On the impact of using different templates on creating and understanding user stories

Catarina Gralha, Rita Pereira[1], Miguel Goulão, João Araujo
NOVA LINCS, Department of Computer Science
Faculty of Science and Technology, Universidade NOVA de Lisboa
{catarina.gralha, mgoul, joao.araujo}@fct.unl.pt
[1]r.casmarrinha@campus.fct.unl.pt

***Abstract*—Context: User stories are often used for elicitation and prioritisation of requirements. However, the lack of a widely adopted user story template, covering benefit and the usage (or not) of a persona, can affect user stories' quality, leading to ambiguity, lack of completeness, or accidental complexity. Objectives: Our goal was to analyse the differences between 4 alternative user story templates when creating and understanding user stories. Methods: We conducted a quasi-experiment. We asked 41 participants to perform creation and understanding tasks with the user story templates. We measured their accuracy, using metrics of task success; their speed, with task duration; visual effort, collected with an eye-tracker; and participants' perceived effort, evaluated with NASA-TLX. Results: Regarding the impact of the different templates in creating user stories, we observed statistically significant differences in some of the metrics for accuracy, speed and visual effort. For understanding user stories, we observed small differences in terms of visual effort. Conclusions: Although some templates outperformed others in a few metrics, no template obtained the best overall result. As such, we found no compelling evidence that one template is "better" than the others.**

***Index Terms*—user story templates, eye-tracking, empirical evaluation**

I. Introduction

Requirements Engineering is the process of requirements elicitation, specification, validation and management. This process includes preparing and maintaining requirements documents [1]. In particular, requirements elicitation can be performed using various techniques, from graphical to textual models. These models help communication between the development team and other stakeholders. A survey conducted under the umbrella of International Software Engineering Research Network (ISERN), with respondents from industry from over 30 countries, identified *user stories* as one of the most often used requirements documentation techniques [2].

A user story is a short and simple description of a system feature, written in natural language, with value to the user or its owner [3]. There are templates to facilitate the understanding of user stories and avoid ambiguity. User story templates vary in their elements and presentation order, which may impact the quality of user stories and the creation and understanding tasks performed with those templates.

In this paper, our goal is to report the impact of four alternative user stories templates, when stakeholders are performing creation and understanding tasks. The first template is called Connextra [3] (we refer to it as CON) and it is written as *"As a (type of user), I want (goal), so that (benefit)"*. A second template is similar to CON but puts the *"benefit"* element first [4]. We called it **BEN**. The last two templates are variants of CON and BEN, where the *"type of user"* is replaced by a *persona*, and we refer to them as **PER** and **PERBEN**, respectively.

A total of 41 participants performed creation and understanding tasks with the user stories templates. We measured their *accuracy*, *speed*, *visual effort* and *perceived effort* to accomplish their tasks, by collecting metrics of success, time, biometric data with an eye-tracker, and perceived effort using the NASA-TLX questionnaire.

The paper is structured as follows. In Section II we present the user stories templates, and the eye-tracking and perceived effort. In Section III we discuss related work. In Section IV we report the experiment planning and in Section V we describe the experiment execution. In Section VI we present the results, which we discuss in Section VII. Finally, in Section VIII we draw conclusions and point directions for future work.

II. Background

A. User stories templates

As the Agile paradigm [5] promotes communication between the development team and other stakeholders, user stories are used to define and prioritise requirements, and they are written in natural language. Although it is expressive, intuitive, and universal (assuming stakeholders speak the same language), natural language is also potentially vague and prone to omissions and ambiguity. User stories templates avoid ambiguity and facilitate understanding in user stories.

Different templates vary in the elements' inclusion and ordering in the user story. In this paper, we will evaluate the following four user stories templates:

- **CON: "As a <*type of user*>, I want <*goal*>, so that <*benefit*>".**
  The most commonly often used template is Connextra (CON) [3], [6]. E.g., "As a *user*, I want to *access my account*, so that *I can manage my money*.".
- **BEN: "In order to <*benefit*>, as a <*type of user*>, I want <*goal*>".**
  Although similar to CON, BEN template puts the

<*benefit*> element first. One possible argument for this change in the order, is that the reader could focus more on the problem and on the importance of its resolution, and not on who needs the functionality [4]. E.g., "In order to *manage my money*, as a *user*, I want to *access my account*.".

- **PER: "As <persona>, I want <goal>, so that <benefit>".**
  In this template, the <*type of user*> is replaced by a <*persona*>. A persona is a fictitious identity with objectives, attitudes and motivations. The characteristics of the persona help to understand the features that real users and customers might want from the software system [3], [7], [8]. E.g., "As *Yasmin*, I want to *access my account*, so that *I can manage my money*.".
- **PERBEN: "In order to <benefit>, as <persona>, I want <goal>".**
  This template is a union of BEN and PER templates, in which a <*persona*> is used to represent the <*type of user*> that wants this functionality in the system or product, and where the <*benefit*> element is placed first. E.g., "In order to *manage my money*, as *Yasmin*, I want to *access my account*.".

B. Eye-tracking and perceived effort

To gain a multi-perspective on how participants interact with user stories templates, we use a combination of techniques to collect: *i*) direct task performance metrics; and *ii*) indirect measures like the effort while performing the tasks, assessed with an eye-tracker, and the participants' perception on their effort, measured with NASA-TLX.

Eye-tracking. It is a technology that measures the activity of the eyes. Eye-movements are essential to collect evidence regarding participants' cognitive processes [9], [10]. Eye-trackers monitor a participant's visual attention by collecting eye-movement data when (s)he looks at a stimulus while working on a specific task. A stimulus is an object, such as text, source code, or diagram, that is necessary to perform that task [11]. Beyond the analysis of visual attention and cognitive processes, eye-data can also be examined to measure the workload of a task (i.e., visual effort) [12]. Plus, the data can be studied for certain areas of the stimuli, which are called areas of interest (AOI). An AOI can either be relevant or irrelevant for a task that is being performed by a participant.

Perceived effort. Cognitive load is a multi-dimensional construct representing the load that a task imposes on a participant [13]. This also refers to the level of *perceived effort* for learning, thinking and reasoning as an indicator of pressure on working memory during task execution [14]. This measure of mental workload represents the interaction between task processing demands and human capabilities or resources. The NASA-Task Load Index (NASA-TLX) [15] is a technique for subjective workload and cognitive load assessment. It uses 6 (six) dimensions to assess workload and cognitive load: mental, physical, and temporal demand, performance, effort, and frustration. The evaluation process consists of a bipolar scale of 20 steps for each of the dimensions, followed by a comparison between the dimensions. This comparison is made by showing, in turn, all possible pairs of dimensions to the participant. He chooses which, in his opinion, is the most relevant of each pair. In the end, the effort spent on the task is calculated based on the result calculated for each dimension, through the scale, and on the weight of each dimension, based on the frequency with which it was considered the most relevant of the pair. Through the results of these 6 dimensions, an overall score is also calculated.

III. Related work

User stories have been the subject of several research. Lucassen et al. [6] analysed if participants use guidelines or templates when creating user stories. They found that 59% of the participants use Connextra and 10% use one of its variants. Another study focused on the substitution of the user for a persona, in a business context, when writing user stories [16]. There were also studies on user stories' granularity [17] and completeness [18], showing it's the importance of each user story being complete, not ambiguous, and showing the stakeholders' needs.

Dalpiaz et al. [19] compared two approaches used to identify ambiguity in user stories. They distributed 57 students into 28 groups to assess the precision and recall of manual inspection (using only pen and paper) versus REVV-Light, an open-source Web 2.0 tool that uses natural language processing and information visualisation to help requirements engineers to identify ambiguity in user stories. The results reveal that manual inspection gives a statistically significant higher recall.

Dalpiaz and Sturm [20] investigated the suitability of use cases and user stories for the manual derivation of a conceptual model, studying the completeness and correctness of such derivation. The experiment consisted of a two-factor, two-treatment controlled experiment with 118 participants. The results reveal that user stories perform better than use cases for deriving conceptual models, probably due to their conciseness and repetitions.

Liskin et al. [21] performed a study with 72 participants that analysed the duration of the implementation of user stories. The objective was to verify that smaller user stories are clearer and give rise to fewer problems than larger user stories. By dividing complex user stories into several smaller ones, some participants encountered problems with clarity and dependencies between user stories. The final results show that different tasks require different granularities, but two concise user stories obtain better results than long user stories equivalent to their composition.

Lin et al. [22] collected and analysed 142 user stories from a project at Beihang University. They found that most of them do not describe the needs of stakeholders and users, but are more focused on the objectives of the programmers. The authors propose a hierarchical model of objectives that facilitate the understanding of stakeholder requirements and avoid problems of ambiguity. This model was implemented within an educational context in order to allow a comparison

with the user stories resulting from the approach currently used in companies. The results showed a significant improvement both in the number of user stories and in their quality.

In terms of eye-tracker devices, they have been used in several studies, to observe how people understand graphic and textual models. In particular, Sharafi et al. [23] compared the efficiency of graphical and textual representations in requirements comprehension, and evaluated the differences in terms of visual effort, duration and precision. The results show that, despite being the preference of most participants, the graphic representation required more visual effort and more time than the textual one.

Regarding graphic models, some authors have investigated patterns in the understanding of UML class diagrams [24], as well as the impact of stereotypes and colours [25]. There were also studies analysing the use of patterns in class diagrams [26]–[28]. Some studies investigated the impact of bad layouts, semantics and different problem-solving styles in different tasks with *i** and iStar 2.0 models [29]–[31].

There are some studies that investigated patterns in reading and understanding source code, focusing on beacons [32], dwelling time [33] and automated code summarization [34]. Katona compared the visual effort in reading semantically equivalent *"clean" vs. "dirty"* source code and found statistically significant evidence that *"dirty"* source code readers had a higher number of fixations, their mean duration was longer, and the mean fixation connections length (that is, saccade length) was higher [35]. All these comparisons suggested a higher visual effort when reading the so-called *"dirty"* code, and were in line with the self-reported assessment of participants, that found *"clean"* code significantly easier to read and understand, as well as more precise, than *"dirty"* code.

There are, however, significant differences in reading and understanding source code and natural language text [36], [37]. User stories templates, due to their structure, may lie somewhere in between source code and natural language text.

All these works bring important insights on the use of user stories covering different aspects, or on the usage of eye-tracking devices to evaluated the visual effort of participants when performing different tasks (over graphics and text). However, none of them analyses the impact of the different templates for user stories, regarding their creation and understanding, considering accuracy, speed, visual effort and perceived effort.

IV. Experiment planning

A. Goals

We describe our research goals following the GQM research goals template [38]. Our first goal (**G1**) is to *analyse* the difference in user story templates, *for the purpose of* evaluation, *with respect to* their effects on the **creation** of user stories, *from the viewpoint of* researchers, *in the context of* an experiment conducted with students and practitioners. Our second goal (**G2**) can be obtained by replacing the term *creation* with *understanding*.

We can break down each goal into four sub-goals, concerning the effect(s) of the different user story templates, in terms of *accuracy*, *speed*, *visual effort* and *perceived effort*. The refined goals can be obtained by replacing the terms *creation* (or *understanding*) with *accuracy to create*, *speed to create*, *visual effort to create*, *perceived effort to create*.

B. Participants

We recruited 41 participants through convenience and snowball sampling. From those, 10 participants performed tasks with CON, 10 with BEN, 10 with PER, and 11 with PERBEN. We leveraged personal contacts and participants were made aware of the study either by direct communication or by e-mail. Some of these participants actively recruited their contacts to participate, hence the snowball sampling. This allowed us to have a more diversified set of participants.

In total, 26 participants were familiar with user stories, and have used them in the context of a course or in a professional setting. The remaining 15 had no *previous experience* with them. Figure 1a presents a detailed characterisation, divided by user story template. Regarding *occupation*, 19 are students, 4 are working-students, 11 are practitioners, 5 are researchers, and 2 are unemployed. Figure 1b presents a detailed characterisation, divided by user story template.

C. Experimental materials

The experimental materials for this evaluation included *i*) a participant consent form; *ii*) a video tutorial about user stories; *iii*) two tasks to be performed; *iv*) a NASA-TLX questionnaire; and *v*) a demographic questionnaire. **All the materials can be found in Zenodo [39]**.

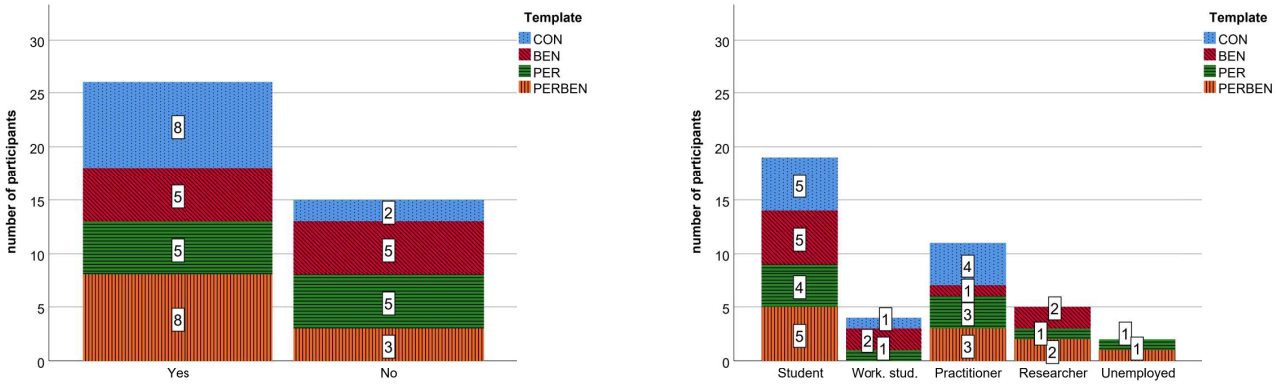
The *participant consent form*, adapted from [40], explained that the participation was entirely voluntary, the participants could refuse to answer any question, and could leave the experiment at any time, and that all the collected data would remain anonymous.

A 2 minutes *video tutorial* introduced user stories to participants. The video described one of the four templates, matching the template the participant would interact with, in the subsequent tasks. The tutorial included visual and audio explanation on how user stories are used, how to create users stories, as well as some examples of user stories with the template used on the tasks. Participants had no control over the video, not being able to pause it or resume it, since having different viewing times and going through specific parts of the tutorial more than once could impact the results.

The *creation and modification tasks* were performed in a custom web-based tool developed by the authors.

The *NASA-TLX questionnaire* collected the participants' perceived effort.

The *demographic questionnaire* collected demographic information on the participants. Additionally, we also needed to know if the participant wore glasses or contact lenses, since it could impact the usage of the eye-tracker, and the extent to which the participant had previous experience with user stories.



(a) Previous experience with user stories.

(b) Current occupation.

Fig. 1: Participants' demographic information.

D. Tasks

There are 4 user story templates and, for each, 1 creation and 1 understanding task. For the creation tasks, the domain was a *booking system for a hotel*. For the understanding task, the domain was a *website for content sharing*. We opted for relatively known domains to reduce the effect of the results being related to difficulties in understanding the domain itself, and not due to the user stories templates under study. However, we are aware that tacit knowledge may play an important role in the performance of the participants. Each participant performed one creation task and one understanding task, with the same template.

In the *understanding* task, participants using CON and BEN had to choose the user stories that best matched the given scenario. There were 7 options, and 4 of those where considered the correct answer. For PER and PERBEN, participants had access to two personas descriptions and 7 user stories. They then had to choose which user stories best suited the needs of each persona.

For each *creation* task, in the ones using CON and BEN, participants had to write user stories based on a given scenario. For PER and PERBEN, the task was similar, but we presented the same scenario and a persona.

The distribution of participants to tasks and templates was random, and the number of participants for each template was balanced.

E. Hypotheses and variables

For each one of the high-level goals presented in Subsection IV-A, we define a null hypothesis (H_0) and the alternative hypothesis (H_1). For G1, concerning the *creation* task, we have the following hypotheses:

> $H_{0Create}$: Different user story templates **do not** impact the *creation* of user stories.
> $H_{1Create}$: Different user stories templates impact the *creation* of user stories.

These hypotheses are further refined to cope with *accuracy*, *speed*, *visual effort* and *perceived effort* of creation. For example, for accuracy:

> $H_{0CreateAcc}$: Different user story templates **do not** impact the *accuracy to create* user stories.
> $H_{1CreateAcc}$: Different user stories templates impact the *accuracy to create* user stories.

We follow the same approach to define hypotheses for G2, concerning the *understanding* task, respectively. These hypotheses are also further refined to cope with *accuracy*, *speed*, *visual effort* and *perceived effort* of understading.

The **independent** variable is the user story template. There are four options for the treatments, which may be *CON*, *BEN*, *PER* OR *PERBEN*. In Table I, we present an overview of this variable. The first column presents its name, and the second column has the scale type. The last column has the options for the values, that is, the treatments, which may be CON, BEN, PER or PERBEN. The **dependent** variables are *accuracy*, *speed*, *visual effort* and *perceived effort*. For each of these variables, there is a set of metrics. From Table II to V, we present an overview of these metrics. The first column shows the name of the variable. The second column presents the range, and the last column has the counting rule or formula for the metric calculation.

TABLE I: Overview of the *independent* variable.

Name	Scale	Values
User story template	nominal	{CON; BEN; PER; PERBEN}

Assessing accuracy. In Table II we present the metrics for the dependent variable *accuracy*. It is evaluated using *precision*, *recall* and *f-measure*. Higher values of these metrics support the claim of a better accuracy when using the corresponding template for creating or understanding user stories.

Assessing speed. In Table III we present the metric for the dependent variable *speed*. The unit for this metrics is the *second*. Lower values of *duration* correspond to better speed,

TABLE II: Overview of the metrics for the *dependent* variable *accuracy*.

Name	Range	Counting rule
Precision	$0 \leq x \leq 1$	$\frac{number\ of\ correct\ answers\ provided}{total\ number\ of\ answers\ provided}$
Recall	$0 \leq x \leq 1$	$\frac{number\ of\ correct\ answers\ provided}{total\ number\ of\ correct\ answers}$
F-measure	$0 \leq x \leq 1$	$\frac{2 * (precision * recall)}{(precision + recall)}$

indicating that the corresponding template help in improving the speed with which the user stories are created or understood.

TABLE III: Overview of the metric for the *dependent* variable *speed*.

Name	Range	Counting rule
Duration	$0 \leq x$	$completion\ time - start\ time$

Assessing visual effort. In Table IV we present the metrics for the dependent variable *visual effort*, collected with the eye-tracker. A fixation is a stabilisation of the eye on a part of the stimulus for a period of time between 200 and 300 ms. A higher *number* and *duration* of *fixations* is associated with a higher visual attention in a given set of AOI. Those areas of interest can be relevant or irrelevant, depending if the area contains a correct answer for the task [10], [41]. Regarding the *average duration of fixation*, a higher value indicates more time and attention devoted to AOI [10], [42], which is correlated with cognitive processes [11]. A saccade is a sudden and quick eye-movement lasting between 40 to 50 ms. A higher *number of saccades* can be associated with a higher visual effort, meaning the participant may be somewhat "lost", making a more erratic navigation [10], [12].

Assessing perceived effort. In Table V we present the metrics for the dependent variable *perceived effort*, assessed through the NASA-TLX questionnaire. It has 6 metrics to assess workload and cognitive load: mental, physical, and temporal demand, performance, effort, and frustration. Higher values, in all the metrics, correspond to a greater perceived effort by the participant. Each metric is weighted, in terms of its importance for the overall effort. The denominator *15* corresponds to the 15 paired comparisons of all the 6 dimensions to access the perceived workload [43].

F. Experimental design

We follow an *experimental design*, since the allocation of participants to the user story template was random. If a participant performed the tasks with a given template, the next one would be allocated to a different template, so that the number of participants using each template would be balanced.

We have a *between-subjects design*. Every participant is subjected to a single treatment, i.e., only performs the tasks with one of the user story templates. We opted for the between-subjects design for 3 main reasons: *i)* time; *ii)* fatigue; and *iii)* the learning effect. In particular, in studies with practitioners, time is a decisive factor. An experiment with multiple tasks with various templates may increase the mortality of the participants, or discourage them from participating, in the first place. Moreover, in a long experiment, the participants may become tired. This could decrease their performance on the last templates. Alternatively, the learning effect may cause them to improve their performance over the course of the studies. Finally, a crossover design, where every participant is subjected to more than one treatment, is complex [45] and it is discouraged based on the risk of performing an incorrect analysis [46].

V. Execution

A. Preparation

The experiment was performed with a laptop connected to an external 22 inch, wide screen, full HD monitor; a The Eye Tribe eye-tracker [47]; and an external mouse and keyboard. We prepared the session on the laptop, and the participant had access to the monitor, mouse and keyboard. Participants sat on a chair without wheels, to avoid movement that could jeopardise the eye-tracker data.

We prepared each session with identical conditions for all participants. The room was only being used for the experiment and there was no interruption while the participant was performing the tasks. Only one session was held at a time. All sessions were scheduled according to participant's availability.

B. Procedure

When a participant arrived, we started by explaining that we were evaluating a textual language, not the participant himself. We informed that (s)he would watch a tutorial video, and perform two tasks. After each task, (s)he would fill a NASA-TLX questionnaire. We further informed the participant that (s)he would answer a demographic questionnaire. We explained that all collected data were anonymous and (s)he would control the entire session. We also explained (s)he could quit at any moment, and there was no time limit for performing the task. Finally, we asked if the participant had any questions, and informed we could not answer questions during the experiment.

We then helped the participant seat comfortably and adjusted the eye-tracker to the participant's eye level. The eye-tracker was placed below the screen, without blocking it. Then, we calibrated the eye-tracker using its software and only accepted *good* or *excellent* calibrations (top levels of a 5 points ordinal scale).

We started the video and audio recording and let the participant begin the session. The participant had control over the entire session. After watching video tutorial, (s)he would click on a continue button and the first task would appear. When the participant felt the task was completed, (s)he clicked on another continue button and the NASA-TLX would appear. The procedure was the same for the second task, the second NASA-TLX, and the demographic questionnaire. In the end, we thanked the participant for being part of the evaluation and answered any questions (s)he might have.

TABLE IV: Metrics for the *dependent* variable *visual effort*: eye-tracking

Name	Abbreviation	Range	Counting rule
Fixation rate on relevant elements	FixRel	$0 \leq x$	$\frac{number\ of\ fixations\ on\ the\ relevant\ AOI}{number\ of\ fixations\ on\ the\ AOG}$
Fixation rate on irrelevant elements	FixIrrel	$0 \leq x$	$\frac{number\ of\ fixations\ on\ the\ irrelevant\ AOI}{number\ of\ fixations\ on\ the\ AOG}$
Average duration relevant fixations	AvgFixRel	$0 \leq x$	$\frac{\Sigma\ duration\ of\ fixations\ on\ the\ relevant\ AOI}{number\ of\ fixations\ on\ the\ relevant\ AOI}$
Average duration irrelevant fixations	AvgFixIrrel	$0 \leq x$	$\frac{\Sigma\ duration\ of\ fixations\ on\ the\ irrelevant\ AOI}{number\ of\ fixations\ on\ the\ irrelevant\ AOI}$
Total number of saccades	Saccades	$0 \leq x$	$\Sigma\ saccades$

TABLE V: Overview of the metrics for the *dependent* variable *perceived effort* [44].

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

C. Deviations from the plan

We detected non preventable flaws in the equipment, which led to the exclusion of some data from the evaluation. While performing an understanding task with PER, the audio recording was interrupted once, resulting in the exclusion of 1 participant speed data and answer to the task. During another task, also with PER, the video recording was interrupted once, resulting in the exclusion of speed and visual effort data for that participant. While performing understanding tasks with PERBEN template, the audio recording stopped working twice, resulting in the exclusion of the speed data of 2 participants. Altogether, in tasks with PERBEN, the eye-tracker stopped working 3 times, resulting in the exclusion of eye-tracking data from 3 participants in the understanding tasks, and 3 participants in the creation tasks. While performing tasks with BEN template, there was one session where the eye-tracker did not record the screen coordinates. Thus, the visual data of this participant were also excluded.

VI. Analysis

A. Data set preparation

For the *creation* task, the answers were written in a text box, and saved in a CSV file. We had a gold-standard user story set, created and validated by experienced user story researchers. These user stories were iterated and changed based on the data analysis, in the case a participant adding a particular user story that was useful and not covered by the initial goal-standard. In the end, all the user stories created by the participants were evaluated based on this gold-standard. In this context, we can regard the user stories in the gold-standard as a closed set, with a complete list of correct user stories. When assessing each user stories set, we counted which of the final goal-standard list were suitably described by the participants. All user stories provided by participants that were not found the final gold-standard were counted as false positives. These data allow us to analyse the participants' *accuracy*.

In the *understanding* tasks, the participant answered out loud and the answers were recorded. For processing the data of the understanding tasks, we created a table with all options and, when listening to the audio, each answer was manually marked as correct or incorrect. The understanding tasks had multiple choice questions, so the participant did not give answers that were not on the list provided. These data allow us to analyse the participants' *accuracy*.

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

For the eye-data, the areas of the stimulus and its elements were mapped into pixel coordinates, and saved in a CSV file. This enabled tagging the eye-tracking data with the elements being gazed at any given moment. The fixations and corresponding durations were saved in another CSV, to compute the normalised fixation durations. These data allow us to analyse the participants' *visual effort*.

The demographic questionnaire and the NASA-TLX questionnaire answers were saved into CSV files. Those files had the structure needed to perform the analysis on the participant's *demographic data* and *perceived effort*.

B. Analysis procedure

We started by collecting descriptive statistics on our variables, to get an overview of their distribution. We collected the *mean*, *standard deviation*, *skewness*, and *kurtosis*. We also used *box plots* and *Q-Q plots* (omitted here for the sake of brevity), to help with the visual analysis of the distributions, in combination with the Shapiro-Wilk normality test.

We then applied the *Levene's test* for homogeneity of variance to assess if each group of the independent variable had the same variance. If the Levene statistic is significant at the $p < 0.05$ level, we reject the null hypothesis that the groups have equal variances.

For testing our hypotheses, we used the *Welch's t-test*. A discussion on the benefits of using Welch *t*-test [48] for comparing distributions to detect statistically significant differences in a robust way (as opposed to one-way ANOVA, or a non-parametric alternative, such as Kruskal-Wallis H Test) is in [49], [50]. We are using $p < 0.05$ for the level of significance and thus rejecting the null hypothesis.

With more than two groups, Welch's t-test does not inform us which groups are different from the others, only that a difference exists. After finding a significant difference, we need to apply a post-hoc test on the factor to examine the differences between the user story templates. We used the *Games-Howell* post-hoc procedure, which is robust for unequal variances in the groups. We use $p < 0.05$ for the level of significance and rejecting the null hypothesis.

C. Descriptive statistics

In Table VI we present the descriptive statistics for the metrics collected. For the sake of brevity, we only present the results for the *creation* task, and for the dependent variables *accuracy*, *speed* and *visual effort*. The metrics are collected for each of the 4 user stories templates (BEN, CON, PER, or PERBEN). The descriptive statistics include **Mean**, **S**tandard **D**eviation, **Skew**ness, and **Kurt**osis. We also include the *p*-value for the **S**hapiro-**W**ilk normality test. The shape of the distributions suggests that, in some of the cases, normality is **not** a reasonable assumption ($p < 0.05$). These distributions are highlighted in **bold**.

D. Hypotheses testing

For the sake of brevity, we only presented the results for the hypotheses testing that are statistically significant.

RQ1: *Do different user story templates have an impact on the creation of user stories?*

In Table VII we present Levene's and Welch's test results for the *creation* task, for metrics with a statistically significant difference in the Welch *t*-test.

In Table VIII we summarise the Games-Howell post-hoc test results, as well as present the mean and standard deviation for the different templates, for all the metrics that had a statistically significant difference.

Assessing accuracy. There was a statistically significant difference among the templates in terms of accuracy, for *recall* and *f-measure*. However, the Games-Howell post-hoc test could not identify those differences. Since participants' sample was small, the differences were too small to be detected.

TABLE VI: Descriptive statistics for the *creation* task with the 4 templates.

	Metric	Template	Mean	S.D.	Skew.	Kurt.	S-W
Accuracy	Precision	CON	.315	.310	.475	-.492	.110
		BEN	.314	.251	-.233	-1.603	.138
		PER	.437	.305	-.327	-1.526	.143
		PERBEN	.197	.253	.807	-1.003	**.004**
	Recall	CON	.151	.172	.689	-.937	**.047**
		BEN	.136	.137	.566	-1.237	.073
		PER	.082	.0682	.131	-1.723	.203
		PERBEN	.023	.029	.711	-1.146	**.004**
	F-Measure	CON	.196	.218	.696	-.592	.053
		BEN	.179	.173	.428	-1.478	.074
		PER	.136	.110	.079	-1.705	.256
		PERBEN	.041	.051	.641	-1.441	**.003**
Speed	Duration	CON	559.204	405.964	1.249	1.167	.112
		BEN	537.305	215.881	.711	1.251	.825
		PER	249.804	109.024	1.135	1.616	.228
		PERBEN	338.912	166.197	1.034	.441	.132
Visual effort	FixRel	CON	1400.143	2172.676	2.440	6.152	**.001**
		BEN	976.111	588.593	.608	.909	.633
		PER	252.200	224.680	.732	-.792	.165
		PERBEN	375.875	175.853	1.079	1.995	.348
	FixIrrel	CON	1770.714	2275.771	1.982	4.289	**.015**
		BEN	1575.888	1406.623	.679	-.995	.290
		PER	1035.400	793.518	1.699	2.969	**.031**
		PERBEN	1531.625	1062.398	.055	-1.520	.409
	AvgFixRel	CON	460.326	170.913	1.176	1.912	.240
		BEN	408.564	122.538	3.000	9.000	**.000**
		PER	494.308	149.004	1.093	1.213	.116
		PERBEN	409.060	29.393	-.105	-.559	.940
	AvgFixIrrel	CON	540.692	100.938	-.881	-.961	.123
		BEN	1334.434	4001.212	3.000	9.000	**.000**
		PER	2093.269	6306.823	3.156	9.969	**.000**
		PERBEN	3301.715	9335.950	2.828	8.000	**.000**
	Saccades	CON	530.000	528.030	1.862	4.194	**.040**
		BEN	492.330	265.687	-.330	-1.442	.309
		PER	266.900	140.717	1.021	1.640	.144
		PERBEN	347.000	217.285	.073	-1.806	.331

TABLE VII: Levene's and Welch's test for the *creation* task.

Metric	Levene Sig.	Welch Sig.
Recall	.000	.011
F-measure	.000	.017
Duration	.028	.006
FixRel	.002	.014

Assessing speed. There was a statistically significant difference among the templates, regarding speed. The task *duration* when using PER was lower ($M = 249.804$, $SD = 109.024$) than when using BEN ($M = 537.305$, $SD = 215.881$).

Assessing visual effort. There was a statistically significant difference among the templates, regarding visual effort. The *fixation rate on relevant elements* was higher for participants using BEN ($M = 976.111$, $SD = 588.59$) than for those using PER ($M = 252.200$, $SD = 224.680$). In Figure 2 we present the boxplots for this metric.

Assessing perceived effort. There was no statistically significant difference among templates, regarding perceived effort. We found no evidence that different templates influence the perceived effort when creating user stories.

RQ2: *Do different user story templates have an impact on the understanding of user stories?*

In Table IX we present Levene's and Welch's test results

TABLE VIII: Games-Howell post-hoc test for the *creation* task.

Metric	Template (I)			Template (J)			Mean difference	Sig.
	Name	Mean	S.D.	Name	Mean	S.D.		
Duration	BEN	537.305	215.881	PER	249.804	109.024	287.501	.011
FixRel	BEN	976.111	588.593	PER	252.200	224.680	723.911	.026

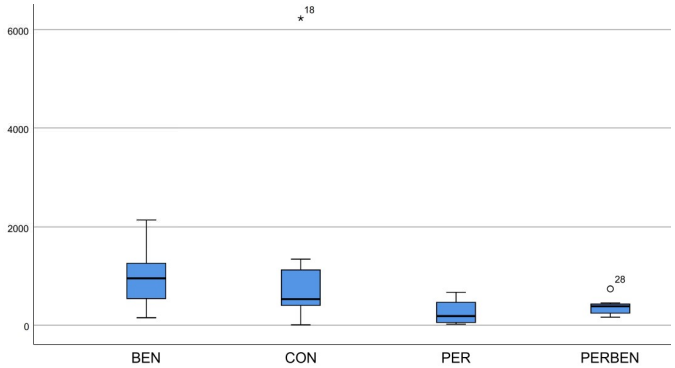


Fig. 2: Boxplots for the *fixation rate on relevant elements*.

for the *understanding* task, for metrics with a statistically significant difference in the Welch *t*-test.

TABLE IX: Levene's and Welch's test for the *understanding* task.

Metric	Levene Sig.	Welch Sig.
FixRel	.001	.002
FixIrrel	.246	.007
AvgFixIrrel	.002	.034

In Table X we summarise the Games-Howell post-hoc test results, as well as present the mean and standard deviation for the different templates, for all the metrics that had a statistically significant difference.

Assessing accuracy. There was no statistically significant difference among templates, regarding accuracy. We found no evidence that different templates influence the accuracy when understanding user stories.

Assessing speed. There was no statistically significant difference among templates, regarding speed. We found no evidence that different templates influence the speed when understanding user stories.

Assessing visual effort. There was a statistically significant difference among the templates, in terms of visual effort. The *fixation rate on relevant elements* was higher for participants using CON ($M = 235.200$, $S.D. = 161.965$) than for those using PER ($M = 22.666$, $S.D. = 27.866$) and for those using PERBEN ($M = 16.125$, $S.D. = 17.431$). The *fixation rate on irrelevant elements* was higher for participants using PER ($M = 1005.555$, $S.D. = 525.215$) than for those using CON ($M = 279.300$, $S.D. = 283.641$) and for those using BEN ($M = 344.000$, $S.D. = 343.753$). Finally, the *average duration of irrelevant fixations* was higher for participants using PER ($M = 532.145$, $S.D. = 95.467$) than for those using CON ($M = 378.926$, $S.D. = 133.624$). In Figures 3 and 4 we present the boxplots for the fixation rates on relevant and irrelevant elements, respectively, collected on understanding tasks.

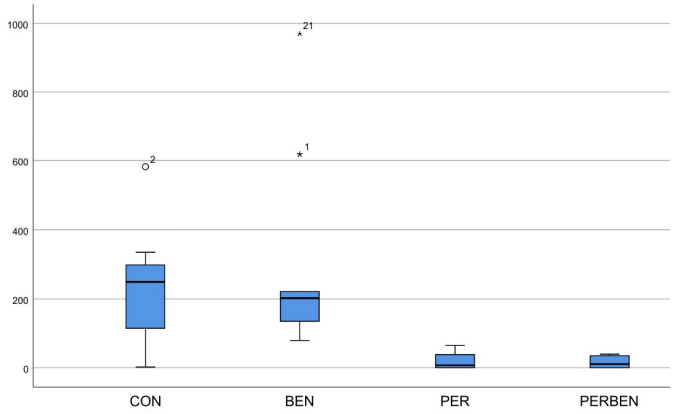


Fig. 3: Boxplots for the *fixation rate on relevant elements*.

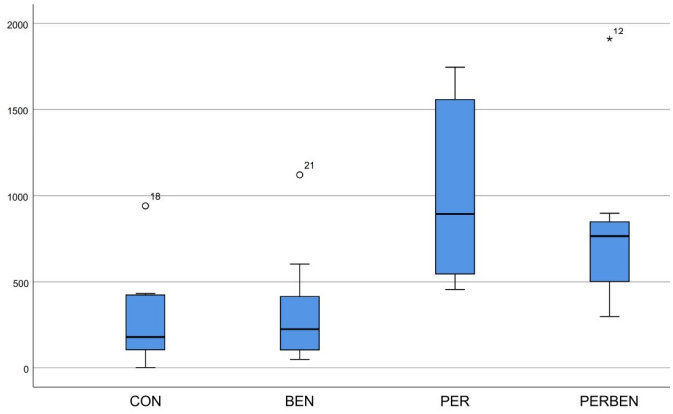


Fig. 4: Boxplots for the *fixation rate on irrelevant elements*.

When analysing the heatmaps generated during the understanding tasks, we observed a tendency for a higher effort in the first user stories (particularly the first and second), when compared to the last ones (i.e., user stories six and seven), regardless of the used template. The user stories were randomly sorted, so this is also independent of the specific user stories order. In Figure 5 we illustrate this observation for the CON template, being analysed by two different participants.

Assessing perceived effort. There was no statistically significant difference among templates, regarding perceived effort. We found no evidence that different templates influence the perceived effort when understanding user stories.

TABLE X: Games-Howell post-hoc test for the *understanding* task.

Metric	Template (I)			Template (J)			Mean difference	Sig.
	Name	Mean	S.D.	Name	Mean	S.D.		
FixRel	CON	235.200	161.965	PER	22.666	27.866	212.533	.011
				PERBEN	16.125	17.431	219.075	.009
FixIrrel	PER	1005.555	525.215	CON	279.300	283.641	726.256	.014
				BEN	344.000	343.753	661.556	.032
AvgFixIrrel	PER	532.145	95.467	CON	378.926	133.624	153.219	.046

(a) Participant 2 understanding task.

(b) Participant 13 understanding task.

Fig. 5: Heatmaps for the understanding task, when using the CON template.

VII. Discussion

A. Evaluation of results and implications

RQ1: *Do different user story templates have an impact on the creation of user stories?*

The Welch *t*-test shows significant differences in terms of **accuracy**, but it was not possible to identify these differences through the Games-Howell *post-hoc* test. The differences are small and the test fails to detect them.

There were significant differences in terms of **speed**. Participants using BEN, a template without persona, were slower to complete the task then the ones using PER, a template with persona. There were also significant differences in terms of **visual effort**. Participants using BEN had a higher *fixation rate on relevant elements*, than the ones using PER. With BEN, participants were looking at the right elements more often, but were taking longer to recognise them as relevant. We argue that not having a persona might have hindered a faster identification of relevant elements by our participants.

There were no significant differences in terms of **perceived effort**, as reported by our participants through the NASA-TLX questionnaire. This suggests that all templates require a similar perceived mental workload and cognitive effort to create user stories.

We also observed that tacit knowledge influenced participants' answers. The user stories created by some of our participants covered features that were not included in the scenario. Due to their experience with this type of systems, participants tended to use their knowledge instead of analysing the presented information. Furthermore, some participants, already familiar with user stories, chose to perform the creation task with CON, instead of using the template presented in the video tutorial. Since this template is typically taught in an academic context, it is possible that participants' previous experience contributed to its use, even when CON was not the intended template.

RQ2: *Do different user story templates have an impact on the understanding of user stories?*

The Welch *t*-test shows significant differences in terms of **visual effort**. Participants using templates without personas, CON and BEN, had a higher *fixation rate on relevant elements*, and a lower *fixation rate on irrelevant elements*, respectively. Our interpretation is that, without personas, participants were able to find the relevant elements more easily. We argue that the higher amount of information presented when having personas may have caused a more comprehensive analysis, and thus a greater visual effort observable through the higher fixations on irrelevant elements, when using PER.

There were no significant differences in terms of **perceived effort**, as reported by our participants through the NASA-TLX questionnaire. This suggests that all templates require a similar perceived mental workload and cognitive effort to understand user stories.

Regardless of the template, and the order of the user stories, which were randomly presented to participants, we observed that participants dedicated more effort to understand the first couple of user stories than the last ones. This suggests that they became increasingly confident in their ability to understand user stories, as the repetitive nature of the template usage fosters familiarity.

B. Threats to validity

For the threats to validity, we are following Wohlin et al.'s guidelines [45].

Internal validity. We used a combination of convenience and snowball sampling. This can cause a selection threat, since the participants tend to be more motivated to be part of the experiments, considering that their participation is entirely voluntary. However, we found no evidence of this in the results. Moreover, we have made available an independent

replication package [39] to colleagues from other organisations and countries.

Conclusion validity. The number of participants is not as high as intended. In order to guarantee an adequate power level, we needed a sample with a minimum of 26 participants per template, 104 in total [51]. However, due to schedule and availability restrictions, both for companies and those contacted, it was not possible to reach this number of participants. One of the consequences was noticeable in the analysis of the results, where the post-hoc test could not detect differences among groups, due to the small size of the sample. We plan to launch a replication of this experiment with a higher number of participants and We encourage replications of the quasi-experiment with a larger group.

External validity. Some of our participants had little to no prior knowledge in user stories. However, we found no different between experienced and non-experienced participants. The number of user stories presented to participants was relatively small. This may not be representative of the number of user stories that a stakeholder needs to analyse in a real-world situation. However, we could not have a long list of user stories, since we were limited by the technical specifications of the eye-tracker device, such as constraints in the external monitor dimensions and in the participant distance to the eye-tracker. The font used had to be big enough for easy visualisation by all participants. In future replications, it is important to vary the number of the presented user stories, to assess whether there is a significant difference on the success and effort of the task. Finally, all tasks were in English. However, our participants have Portuguese as their mother tongue. We decided to create all the materials in English so they could be used in independent replications by international researchers. However, limited English proficiency could have impacted the results. Nevertheless, all the participants were at ease with the English language and we found no impact of this decision in the results obtained.

Construct validity. We showed a video tutorial about the user stories template used in the task. As such, participants might have felt they were being evaluated. This might have caused an evaluation apprehension threat, where participants try to look better and thus confound the results. To mitigate this threat, we have not informed them about what exactly was being tested. The video allowed all participants to have the same information, avoiding having participants with different levels of knowledge, which could impact on the results. Furthermore, for the creation task, there is a risk that the final gold-standard user stories (explained in Subsection VI-A) is not complete and there may be further valid user stories that were not considered. We mitigate this threat by having experienced user stories researchers creating and validating the gold-standard, as well as changing the user stories set based on the data analysis.

VIII. Conclusions and future work

We performed a quasi-experiment to analyse the impact of 4 user story templates (CON, BEN, PER and PERBEN) when creating and understanding user stories. These templates vary in the order of their elements, and in the usage, or not, of a persona. We measured accuracy, speed, visual effort and perceived effort of 41 participants. We used metrics of tasks success; time; visual effort, collected with an eye-tracker; and participant's feedback through NASA-TLX questionnaire. Each participant performed the tasks with only one of the templates.

For the creation task, there were statistically significant differences for some of the collected metrics, namely *recall* and *f-measure* (accuracy), *duration* (speed) and *fixation rate on relevant elements* (visual effort). For the understanding task, there were statistically significant differences in three of the collected visual effort metrics: *fixation rate on relevant elements*, *fixation rate on irrelevant elements* and *average duration of irrelevant fixations*. However, in practice, these differences seem to have a negligible effect on the perceived effort, as reported by our participants through the NASA-TLX questionnaire.

In addition, the results showed that, although some templates outperformed others in a few metrics, no template obtained the best overall result. As such, there was not a real difference among the templates, and we found no compelling evidence that one template is "better" than the others. In practice, this means that, at least for problem descriptions of this size and for this number of user stories, it does not make a difference which user story template (CON, BEN, PER, PERBEN) is being used.

However, while using personas, we observed a greater visual effort. This might stem from the added information that those templates carry, when compared to the ones without personas. On the other hand, this was not perceived by the participants in terms of their subjective effort. Further studies may shed some light in terms of the actual benefits and shortcomings of adopting personas in these templates.

We plan to replicate the experiment in other contexts, and with more complex scenarios. This would allow us to understand whether the lack of significant differences is attributable to the size of the tasks we present to the participants. We also plan to perform other studies focusing on different tasks performed with these user story templates, such as modification and reviewing.

Acknowledgment

We thank NOVA LINCS (UIDB/04516/2020) for the financial support of FCT – Fundação para a Ciência e a Tecnologia, through national funds.

References

[1] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *Conference on the Future of Software Engineering*, 2000, pp. 35–46.

[2] D. Méndez and S. Wagner, "Naming the pain in requirements engineering," Access: December 2020. [Online]. Available: http://napire.org/#/home

[3] M. Cohn, *User Stories Applied: For Agile Software Development*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 2004.

[4] A. Marcano, "Old favourite: Feature injection user stories on a business value theme," Access: Dec. 2020. [Online]. Available: http://antonymarcano.com/blog/2011/03/fi_stories/

[5] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries *et al.*, "Manifesto for agile software development," Access: December, 2020. [Online]. Available: https://www.agilealliance.org/agile101/the-agile-manifesto/

[6] G. Lucassen, F. Dalpiaz, J. van der Werf, and S. Brinkkemper, "The use and effectiveness of user stories in practice," in *REFSQ 2016*, 2016, pp. 205–222.

[7] R. Pichler, "From personas to user stories," Access: December 2020. [Online]. Available: https://www.romanpichler.com/blog/personas-epics-user-stories/

[8] ——, "10 tips for writing good user stories," Access: December 2020. [Online]. Available: https://www.romanpichler.com/blog/10-tips-writing-good-user-stories/

[9] R. Radach, J. Hyona, and H. Deubel, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, 1st ed. Elsevier, 2003.

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

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

[12] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Engineering ICSE 2014*. ACM, 2014, pp. 402–413.

[13] F. Paas and J. van Merriënboer, "The efficiency of instructional conditions: An approach to combine mental effort and performance measures," *Human Factors*, vol. 35, no. 4, pp. 737–743, 1993.

[14] Y. Yeh and C. Wickens, "Dissociation of performance and subjective measures of workload," *Human Factors*, vol. 30, no. 1, pp. 111–120, 1988.

[15] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[16] J. Choma, L. Zaina, and D. Beraldo, "Userx story: incorporating ux aspects into user stories elaboration," in *Human-Computer Interaction HCI 2016*. Springer, 2016, pp. 131–140.

[17] O. Liskin, R. Pham, S. Kiesling, and K. Schneider, "Why we need a granularity concept for user stories," in *International Conference on Agile Software Development*. Springer, 2014, pp. 110–125.

[18] J. Lin, H. Yu, Z. Shen, and C. Miao, "Using goal net to model user stories in agile software development," in *SNPD 2014*. IEEE, 2014, pp. 1–6.

[19] F. Dalpiaz, I. van der Schalk, S. Brinkkemper, F. Aydemir, and G. Lucassen, "Detecting terminological ambiguity in user stories: Tool and experimentation," *Inf. Softw. Technol.*, vol. 110, pp. 3–16, 2019.

[20] F. Dalpiaz and A. Sturm, "Conceptualizing requirements using user stories and use cases: A controlled experiment," in *REFSQ 2020*, vol. 12045, 2020, pp. 221–238.

[21] O. Liskin, R. Pham, S. Kiesling, and K. Schneider, "Why we need a granularity concept for user stories," in *XP 2014*, vol. 179. Springer, 2014, pp. 110–125.

[22] J. Lin, H. Yu, Z. Shen, and C. Miao, "Using goal net to model user stories in agile software development," in *SNPD 2014*. IEEE Computer Society, 2014, pp. 1–6.

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

[24] Y.-G. Guéhéneuc, "Taupe: Towards understanding program comprehension," in *Proceedings of the 2006 Conference of the Center for Advanced Studies on Collaborative Research*, ser. CASCON '06. Riverton, NJ, USA: IBM Corp., 2006.

[25] S. Yusuf, H. Kagdi, and J. I. Maletic, "Assessing the comprehension of uml class diagrams via eye tracking," in *5th IEEE International Conference on Program Comprehension, 2007. ICPC'07*. IEEE, 2007, pp. 113–122.

[26] S. Jeanmart, Y.-G. Guéhéneuc, H. Sahraoui, and N. Habra, "Impact of the visitor pattern on program comprehension and maintenance," in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '09. IEEE Computer Society, 2009, pp. 69–78.

[27] B. Sharif and J. I. Maletic, "An eye tracking study on the effects of layout in understanding the role of design patterns," in *IEEE International Conference on Software Maintenance (ICSM), 2010*. IEEE, 2010, pp. 1–10.

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

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

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

[31] C. Gralha, M. Goulão, and J. Araújo, "Analysing gender differences in building social goal models: a quasi-experiment," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019, pp. 165–176.

[32] M. E. Crosby, J. Scholtz, and S. Wiedenbeck, "The roles beacons play in comprehension for novice and expert programmers," in *14th Workshop of the Psychology of Programming Interest Group*, 2002, pp. 58–73.

[33] T. Busjahn, R. Bednarik, and C. Schulte, "What influences dwell time during source code reading?: Analysis of element type and frequency as factors," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14, 2014, pp. 335–338.

[34] P. Rodeghero, C. McMillan, P. W. McBurney, N. Bosch, and S. D'Mello, "Improving automated source code summarization via an eye-tracking study of programmers," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014, 2014, pp. 390–401.

[35] J. Katona, "Clean and Dirty Code Comprehension by Eyetracking Based Evaluation using GP3 Eye Tracker," *Acta Polytechnica Hungarica*, vol. 18, no. 1, pp. 79–99, 2021.

[36] T. Busjahn, C. Schulte, and A. Busjahn, "Analysis of code reading to gain more insight in program comprehension," in *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*, 2011, pp. 1–9.

[37] D. Binkley, M. Davis, D. Lawrie, J. I. Maletic, C. Morrell, and B. Sharif, "The impact of identifier style on effort and comprehension," *Empirical Software Engineering*, vol. 18, no. 2, pp. 219–276, 2013.

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

[39] C. Gralha, R. Pereira, M. Goulão, and J. Araujo, "On the impact of different templates on creating and understanding user stories - Supplemental Material," 2021, (Access: July, 2021). [Online]. Available: http://doi.org/10.5281/zenodo.5113924

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

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

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

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

[44] TLX@NASA, "Nasa tlx paper/pensil version," (Access: December 2020). [Online]. Available: https://humansystems.arc.nasa.gov/groups/TLX/tlxpaperpencil.php

[45] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[46] S. Vegas, C. Apa, and N. Juristo, "Crossover designs in software engineering experiments: Benefits and perils," *IEEE Trans. Software Eng.*, vol. 42, no. 2, pp. 120–135, 2016.

[47] The Eye Tribe eye-tracker, "The eyetribe," Access: December 2019. [Online]. Available: https://theeyetribe.com/

[48] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

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

[50] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

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