-experiment could be considered appropriate, as suggested in the literature [15]. Even that two different techniques were used and there was a third control group, the execution of a single case study could limit the scope of the conclusions. In the future, the experiment could be replicated again, incorporating additional case studies from other domains to compare if there is an effect between the techniques and the domains. The possibility of contamination between groups, whereby students in one group may have shared information with those in another, may be considered a threat to the validity of the study. However, we made special efforts to ensure that this did not occur. Even so, if it had occurred, the effect of prior knowledge on the brainstorming process is likely to be attenuated because it were conducted in groups, so the influence of any one participant was diluted.
- Threats to internal validity are to some extent mitigated by the design of the quasi-experiment. In our case, both the support materials and the exercise were the same for all the groups, but an additional technique was presented to Group A (P+B

treatment) and Group B (EM+B treatment). Due to time constraints some parts of the experiment tasks were completed *a posteriori*, outside of the controlled environment. Although this has occurred in all groups, i.e., for all treatments, it is an aspect that may have affected the results. We will therefore take it into account in future replications and experiments.

- Conclusion validity concerns the data collection, the reliability of the measurement, and the validity of the statistical tests. Statistical tests appropriate to the type of measures of the dependent variables have been used to test the hypotheses. It has been explicitly mentioned and discussed whenever non-significant differences were found to be present. It is also necessary to state that the conclusion validity could also be affected by the number of observations. Further replications with larger datasets are, therefore, required to confirm or contradict the results shown herein.
- Construct validity may be influenced by the measures used to attain a quantitative evaluation of the ideas generated, the comprehension of the techniques explained, and the experimental tasks. The number of ideas of requirements were measure, to avoid any subjectivity as regards the way in which they were written. Since participants were asked to generate ideas of requirements that still need to be negotiated and validated with the clients, we paper understand that it is an interesting and valuable result for the requirements engineer since the expected result of the use of the Brainstorming technique as an elicitation tool is the generation of a large flow of relevant ideas of requirements but not necessarily high-quality ideas of requirements, which may be performed during the negotiation and validation of the requirements of the software system to be developed.

6 Conclusions

This paper presents the results of a quasi-experiment carried out to evaluate the improvement in the *Effectiveness*, measured as quantity of ideas of requirements generated by the students, and *Identified Stakeholders*, measured as quantity of different stakeholders identified by the students, of Brainstorming sessions when are complemented with Empathy Maps or Personas techniques. The quasi-experiment was carried out with 74 undergraduate students enrolled on a Bachelor's degree in Computer Engineering at the Universitat Politècnica de València in October 2021.

The main findings obtained are the following: (1) The descriptive analysis reveals an increase in number of ideas of requirements generated in Brainstorming sessions when using Empathy Maps and Personas techniques even when there is no statistically significant difference. Moreover, when using Empathy Maps, there is an increase in the number of functional and a reduction in that of non-functional ideas of requirements, which evidences that the technique influences the type of idea identified, as was already stated in previous studies. This result indicates that it may be useful to complement Brainstorming sessions with the use of DT-based techniques. (2) The use of a US template seems to contribute to the generation of a greater number and more precise ideas of requirements. This finding emerges from comparing the number of ideas in this quasi-experiment with the results obtained in our previous study. (3) The number of

stakeholders identified, according to the baseline defined by the authors of this paper, is greater when applying Personas and Empathy Maps as opposed to using Brainstorm-ing alone, with a statistically significant difference in favour of Empathy Maps. This result must be evaluated with caution, since the treatment groups, particularly group C (Brainstorming Treatment), were very small. (4) When analyzing the distribution of ideas among internal, external, and internal and external stakeholders, we found a sta-tistically significant difference when applying Empathy Maps on the ideas among in-ternal stakeholders' variable. When applying Personas on the internal stakeholder and internal and external stakeholders' variables, even when there is no statistical signifi-cance, the p-value and observed power allows to evaluate a strong correlation between the treatment. This would indicate that the techniques favour the identification of stake-holders in domains in which they are not experts. (5) It would not seem to be worth using both techniques together, but rather to use either Empathy Maps or Personas. However, it could not be concluded whether Empathy Maps or Personas is better in terms of *Effectiveness or Identified stakeholders* of the ideas of requirements. These results may be useful to practitioners as well as to RE and Software Engineering lec-turers since there are various techniques available for requirements elicitation but very little evidence about how to combine them to improve the quantity and quality of re-quirements obtained.

As future work we plan to replicate this quasi-experiment to corroborate the findings and to obtain more conclusive results. We will consider using experimental objects re-lated to different domains and larger samples. Also, we will be considering the incor-poration of additional dependent variables, which allow a better evaluation of the qual-ity of the validated requirements ideas, such as a measure of correctness. In addition, we want to explore whether there is any significative difference between Empathy Maps and Personas and to identify under which circumstances (e.g., type of problem domain, number or type of stakeholders, team experience, team composition) these dif-ferences might appear.

Acknowledgment

The research presented in this paper is part of the following projects: ADAGIO (Consejería de Educación, Cultura y Deportes de la JCCM, SBLPY/21/180501/000061), AETHER-UCLM (MICINN, PID2020-112540RB-C42) y Proyecto de investigación en Procesos de desarrollo de Software (línea de investigación en Ingeniería de requisitos, UNTREF, Argentina).

References

1. Vetterli, C., Brenner, W., Uebernickel, F., Petrie, C.: From Palaces to Yurts: Why Require-ments Engineering Needs Design Thinking. IEEE Internet Computing. 17, 91–94 (2013).
2. Ferreira Martins, H., Carvalho de Oliveira Junior, A., Dias Canedo, E., Dias Kosloski, R.A., Ávila Paldês, R., Costa Oliveira, E.: Design Thinking: Challenges for Software Require-ments Elicitation. Information. 10, 371 (2019).

3. Hehn, J., Uebernickel, F., Mendez Fernandez, D.: DT4RE: Design Thinking for Requirements Engineering: A Tutorial on Human-Centered and Structured Requirements Elicitation. In: 2018 IEEE 26th International Requirements Engineering Conference (RE). pp. 504–505. IEEE, Banff, AB (2018).
4. Kahan, E., Genero, M., Oliveros, A.: Challenges in Requirement Engineering: Could Design Thinking Help? In: Piattini, M., Rupino da Cunha, P., García Rodríguez de Guzmán, I., and Pérez-Castillo, R. (eds.) Quality of Information and Communications Technology. pp. 79–86. Springer International Publishing, Cham (2019).
5. Kahan, Ezequiel, Isfrán, Emilio, Genero, Marcela, Oliveros, Alejandro: Studying the Influ-ence of Empathy Maps on Brainstorming for Requirements Elicitation: A Quasi-Experi-ment, (2021).
6. Hehn, J., Uebernickel, F., Stoeckli, E., Brenner, W.: Designing Human-Centric Information Systems: Towards an Understanding of Challenges in Specifying Requirements within De-sign Thinking Projects. 12 (2018).
7. Pacheco, C., García, I., Reyes, M.: Requirements elicitation techniques: a systematic liter-ature review based on the maturity of the techniques. IET Software. 12, 365–378 (2018).
8. Gasparini, A.: Perspective and Use of Empathy in Design Thinking. ACHI 2015 : The Eighth International Conference on Advances in Computer-Human Interactions. 49–54 (2015).
9. Hoda, R., Salleh, N., Grundy, J.: The Rise and Evolution of Agile Software Development. IEEE Softw. 35, 58–63 (2018).
10. Dalpiaz, F., Brinkkemper, S.: Agile Requirements Engineering with User Stories. In: 2018 IEEE 26th International Requirements Engineering Conference (RE). pp. 506–507. IEEE, Banff, AB (2018).
11. Lucassen, G., Dalpiaz, F., van der Werf, J.M.E.M., Brinkkemper, S.: Improving agile re-quirements: the Quality User Story framework and tool. Requirements Eng. 21, 383–403 (2016).
12. Ferreira Martins, H., Carvalho de Oliveira Junior, A., Dias Canedo, E., Dias Kosloski, R.A., Ávila Paldês, R., Costa Oliveira, E.: Design Thinking: Challenges for Software Require-ments Elicitation. Information. 10, 371 (2019). .
13. Meireles, M., Souza, A., Conte, T., Maldonado, J.: Organizing the Design Thinking Toolbox: Supporting the Requirements Elicitation Decision Making. Presented at the ACM International Conference Proceeding Series (2021).
14. Teixeira, G., Zaina, L.: Using Lean Personas to the Description of UX-related Require-ments: A Study with Software Startup Professionals. Presented at the International Confer-ence on Enterprise Information Systems April 25 (2022).
15. Costa Valentim, N.M., Silva, W., Conte, T.: The Students' Perspectives on Applying De-sign Thinking for the Design of Mobile Applications. In: 2017 IEEE/ACM 39th Interna-tional Conference on Software Engineering: Software Engineering Education and Training Track (ICSE-SEET). pp. 77–86. IEEE, Buenos Aires (2017).
16. Ferreira, B., Silva, W., Barbosa, S.D.J., Conte, T.: Technique for representing requirements using personas: a controlled experiment. IET Software. 12, 280–290 (2018).
17. Kahan, E., Isfrán, E., Genero, M., Oliveros, A.: Experimental Material WER 2023. Availa-ble online: https://tinyurl.com/2zv8ek2s
18. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimenta-tion in Software Engineering: An Introduction. Springer US (2000).
19. Basili, V., Rombach, H.D.: Towards a comprehensive framework for reuse: A reuse-ena-bling software evolution environment. undefined. (1988).
20. Sommerville, I.: Software engineering. Addison-Wesley Pub. Co., Wokingham, England; Reading, Mass. (1992).
21. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimenta-tion in Software Engineering. Springer (2012).