What is the Impact of Bad Layout in the Understandability of Social Goal Models?

Mafalda Santos, Catarina Gralha, Miguel Goulão, João Araújo, Ana Moreira and João Cambeiro*
NOVA LINCS, DI, FCT, Universidade NOVA de Lisboa
{mcd.santos, acg.almeida, *jmc12976}@campus.fct.unl.pt, {mgoul, joao.araujo, amm}@fct.unl.pt

***Abstract*—The *i** community has published guidelines, including model layout guidelines, for the construction of models. Our goal is to evaluate the effect of the layout guidelines on the *i** novice stakeholders' ability to understand and review *i** models. We performed a quasi-experiment where participants were given two understanding and two reviewing tasks. Both tasks involved a model with a bad layout and another model following the *i** layout guidelines. We evaluated the impact of layouts by combining the success level in those tasks and the required effort to accomplish them. Effort was assessed using time, perceived complexity (with NASA TLX), and eye-tracking data. Participants were more successful in understanding than in reviewing tasks. However, we found no statistically significant difference in the success, time taken, or perceived complexity, between tasks conducted with models with a bad layout and models with a good layout. Most participants had little to no prior knowledge in *i**, making them more representative of stakeholders with no requirements engineering expertise. They were able to understand the models fairly well after a short tutorial, but struggled when reviewing models. In the end, adherence to the existing *i** layout guidelines did not significantly impact *i** model understanding and reviewing performance.**

***Index Terms*—social goal models; i*; diagram layout; task complexity; eye-tracking**

I. Introduction

The success of Requirements Engineering depends critically on effective communication between requirements engineers and other stakeholders [1]. However, the extent to which software engineering visual modelling languages are adequate for communication purposes has been somewhat neglected [2]. Indeed, the communication potential of these languages is not fully explored, as their cognitive effectiveness is often not optimised. Several modelling languages have been criticised for their lack of semantic transparency, making it hard for non-experts to correctly recognise their symbols (e.g., UML [3] and *i** [1]). Even if one correctly recognises the symbols of a language, understanding the domain concerns in the models is yet another challenge. Thus, choosing an adequate layout for requirements models may be a relevant issue, as a bad layout may compromise the adequacy of the models. In the long run, poorly understood requirements may lead to problems in artifacts produced in later stages of software development.

Although the visual syntax of a language is usually subordinate to its semantics [2], organisations and languages' creators may propose what is understood as a good layout, as a standard to follow. It is believed that these guidelines are important to improve the readability and, consequently, the overall understandability of models used in software development. But one may ask: is that true? What is the actual impact of adhering to such guidelines, producing models with good or bad layouts, on understanding and reviewing those models?

We evaluate the adequacy of such guidelines and the potential impact of not following them, for the case of *i** models. The *i** guide, available at the *i** wiki, offers several layout rules proposed by experts. Our evaluation studies the success of understanding and reviewing *i** models, collecting measures such as precision, recall, and F-measure, the duration of those tasks, the visual effort while performing them (assessed with eye-tracking [4]), and the participants perceptions on their effort while performing the tasks, measured with a NASA Task Load Index (NASA TLX) questionnaire [5].

We define the main goal of this paper following the goal template formulated in [6]. Our goal is to **analyse** the *i** layout guidelines **for the purpose of** their evaluation **with respect to** their impact on the effort required for understanding and reviewing *i** Strategic Rationale models, **from the point of view of** potential system stakeholders, whose experience on *i** is expected to be reduced or non-existent, **in the context of** a research project conducted in the Informatics Department of Universidade Nova de Lisboa (UNL).

Participants were undergraduate and graduate students, postdocs, and staff members of UNL, with little to no prior knowledge in *i**. They were recruited as surrogates for stakeholders without *i** expertise to perform the tasks of understanding and reviewing *i** models.

The paper is organised as follows. Section II describes background information about *i**, the *i** models design recommendations, related studies, and relevance to practice. Section III reports the experiment planning, including goals, participants, experimental material, tasks, hypotheses, design, procedure, and analysis procedure. Section IV describes the experiment execution, with the preparation and deviations from the plan. Section V analyses the results, including descriptive statistics, data set preparation and hypothesis testing. Section VI presents and discusses the results, also reporting threats to validity and inferences. Section VII presents conclusions and future work.

II. Background

A. The *i** framework

The *i** framework was developed for modelling and reasoning about organisational environments and their information systems, covering both agent and goal-oriented modelling [7]. It focuses on the concept of *intentional actor*. Actors, in their

organisational environment, are viewed as having intentional properties, such as *goals*, *abilities* and *commitments*.

This framework has two main modelling components: the *Strategic Dependency* (SD) model and the *Strategic Rationale* (SR) model. The SD model describes the dependency relationships, through *dependency links*, among the actors in an organisational context. An actor (or *depender*) depends on another actor (the *dependee*) to achieve *goals* and *softgoals*, to perform *tasks*, to obtain *resources*, and to express *beliefs*. The SR model provides a more detailed level of modelling than the SD model, as it focuses on modelling intentional elements and relationships internal to actors. Intentional elements (goals, softgoals, tasks, resources and beliefs) are related by *means-ends* or *decomposition links*. *Means-ends* links can be perceived as decomposition links that are used to link *goals* (ends) to *tasks* (means) to specify alternative ways to achieve goals. *Decomposition links* are used to decompose tasks into a *subgoal*, a *subtask*, a *resource*, and/or a *softgoal*. Additionally, *contribution links*, which can be *positive* or *negative*, and are used to link intentional elements to *softgoals*.

B. i* models design recommendations

The *i** community gathers a set of modelling recommendations in the *i** wiki[1], namely the *i* Guide*, intended to be both an introduction to *i** for new users and a reference for experienced users. Each guideline is annotated with a set of attributes of the form *(Level, Type)*. *Level* can be Beginner, Intermediate, or Advanced, and *Type* can be Concept, Naming, Notation, Layout, Methodology, or Evaluation. Since our goal is to analyse the *i** layout, we selected the guidelines within the **Layout** type. According to the *i** Guide, layout *"deals with the arrangement and organisation of i* models and the way the contents of the models appear and are placed. It also covers issues related to modelling space and complexity."*.

For designing the models, we used OpenOME[2], one of the most popular *i** tools. Table I shows *(i)* the layout guidelines from the *i* Guide*, *(ii)* if the guideline is enforced by the chosen tool, and *(iii)* whether we have considered it as a layout guideline or not. Although all of the presented guidelines are categorised with the layout type by the *i** wiki, we have a different interpretation for some of them. We consider guidelines #9, #11 and #17 as being concerned with the well-formedness of the models, not with layout. For guidelines #12 and #16, we considered they are related with the completeness of the model, not with layout. Hence, those guidelines were not considered when preparing the models for this experiment.

C. Related studies

Störrle has found a significant impact of the usage of good *vs.* bad diagram layouts on model comprehension tasks when using UML analysis models often used during requirements elicitation, namely, use case, class, and activity diagrams [8]. He reported on three controlled experiments with 77 participants to support this claim. He also noted that novices benefit considerably more than experts from the usage of a layout adhering to several layout heuristics applied simultaneously, when compared to a layout violating such heuristics.

Eye-tracking has been used on some occasions to assess the effort involved in the comprehension of software models, by monitoring participant's visual attention through fixations and other indicators [4]. Fixation is the stabilisation of the eyes on a part of a stimulus (object of interest presented on screen) during a period of time (200-300 ms). Psychology studies reveal that most information acquisition and cognitive processing occur during fixations. Thus fixation data is used to calculate metrics that estimate visual effort based on fixations number and duration in a certain area of interest (AOI) of the stimulus. Yusuf et al. [9] used eye-tracking to compare the visual effort involved in answering questions about UML class diagrams containing the same information, but designed following 3 different layout strategies: multiple-cluster (classes with related functionality are in clusters); three-cluster (positions classes based on their stereotype role) and orthogonal layout (minimises edge crossings and bending). They concluded that *multiple-cluster* outperformed *three-cluster* and *orthogonal* layouts, as participants had to make, on average, a smaller number of fixations on the diagram. This study was later extended without eye-tracking, confirming the previous results [10]. Again including eye-tracking data, the effect of different layouts was also studied for design pattern roles identification in UML class diagrams [11], [12]. Another eye-tracking study showed that although the presence of a visitor pattern and its layout had no significant impact on the comprehension of UML class diagrams, it did have a significant impact in modification tasks [13]. A common feature in all these studies is the concern with the importance of some aspect of a UML diagram layout (be that a layout heuristic, or the explicit usage of a particular design pattern).

Other studies with eye-tracking focused on BPMN [14], ER [15], and TROPOS diagrams [16]. The later contrasted the effectiveness of a textual and the TROPOS diagrammatic requirements language for requirements comprehension purposes and the textual language turned out to be more effective. Our requirements understanding tasks are similar in complexity to those in [16], but using *i**, and changing the contrasted treatments (in our case, good *vs.* bad layout, rather than a textual notation *vs.* a diagrammatic one).

D. Relevance to practice

The exploitation of human factors is increasingly regarded as a relevant topic in Software Engineering [17] in general, and modelling in particular. The Requirements Engineering community is concerned with bridging the perceived gap between sophisticated requirements engineering approaches, such as *i**, and the stakeholders which requirements engineers need to interact with. The most common requirements notation remains to be natural language. The problem is that natural language often leads to ambiguous requirements

[1] *i** wiki: http://istarwiki.org/

[2] OpenOME: http://www.cs.toronto.edu/km/openome/

TABLE I: Guidelines for *i** models' layout

#	Guideline	Enforced	Layout
1	Avoid or minimise drawing intersecting Links and overlapping Links with other Links and elements' text	No	Yes
2	Make both sides of a Dependency Link look like a single, continuous curve as it passes through the Dependum	Yes	Yes
3	Spread the connection points of Dependency Links out on an Actor	Yes	Yes
4	Keep elements horizontal. Do not tilt or twist them	Yes	Yes
5	Avoid or minimise overlapping boundaries of Actors where possible	No	Yes
6	Keep Dependency Links outside the boundaries of Actors to improve the readability of the models	No	Yes
7	Use the conventional Actors' boundaries (circles) unless other shapes such as rectangles can improve models' layout	Yes	Yes
8	Avoid overlapping elements inside or outside Actors	No	Yes
9	Connect each Dependency Link in an SR model to the correct element within the actor	No	No
10	Adopt or follow a consistent direction for the goal refinement/ decomposition hierarchy as much as possible	No	Yes
11	Do not draw SR model elements outside the boundaries of the corresponding actors	No	No
12	Unconnected elements within an Actor is indicative of an incomplete model	No	No
13	Don't extend the text of the name of the element beyond the element's border	Yes	Yes
14	Split a large and complex model into consistent pieces to facilitate easier presentation and rendering	No	Yes
15	Don't extract or zoom into a section of an Actor in a model without showing the incoming and outgoing dependencies with other actors or parts of the model	No	Yes
16	Use the specialised actors notation to the degree that you can gain advantage in instantiating the actual stakeholders	Yes	No
17	Use the leaf-level tasks as the system requirements, not the high level Goals and Softgoals	No	No

specifications. On the other hand, specialised requirements engineering frameworks support reasoning about the requirements, but are poorly understood by relevant stakeholders. So, devising ways of making these requirements languages more accessible is perceived as very important. This can be achieved, for example, by improving the visual metaphors used by a language (see, e.g., [1]). In this paper, we take on a complementary approach by assessing the impact of layout in the understandability of models, and in the ability to review them. As discussed in section II-C, previous studies pointed to a significant impact of layout style in model understandability.

III. Experiment Planning

A. Goals

We describe our two research goals following the GQM research goals template [6]. Our first goal (G1) is to **analyse** the effect of *i** model layout, **for the purpose of** evaluation, **with respect to** its impact on the *understandability* of *i** SR models, **from the viewpoint of** researchers, **in the context of** an experiment conducted with participants with limited or no experience with *i** at UNL. Our second goal (G2) is to **analyse** the effect of *i** model layout, **for the purpose of** evaluation, **with respect to** its impact on the *review* of *i** SR models, **from the viewpoint of** researchers, **in the context of** an experiment conducted with participants with limited or no experience with *i** at UNL.

B. Participants

The participants in this experiment were recruited through convenience sampling. They were made aware of this ongoing study and volunteered to participate. As members of the UNL community, they are aware of the importance of performing evaluations with participants and, therefore, willing to volunteer. Several of them have conducted, or will conduct in a near future, evaluations in the context of their own research projects, so motivating them to participate was not a problem. Participants read a *"Participant consent letter"*, adapted from the one in [18], where we explained that they could leave the experiment at any point, if they desired to do so, and that all the collected data would remain anonymous. They were offered the possibility of receiving a notification on the results of the study, when the final report became available.

The experiment was initially performed by 24 participants. However, 6 cases were excluded due to problems with the data collection, detected at the end of the experiment. We used the data of the remaining 18 participants.

We collected demographic data on *age*, *gender*, *nationality*, *field of studies*, *usage of medical devices* (glasses, or contact lenses), *completed education level*, *current occupation*, and previous *experience* with *i**. Concerning the usage of *medical devices*, 2 of the participants had contact lenses, and 4 were wearing glasses. With regard to previous *experience* with *i**, 4 learnt *i** in the context of a course and 14 did not know it.

C. Experimental material

In this evaluation, we wanted to assess the effect of good *vs.* bad layout in *i** models in two particular tasks: *understanding* and *reviewing* *i** SR models. We designed our experiment to test good and bad layouts for *understanding* correct models, and for *reviewing* incorrect models. So, each participant had 4 models to examine, all from different domains. We designed 4 different models, 2 correct and 2 incorrect. Each of those models was then depicted in two versions: with a good layout, and with a bad layout. All the used models are similar in size, and use the same model elements, just varying the domain. Concerning the bad layout and the injected model defects, their size was also similar. Thus, the models were very similar, but from very distinct domains, to minimise learning effects in the experiment. The chosen domains were *gaming centre*, *tolls system*, *patient wellness tracking*, and *goods acquisition*. Table II summarises the models' size, namely their number of actors, intentional elements, and dependencies between actors.

Fig. 1 provides an example of the tasks made available to participants, all following a common structure. At the top, we have the question the participant is supposed to answer. On the left, we have a key with the main *i** elements in the models, so

TABLE II: Basic metrics about the models

	#actors	#elements	#dependencies
Gaming Centre	2	21	2
Tolls System	2	21	2
Patient Wellness	2	21	2
Goods Acquisition	2	17	1

that the participant can check what a particular symbol means. The main area of the screen has an SR model, about which the question is asked. Each model area has a relevant area, containing the model elements corresponding to the correct answer, and an irrelevant area, with all the remaining elements.

D. Tasks

Each participant performed four tasks. Two of the performed tasks were aimed at evaluating the effort in *understanding* a correct *i** model. In one of them, the model had a good layout (according to the guidelines discussed in section II-B), while on the other the layout was bad (violating several layout guidelines). Each participant also performed two *reviews*, to detect seeded defects in *i** models. Again, one of the models used in the review session had a good layout, while the other had a bad one. Both in the understanding and in the reviewing tasks, the participants had to answer orally a simple question. For example, an *understanding* task would be to enumerate the tasks involved in making payments, in the goods acquisition system. The *reviewing* task consisted in describing all the defects the participant could identify in a given model. The answers were recorded in audio, for further analysis, along with a video with the contents of the screen during task execution, annotated with eye-tracking data. No eye-tracking feedback was visible to the participant while performing the evaluation, to avoid an unnecessary validity threat. Likewise, no feedback on the success of the tasks was provided to the participants, preventing contamination of subsequent tasks.

Each task was followed by a corresponding NASA TLX questionnaire, so that participant's feedback was collected on their experience while performing the task. The participant filled the form made available through the web browser.

E. Hypotheses, parameters and variables

For each of the two goals, we have defined the null hypothesis (H_0) and the alternative hypothesis (H_1).

$H_{0Understand}$: Adherence to layout guidelines do not influence the *understandability* of requirements expressed in *i**.

$H_{1Understand}$: Adherence to layout guidelines influences the *understandability* of requirements expressed in *i**.

$H_{0Review}$: Adherence to layout guidelines do not influence the performance when *reviewing* requirements expressed in *i**.

$H_{1Review}$: Adherence to layout guidelines influences the performance when *reviewing* requirements expressed in *i**.

The independent variable is the *layout*, which may be *bad*, or *good*. The dependent variables are:

- *Precision* — the fraction of model elements retrieved by participants (for the first hypothesis) or of defects (for the second hypothesis) which are relevant.
- *Recall* — the fraction of relevant model elements (or of relevant defects) retrieved by participants, over the total number of model elements (or potential defects) retrieved.
- *F-measure* — a measure that combines precision and recall, computed as $\frac{2*(Precision*Recall)}{(Precision+Recall)}$; this measure provides an harmonic mean of precision and recall.
- *Duration* — the time taken by the participants to complete the task.
- *NASA TLX score* — an overall weighted score resulting from the application of the TLX questionnaire, covering perceived mental, physical and temporal demand, performance, effort and frustration while performing a task.
- *FixRel* — Fixation Rate on Relevant elements; the fraction of number of fixations in an given AOI over the total number of fixations in the AOG (Area of Glance).
- *FixIrrel* — Fixation Rate on Irrelevant elements; the fraction of number of fixations in an given AOI over the total number of fixations in the AOG.
- *AvDurFixRel* — Average Duration of Relevant Fixation; the fraction of total duration of fixations for relevant AOIs over the number of elements of the relevant AOIs.
- *AvDurFixIrrel* — Average Duration of Irrelevant Fixation; the fraction of total duration of fixations for relevant AOIs over the number of elements of the relevant AOIs.

Precision, *Recall*, and the *F-measure* are used as metrics for the success of the tasks being performed. *Duration* is used to assess the efficiency in performing the task. *NASA TLX score* measures the self-perception of effort made by the participants, while performing the tasks. Finally, the eye-tracking metrics *FixRel*, *FixIrrel*, *AvDurFixRel*, and *AvDurFixIrrel* evaluate visual effort in the understanding tasks. A higher number and duration of fixations can be associated with a higher visual attention in a given set of AOIs (in this case, relevant *vs.* irrelevant model elements) of the stimulus (the screen) presented to the participant [19]. We are only using these eye-tracking metrics for the first hypothesis. Although we collected eye-tracking data for the whole session, the error margin of the eye-tracking data makes it unreliable to tag really small elements in the screen (such as the annotations on dependencies), which were crucial for the reviewing task.

F. Design

To reduce learning effects, the order of the 4 tasks changed from participant to participant. All participants had two understanding tasks and two review tasks. Each of the tasks used one of 4 models. Each of these models had 2 versions, one with a layout following the guidelines, another with a layout violating several of these guidelines. Other than the layout, the 2 model versions were exactly the same. We balanced the number of times each task was performed as first, second, third, or fourth task. We also balanced the number of times each model was used with a good or bad layout. Finally, the particular sequence of tasks that was used by each participant was selected from the ones not used yet by other participants. There was no pre-defined sequence for ordering participants.

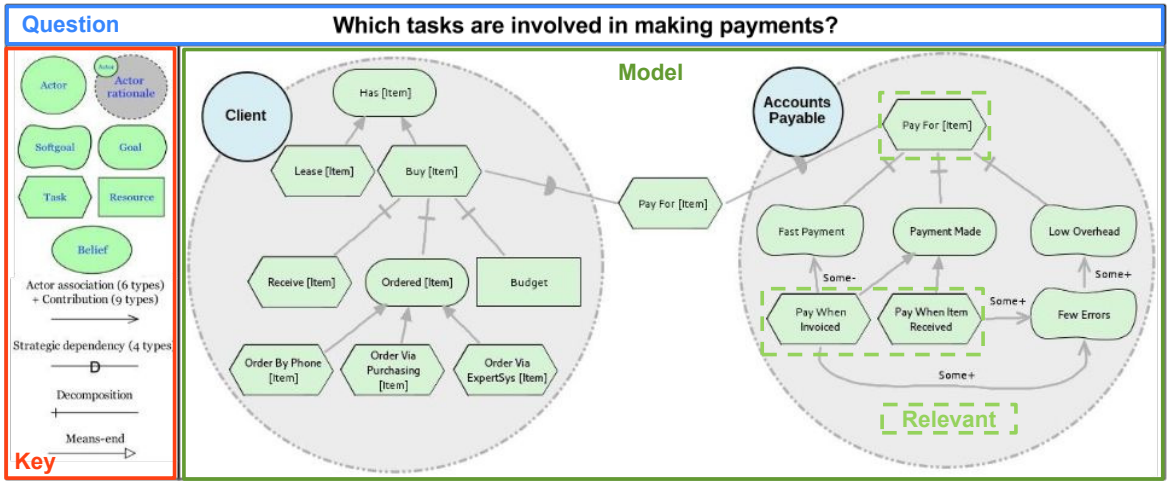


Fig. 1: Example of an understanding task to solve, illustrating the areas of interest

The sequence experienced by each participant is different and illustrated in Table III, where each line represents one participant. The structure is similar to all participants, except in the sequence of tasks (U stands for understand, R stands for review, the number represents the model, G represents a good layout and B a bad layout).

TABLE III: Crossover experimental design

Participant	Letter	Tutorial	T1	T2	T3	T4	Back.
1	✓	✓	U1G	U2B	R3G	R4B	✓
2	✓	✓	R4G	U1B	U2G	R3B	✓
3	✓	✓	U2B	R3G	R4B	U1G	✓
4	✓	✓	R3B	R4B	U1B	U2G	✓
...	✓	✓	...	...	...	...	✓

G. Procedure

We prepared the lab setting so that all participants could have similar conditions. There was only one participant in each evaluation session. We informed the participant that the tasks consisted in watching a short tutorial on a requirements language, analysing requirements expressed in that language, and answering questions about those requirements. We further informed the participants that we would be recording their voice, the contents of the screen, and tracking their eyes movements while they were analysing the requirements and (orally) answering questions about them. Finally, we explained the participants that they could quit at any moment, if they so desired. They then read the *"Participant consent letter"* and gave their free and informed consent to participate in the study.

We helped the participant seating comfortably so that his eyes would be around 50 cm away from the screen. The eye-tracker was placed below the screen, without blocking it. We adjusted the eye-tracker's angle to cope with physical differences among the participants (e.g., the eye-tracker must point towards the subject's eyes, so the participant's height determines the ideal eye-tracker angle). During this process, we explained the procedures, the participant put the headphones, equipped with a microphone, and the session started.

We asked each participant to watch a video tutorial[3] of 7 minutes and 15 seconds, explaining the elements of an *i** model. The tutorial includes the construction of a correct model, similar to those used in the experiment, along with an audio description of both the modelling elements, as they are being introduced, and their role in the model under construction. The modelling elements were described by using the exact phrases and explanations present in the *i** wiki.

After watching this tutorial, we proceeded with the calibration of the eye-tracker, using 9 points. Once the eye-tracker was calibrated, the evaluation session started. Each participant was asked to perform a sequence of four tasks. Each task consisted in either understanding or reviewing an *i** model, and then answering the NASA TLX questionnaire concerning the effort on that task. This was repeated for each of the combinations of good and bad layout with correct and incorrect models. The task (and corresponding model) sequence varied from one participant to the next, as discussed in section III-F. Finally, each participant answered a short questionnaire about demographic information, so that we could better characterise her or him. For each session, we recorded a video with the contents of the screen, synchronised with the voice of the subject during the whole session. We also recorded the 4 NASA TLX sets of answers, one for each task, as well as the answers to the profiling questionnaires.

H. Analysis procedure

We start by collecting descriptive statistics on our variables, namely the *mean*, *standard deviation*, *skewness* and *kurtosis*, to get an overview of their distribution. We are using a crossover design, where our participants apply more than one treatment. This design choice allows controlling the variability among subjects and dealing with the relatively low number of participants (18), but requires controlling the potential

[3]Video tutorial: https://goo.gl/me1jJ2

carryover effect (i.e., the potential persistence of the effect of a treatment when another treatment is applied later). In this case, there is a potential learning effect to consider. Following the guidelines of [20], we use a linear mixed model, which is adequate for mitigating this threat.

IV. Execution

A. Preparation

The data collection was carried out with a core i7 Windows 10 laptop, connected to an external 22 inch, wide screen, full HD monitor, an EyeTribe eye-tracker[4], a headset with a microphone, and an external mouse and keyboard. All data collection was performed in this platform. The experimenter controlled the session on the laptop, while the participant used the eye-tracker and headset microphone to perform the models' analysis, viewing the tasks in the external monitor.

After reading the consent letter, each participant watched the video tutorial on the *i** framework. That was the only source of information on *i** the participant would have for the duration of the experiment, other than an *i** key (always present while the participants were analysing the models). Finally, we recorded the audio and video of the whole section, so that the answers were collected with a "think aloud" approach.

We proceeded with the calibration of the eye-tracker, which consists of having the participant following with her gaze a target as it moves and fixates in predetermined screen coordinates. We used the EyeTribe calibration application, only accepting *good* or *excelent* calibrations (top levels of a 5 points ordinal scale) to proceed to the actual data collection.

B. Deviations

A technical problem with the software for controlling the audio capturing the participants' answers lead to the exclusion of 6 cases. Another situation, where we were not able to determine when the participant started viewing each one of the models, led to the partial exclusion of one case. This can be seen in Table IV, in the lines corresponding to duration, where the total number is 17. Two participants were also excluded from the eye-tracking analysis due to technical problems with the eye-tracker (leaving 16 cases in eye-tracking metrics).

V. Analysis

A. Descriptive statistics

Table IV outlines the descriptive statistics of the collected variables, covering the relative success in the tasks performed by our participants, as well as their perceptions on their performance. Task success was measured by computing the *precision* and *recall* of the answers provided by the participants, the *F-Measure* aggregating precision and recall, and the *duration* of the task in seconds. We further collected the participants' perception of their effort while performing the tasks, through the NASA TLX survey, of which the overall weighted score is also presented in Table IV. For each of these variables, the table is split by task (either understanding (*Und.*), or reviewing (*Rev.*)), layout (*Bad*, or *Good*), number of cases (*#*), mean, standard deviation (*S.D*), skewness (*Skew*), kurtosis (*Kurt*), and the *p-value* of the Shapiro-Wilk normality test, in which the null-hypothesis is that the population is normally distributed. We used an alpha value of 0.05 in the Shapiro-Wilk test. Most variables do not have a normal distribution.

TABLE IV: Descriptive statistics

	Task	Layout	#	Mean	S.D.	Skew	Kurt	S-W
Prec.	Und.	Bad	18	.548	.363	-.461	-1.150	.022
		Good	18	.678	.355	-.828	-.590	.005
	Rev.	Bad	18	.206	.311	1.394	1.036	.000
		Good	18	.178	.341	1.853	2.302	.000
Recall	Und.	Bad	18	.492	.376	.154	-1.348	.021
		Good	18	.622	.344	-.354	-.964	.024
	Rev.	Bad	18	.069	.098	1.031	-.445	.000
		Good	18	.049	.087	1.613	1.405	.000
F-Meas.	Und.	Bad	18	.492	.331	-.296	-1.070	.097
		Good	18	.607	.307	-.688	-.278	.145
	Rev.	Bad	18	.103	.147	1.097	-.247	.000
		Good	18	.073	.130	1.603	1.394	.000
Duration	Und.	Bad	17	216.1	142.8	1.385	1.704	.018
		Good	17	170.1	85.1	.780	-.112	.104
	Rev.	Bad	17	342.1	275.2	2.739	9.031	.000
		Good	17	317.7	217.0	1.376	1.173	.007
TLX	Und.	Bad	18	52.9	20.0	-.671	-.734	.107
		Good	18	50.6	17.5	-.455	-.905	.069
	Rev.	Bad	18	62.4	10.4	-.779	.241	.251
		Good	18	62.5	20.4	-.422	-.308	.881
RelFix	Und.	Bad	16	.096	.111	1.024	-.092	.007
		Good	16	.079	.081	1.174	.729	.011
	Rev.	Bad	NA					
		Good	NA					
IrrelFix	Und.	Bad	16	.169	.093	.470	-1.149	.150
		Good	16	.168	.141	.864	-.615	.023
	Rev.	Bad	NA					
		Good	NA					
AvRelDur	Und.	Bad	16	211.0	138.9	.596	.085	.346
		Good	16	202.2	135.2	1.007	1.314	.193
	Rev.	Bad	NA					
		Good	NA					
AvIrrelDur	Und.	Bad	16	251.8	96.3	.350	-1.169	.182
		Good	16	252.2	98.6	.287	-.488	.896
	Rev.	Bad	NA					
		Good	NA					

The information in Table IV is further illustrated through boxplot diagrams, Figs. 2 to 8, contrasting good and bad layouts for understanding and reviewing tasks. Each figure highlights a different perspective (and corresponding metric). For the reviewing tasks, heat maps are presented in Fig. 9. Fig. 2 presents the *precision*, which is much higher for the understanding task than for the revision task. At a first glance, it seems that precision is slightly better with a good layout than with a bad one, but we will test whether this is significant in section V-C. The distributions seem similar for the review tasks with good and bad layouts. Fig. 3 presents the *recall*. A good layout seems slightly better than a bad one, in the boxplot, for the understanding task, but no difference is observable for the review task. Again, these will be tested later for significance. Fig. 4 presents the *F-Measure*, a combination of *Precision* and *Recall*, that leads to the same observations. Fig. 5 presents the task duration in seconds. There seems to be no clear difference between good and bad layouts, for both tasks. Fig. 6 presents the *NASA TLX*

[4]http://www.theeyetribe.com/

results, showing no significant difference in the perceptions of the overall complexity of the tasks, when contrasting good and bad layouts. Fig. 7 presents both relevant and irrelevant *fixation rates*. Fig. 8 presents the *average fixation duration*. Fig. 9 presents the heat maps with the duration of fixations during reviewing tasks, for both good and bad layout.

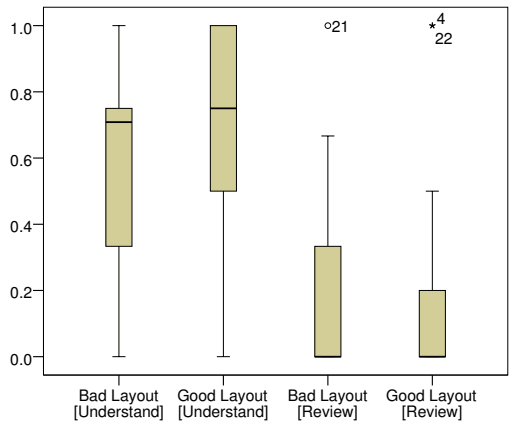


Fig. 2: Results for precision

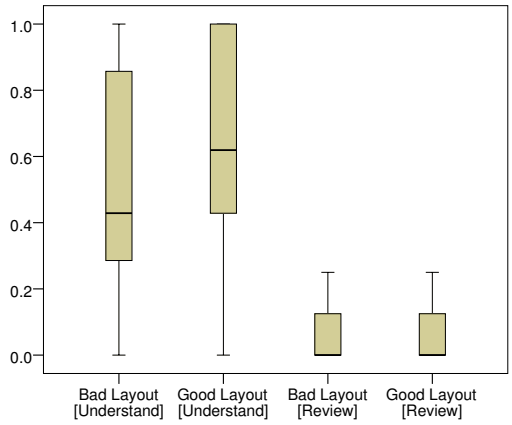


Fig. 3: Results for recall

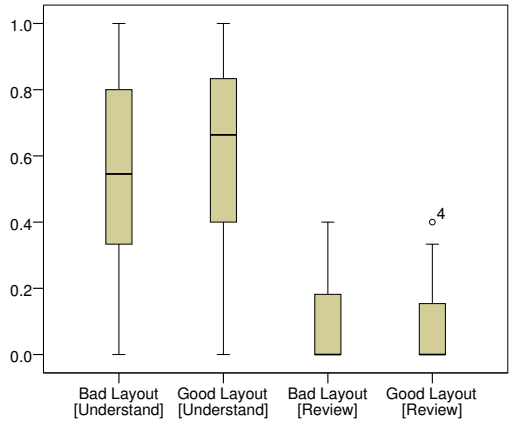


Fig. 4: Results for F-Measure

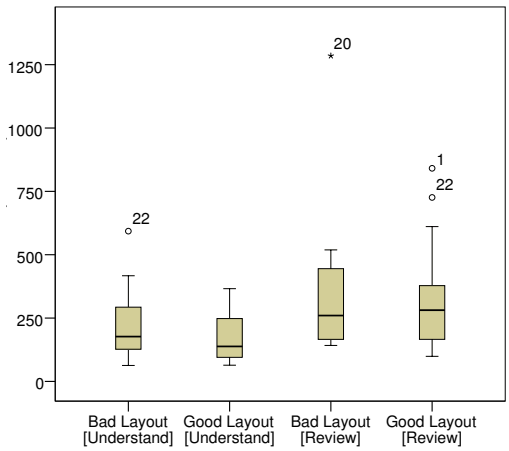


Fig. 5: Results for duration (in seconds)

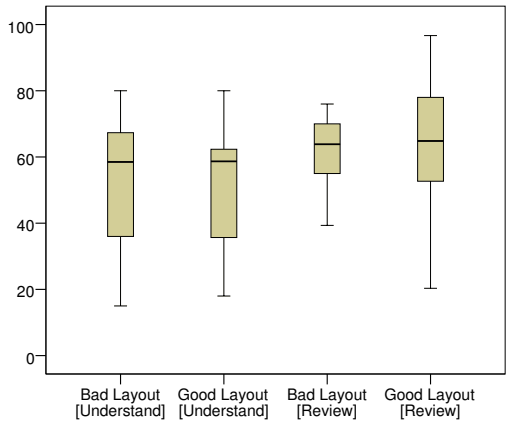


Fig. 6: NASA TLX scores

B. Data set preparation

In each session, we recorded without pausing the video and audio between the models and the NASA TLX, to disturb the participant as little as possible. *Ergo*, we had to manually collect the times when the participant started and ended the visualisation of a given model. Since the answers were given orally, a preparation of that data was also necessary. For the *understanding* tasks, we had a table with all the elements present in the model, one per column. When listening to the answers, elements that a participant described as being the correct ones were marked with 1, in a row dedicated to each participant. For the *reviewing* tasks, the procedure was the same, but when the answer was different from the expected, we added a column with that answer, if it was not already present. In the end, the table contained all the answers given by the participants, and their frequency. Concerning the eye-tracking data, the main areas of the stimulus and its elements were mapped into pixel coordinates to determine which regions and elements the participants were looking at. This allowed tagging the eye-tracking data with the elements being gazed at any given moment, which was a necessary step for computing the eye-tracking metrics used in this paper.

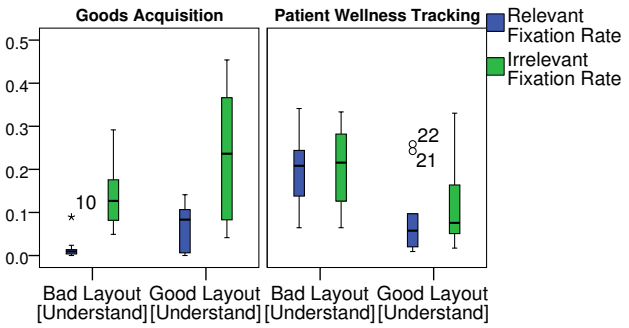


Fig. 7: Fixation rates

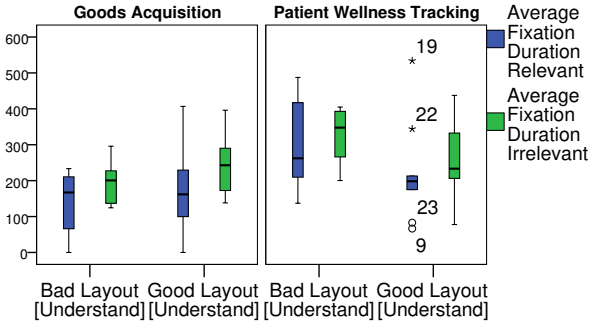


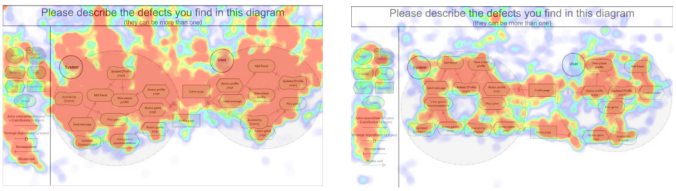
Fig. 8: Average fixation duration

C. Hypotheses testing

We used linear mixed models for testing our hypotheses. The models included the following terms: layout and domain, as fixed factors, and participant as random factor within sequence. Although the residuals of some of these models depart from normality, mixed linear models have been shown robust to violations of the residuals normality assumption [21].

RQ1: Does adherence to layout guidelines influence the understandability of i models?* According to the tests of fixed effects presented in table V, the layout was not a significant factor ($sig. > 0.05$), for any of the dependent variables. However, we did observe significant (but small) differences concerning the impact of the particular domain in three of the eye-tracking metrics. Further, we found no evidence of any influence of the sequence in which the tasks were performed in the results co-variate variance ≈ 0. These results suggest that adherence to layout guidelines has no significant effect on understanding the diagram, on the participants' perception of their performance, or even on the visual effort taken, for the size and complexity of the models used in this evaluation.

RQ2: Does adherence to layout guidelines influence the ability to review of i models?* Similar to RQ1, according to the tests of fixed effects presented in Table VI, neither the layout nor the domain are significant ($sig. > 0.05$) for any of the dependent variables. Further, we found no evidence of any influence of the sequence in which the tasks were performed in the results co-variate variance ≈ 0. These results suggest that adherence to layout guidelines had no effect



(a) Heat map for bad layout (b) Heat map for good layout

Fig. 9: Heat maps for fixations during reviewing tasks

TABLE V: Understanding - Type III Tests of Fixed Effects

Dep.Var.	Source	Num. df	Den. df	F	Sig.
Precision	Intercept	1	16.812	110.272	.000
	Domain	1	16.728	1.973	.178
	Layout	1	16.716	1.394	.254
Recall	Intercept	1	17.946	72.676	.000
	Domain	1	17.955	2.306	.146
	Layout	1	17.956	1.803	.196
F-Measure	Intercept	1	17.924	96.551	.000
	Domain	1	17.817	.000	.984
	Layout	1	17.756	1.315	.267
Duration	Intercept	1	24.027	74.124	.000
	Domain	1	18.233	.017	.899
	Layout	1	17.256	2.243	.152
TLX	Intercept	1	14.682	161.954	.000
	Domain	1	16.003	.348	.563
	Layout	1	16.053	.059	.812
FixRel	Intercept	1	22.140	30.172	.000
	Domain	1	16.406	17.078	.001
	Layout	1	15.674	.707	.413
FixIrrel	Intercept	1	22.874	45.195	.000
	Domain	1	15.686	.770	.393
	Layout	1	14.419	.171	.685
AvDurFixRel	Intercept	1	22.771	62.158	.000
	Domain	1	15.711	9.188	**.008**
	Layout	1	14.501	.142	.712
AvDurFixIrrel	Intercept	1	22.653	197.947	.000
	Domain	1	16.072	6.894	**.018**
	Layout	1	15.034	.009	.924

on the success of the diagram reviewing tasks, and on the participants' perception of their performance, for the size and complexity of the models used in this evaluation.

VI. Discussion

A. Evaluation of results and implications

RQ1: Does adherence to layout guidelines influence the understandability of i models?* Concerning *precision* and looking at only the answers given by the participants, the median value is similar with a good and a bad layout. There is no statistically significant difference between both distributions, so we found no evidence that the impact of the layout is relevant, for that size and complexity of models. Concerning *recall*, the median values are lower than the ones from precision. Although the good layout had a higher median than the bad one for recall (by approximately 20%), the difference in the distributions is not statistically significant.

TABLE VI: Reviewing - Type III Tests of Fixed Effects

Dep.Var.	Source	Num. df	Den. df	F	Sig.
Precision	Intercept	1	16.812	110.272	.000
	Domain	1	16.728	1.973	.178
	Layout	1	16.716	1.394	.254
Recall	Intercept	1	17.946	72.676	.000
	Domain	1	17.955	2.306	.146
	Layout	1	17.956	1.803	.196
F-Measure	Intercept	1	17.924	96.551	.000
	Domain	1	17.817	.000	.984
	Layout	1	17.756	1.315	.267
Duration	Intercept	1	17.000	74.883	.000
	Domain	1	17.000	.140	.712
	Layout	1	17.000	1.923	.183
TLX	Intercept	1	14.682	161.954	.000
	Domain	1	16.003	.348	.563
	Layout	1	16.053	.059	.812

The *F-Measure* results show a higher median value for the good layout (around 10%), but, again the distributions are not significantly different. Overall, in spite of the generally better performance of the good layout in terms of the median value, we found no evidence that the distribution, as a whole, is significantly different when we contrast good and bad layouts. This contradicts our initial expectations, which were based on our intuition, as well as on findings with other modelling languages (see examples in section II-C).

The layout quality does not seem to have a significant impact on the duration for performing the tasks for this size/complexity of models. This was also supported by the hypotheses tests. Hence, a good layout does not seem to contribute to a faster understanding of the model.

Regarding the participants overall perception of effort and complexity, that is practically the same when performing understanding tasks for both good and bad layouts. Thus, participants did not perceive an improvement while using a good layout, when compared with a bad one. This perception is reinforced by the eye-tracking data, which also provided no evidence of a significantly different distribution of visual effort in understanding diagrams with good and bad layouts.

The short answer for RQ1 is that we found no evidence that adherence to layout guidelines had influence in the understandability of the *i** models. As a side note, we also observe that the understanding tasks were much more accessible to our participants than the reviewing tasks. This is not really surprising, as proper reviewing requires both understanding the model being reviewed and knowing the modelling language, to detect when it is misused.

RQ2: Does adherence to layout guidelines influence the ability to review i models?* The reviewing in general had worse results compared to the understanding task. This was expected. Reviewing a model is harder for non-*i** practitioners than just understanding what an *i** model represents. It involves not only reasoning about what the model represents (as in the understanding task) but also about what it does not represent (and should), and what it misrepresents. Our participants clearly struggled with this task.

Concerning *precision*, *recall*, *F-Measure* and *duration*, we found no evidence that the impact of a good *vs.* a bad layout is neither relevant nor significant. The layout seemed to be irrelevant for the performance of our participants in this task. This is consistent with the participants overall perceived effort and complexity of the task, which is practically the same when performing reviewing tasks for both good and bad layouts.

In Fig. 9, we observe that the heat map for good layout is less scattered than the one for bad layout. In the latter, participants spend more time in the model as a whole, while in the former the fixations are more focused in particular model elements. Furthermore, the time participants spend in the key area is higher for links than for other model elements, which hint that participants struggle with this notation. This justifies, in part, the worse results in the reviewing tasks, since most of the problems were related with links.

B. Threats to validity

Conclusion validity. In this experiment, we have a low number of participants to be able to have sound statistical inferences and to reveal a true pattern in the data. Therefore, there is a risk of having erroneous conclusions. For example, although the differences of the distributions with a good and bad layout were not statistically significant for RQ1, for some of those measures the median values with a good layout were better. So, it may be the case that a larger number of participants will lead to the identification of a small, but significant improvement. We plan to mitigate this *low statistical power* threat by replicating this experiment with a higher number of participants. Elements outside the experimental setting might have disturbed the results, such as noise outside the lab. However, this *random irrelevancies in experimental settings* threat was not detected during the course of the experiment nor in the overall results.

Internal validity. Since the participants had 4 tasks to perform, they could have learnt in the process. We tried to mitigate this *maturation* threat by changing the order of the 4 tasks and models from participant to participant. The assignment of the version to a particular participant was random. We have used convenience sampling. To mitigate this *selection* threat we are launching a replication of this experiment with participants selected through a recruitment call, as well as with replications conducted independently by colleagues from other universities and from different countries.

External validity. Since most subjects had little to no prior knowledge in *i**, they are representatives of stakeholders with no requirements engineering expertise. By having participants with a greater level of experience, we could have a representation of stakeholders with requirements engineering expertise, and could analyse the differences between these two profiles. We plan to mitigate this *interaction of selection and treatment* threat by replicating this experiment with a more heterogeneous group in terms of experience with the models. Since the models are not large and have a low complexity level, they may not be representative of models used in industry. However,

the relatively low success in performing the reviewing tasks with these models shows that they were not too simple. Even with this apparently (for requirements experts, at least) simple models, non-requirements experts already found them challenging. Nevertheless, we plan to mitigate this *interaction of setting and treatment* threat by varying the complexity of these models in a future replication, to assess whether there is a significant variation on the success of these tasks as models become more complex. In particular, we expect bad layout to become increasingly a penalising factor as models become more complex and bigger. In this experience, though, we could not use larger models due to technical problems with the eye-tracker device, such as limitations in the external monitor dimensions and distance to the eye-tracker. We need to resolve those technical problems before replicating this experience with more complex models.

Construct validity. Since we have showed a video tutorial about *i**, and afterwards participants answered questions about this modelling language, they might have felt that they were being evaluated. This fact may have caused an *evaluation apprehension* threat, where participants try to look better, which is confounded to the outcome of the experiment. However, since the success level of the tasks was not high, this threat was not detected in the overall results.

C. Inferences

The success level of understanding tasks was, in general, higher than the one in reviewing tasks. As for our research questions, for models of this size and complexity, and for stakeholders with little to none experience with this requirements modelling language, we found no significant difference resulting from using a good, or a bad layout. Further research is necessary to confirm these findings and generalising them to a wider population (both in terms of stakeholders expertise level and of models complexity and size).

VII. Conclusions and future work

We evaluated the impact of adhering, or not, to guidelines for producing good layouts on the understanding and reviewing of *i** models. A series of 4 tasks (2 for understanding and 2 for reviewing) were designed and applied to a set of participants, in order to evaluate the impact of the layout quality. All models had a similar size and complexity. We found that a good or bad layout had no significant impact in the performance of the participants in the assigned tasks. This was somewhat surprising, considering our intuition and evidence found in other experiments conducted with requirements models precisely on the impact of layout in model understanding. It may be the case that the models used in this study were too simple, or too small, for layout to play a significant role, although their size was comparable to that of models used in some of the earlier related studies for other notations. In heat maps for the review tasks, participants' gaze seemed more dispersed by the noise caused by bad layouts.

The evidence collected here suggests that the impact of a good *vs.* a bad layout is not significant, at least for diagrams of this size and complexity. However, we expect layout quality to have a stronger impact as diagrams increase in size and complexity, in line with findings on software models expressed with other languages (e.g. with UML).

We plan to replicate the experiment in other institutional contexts and apply it to bigger and more complex models.

Acknowledgment

We thank NOVA LINCS UID/CEC/04516/2013 and FCT-MCTES research grant SFRH/BD/108492/2015, for the financial support. We also thank A. Karbowy, L. Golebiowski for part of the tool implementation, S. Roldão for the *i** video tutorial, C. Marques for her comments on this paper, and all the volunteers who participated in this evaluation.

References

[1] P. Caire, N. Genon, P. Heymans, and D. L. Moody, "Visual notation design 2.0: Towards user comprehensible requirements engineering notations," in *RE'13*. IEEE, 2013, pp. 115–124.

[2] D. L. Moody, "The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering," *Software Engineering*, vol. 35, no. 6, pp. 756–779, 2009.

[3] H. Störrle and A. Fish, "Towards an operationalization of the "physics of notations" for the analysis of visual languages," in *Model-Driven Engineering Languages and Systems*. Springer, 2013, pp. 104–120.

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

[5] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988.

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

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

[8] H. Störrle, "On the impact of layout quality to understanding uml diagrams," in *VL/HCC, 2011*. IEEE, 2011, pp. 135–142.

[9] S. Yusuf, H. Kagdi, J. Maletic *et al.*, "Assessing the comprehension of uml class diagrams via eye tracking," in *ICPC'07*. IEEE, 2007, pp. 113–122.

[10] B. Sharif and J. Maletic, "The effects of layout on detecting the role of design patterns," in *23rd CSEE&T, 2010*. IEEE, 2010, pp. 41–48.

[11] ——, "An eye tracking study on the effects of layout in understanding the role of design patterns," in *ICSM'10*. IEEE, 2010, pp. 1–10.

[12] B. Sharif, "Empirical assessment of uml class diagram layouts based on architectural importance," in *ICSM'11*. IEEE, 2011, pp. 544–549.

[13] 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.

[14] R. Petrusel and J. Mendling, "Eye-tracking the factors of process model comprehension tasks," in *CAiSE'13*. Springer, 2013, pp. 224–239.

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

[16] 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 *ICPC'13*. IEEE, 2013, pp. 33–42.

[17] A. J. Ko, S. Krishnamurhti, G. C. Murphy, and J. Siegmund, "Human-centric development of software tools," *Dagstuhl Reports*, vol. 5, no. 5, 2016.

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

[19] T. Shaffer and B. Sharif, "Eye-tracking Metrics in Software Engineering," in *APSEC'15*, 2015.

[20] S. Vegas, A. Cecilia, and N. Juristo, "Crossover designs in software engineering experiments: Benefits and perils," *Software Engineering*, vol. 42, no. 2, pp. 120–135, 2016.

[21] W. H. Greene, *Econometric Analysis*, 7th ed. Pearson Education, 2012.