On the Impact of Semantic Transparency on Understanding and Reviewing Social Goal Models

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

***Abstract*—*Context:* *i** is one of the most influential languages in the Requirements Engineering research community. Perhaps due to its complexity and low adoption in industry, it became a natural candidate for studies aiming at improving its concrete syntax and the stakeholders' ability to correctly interpret *i** models. *Objectives:* We evaluate the impact of semantic transparency on understanding and reviewing *i** models, in the presence of a language key. *Methods:* We performed a quasi-experiment comparing the standard *i** concrete syntax with an alternative that has an increased semantic transparency. We asked 57 novice participants to perform understanding and reviewing tasks on *i** models, and measured their *accuracy*, *speed* and *ease*, using metrics of task success, time and effort, collected with eye-tracking and participants' feedback. *Results:* We found no evidence of improved accuracy or speed attributable to the alternative concrete syntax. Although participants' perceived ease was similar, they devoted significantly less visual effort to the model and the provided language key, when using the alternative concrete syntax. *Conclusions:* The context provided by the model and language key may mitigate the *i** symbol recognition deficit reported in previous works. However, the alternative concrete syntax required a significantly lower visual effort.**

***Index Terms*—social goal models, i*, physics of notations, eye-tracking**

I. Introduction

Requirements Engineering (RE) success depends on, among several other factors, the quality of the communication between requirements engineers and other stakeholders. Indeed, communication flaws are among the most frequently reported RE problems that may lead to project failure [1]. One of the key elements of an effective communication is the language used. Visual notations are often adopted, as they are perceived as more effective for conveying information to non-technical stakeholders than text [2]. However, the visual syntax of software engineering languages has historically played a secondary role when comparing alternative visual notations for Software Engineering [3]. The confounding effect potentially played by language syntax is often **not** considered, when comparing languages. In his seminal paper on the *"Physics" of Notations* (PoN) [3], Moody proposed a set of principles to support the evaluation, comparison, improvement and construction of visual notations for Software Engineering. His proposal focused on how to visually represent a set of constructs whose semantics had been previously defined. A core concept, adopted from [4], is the notion of **cognitive effectiveness**, which can be defined as the **accuracy**, **speed**, and **ease** with which a representation can be processed by the human mind. **Semantic transparency**, together with the remaining 8 PoN principles, can lead to cognitive effectiveness. It is defined as *"the extent to which the meaning of a symbol can be inferred from its appearance"* [3].

Several studies were conducted on languages such as UML [5, 6], BPMN [7, 8], KAOS [9] or *i** [10, 11], to identify improvement opportunities for those languages, by detecting problems concerning their concrete syntax and proposing solutions to mitigate them. Those studies focused on the stakeholders' ability to correctly recognise individual language symbols. However, software engineers use **models**, rather than their individual symbols, for communication.

In this paper, our objective is to compare the ability of stakeholders to *understand* and *review* social goal models using two concrete syntaxes: (i) the "official" *i** concrete syntax, and (ii) an alternative *i** concrete syntax, with an increased semantic transparency (that resulted from the series of experiments reported in [11]). In particular, we performed a quasi-experiment to analyse the effect of changing the *i** concrete syntax, evaluating the semantic transparency impact on both the *understandability* and the ability to *review i** models, in the presence of a language key. Differently from previous studies, we perform our evaluation at the model level, rather than through isolated symbol recognition tasks.

A total of 57 novice participants (surrogates for stakeholders other than requirements engineers) performed understanding and reviewing tasks on *i** models. We measured the *accuracy*, *speed*, and *ease* with which they accomplished their tasks. We found no evidence of improved accuracy or speed attributable to the alternative *i** concrete syntax, but found that working with this concrete syntax required significantly lower visual effort. This suggests that the usage of those symbols, in the context of models, and the presence of a language key, may have mitigated the *i** symbol recognition deficit consistently reported in previous works, to the point that it had no observable effect in the accuracy, or speed, with which our participants performed understanding and reviewing tasks.

Section II presents the two concrete syntaxes contrasted in this paper. 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, dataset preparation, and the results of hypothesis testing. Section VI discusses the results and reports threats to validity and inferences. Section VII presents the related work. Finally, Section VIII draws conclusions and points directions for future work.

II. i* STANDARD AND CANDIDATE NOTATIONS

The *i** [12] framework was designed for modelling and analysis of organisational environments and their information systems. *Intentional actor* is the central concept of the approach. Actors are viewed as having intentional properties such as *goals*, *beliefs*, *abilities* and *commitments*. *i** has two main models: the Strategic Dependency (SD) and the Strategic Rationale (SR). The SD model describes the dependency relationships among the actors in an organisational context. An actor (the *depender*) depends on another actor (the *dependee*) to achieve goals and softgoals, to perform tasks and to obtain resources. The SR model focuses on modelling intentional elements and relationships internal to actors.

Although well-known in the RE community, *i** is, as most other requirements languages, poorly understood by novice users [11]. In this paper we explore the extent to which this problem can be mitigated by using an alternative concrete syntax for *i**, with an increased semantic transparency. Research on visual languages design principles and evaluation has the potential for significantly improving the languages' adoption. One of the ways to improve the cognitive effectiveness of a notation is to increase the semantic transparency of its symbols. **Semantic transparency** defines the degree of association between the syntax (form) and semantic (content) of a symbol [3]. However, the *i** language concrete syntax, as described in the *i** Wiki[1], has been shown to be **semantically opaque** [10], as its symbols are abstract geometrical shapes (Fig. 1). In this paper, we will refer to this concrete syntax as "standard" *i**[2].

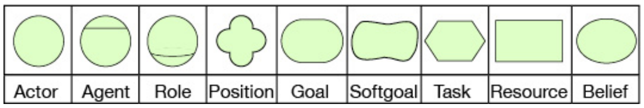


Fig. 1: Standard *i** symbol set

The identification of this shortcoming in the concrete syntax of *i** led to the development of an alternative concrete syntax built upon the PoN principles, to make it more semantically transparent [10]. Later, Caire et al. reported a series of empirical evaluations involving the "standard" *i** and three alternative candidates: *(i)* the symbols designed by experts following the PoN principles [10], *(ii)* the symbols more frequently designed by novices in the context of a symbolisation experiment, and *(iii)* the symbols more frequently chosen by subjects, among those designed by other novices [13].

Those four concrete syntaxes were then tested to determine which symbols were more frequently correctly identified in a blind interpretation experiment to evaluate the semantic transparency and cognitive load involved in recognising the symbols from the various *i** concrete syntaxes. The semantic transparency significantly increased symbol recognition and decreased interpretation errors [11]. The symbol interpretation results were reported in such a way that it allowed us to choose the most frequently recognised symbol for each language construct (i.e., those with the highest semantic transparency coefficient), leading to a proposed *i** notation, referred in this paper as "new" *i** concrete syntax (Fig. 2).

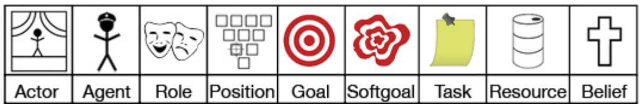


Fig. 2: New *i** symbol set [11]

We selected the best evaluated symbols in Caire et al.'s experiments for each *i** construct [11]. As these symbols were selected independently from each other, they do not necessarily form a consistent set, in terms of the chosen visual metaphors, when compared to what an expert designer would be able to produce. Further research is required to study how an inconsistent set of symbols impacts the overall model understanding. Furthermore, these symbols might be difficult to draw by hand. In this paper, we are only covering model reading, with models built using an *i** editor. Thus, although important, the difficulty in drawing symbols by hand was not an issue in the present study.

III. EXPERIMENT PLANNING

A. Goals

We describe our two research goals following the GQM research goals template [14]. Our first goal (G1) is to *analyse* the effect of changing the *i** concrete syntax, *for the purpose of* evaluation, *with respect to* its semantic transparency 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 our University. Our second goal (G2) is to *analyse* the effect of changing *i** concrete syntax, *for the purpose of* evaluation, *with respect to* its semantic transparency impact on the **review** of *i** SR models, *from the viewpoint of* researchers, *in the context of* an experiment conducted with the same participants.

Because we are comparing the effect of *semantic transparency* of two concrete syntaxes for the same abstract syntax, we can break down each goal into three sub-goals (G1.1, G1.2, G1.3, G2.1, G2.2, and G2.3), concerning the effect of those two concrete syntaxes, in terms of *speed*, *accuracy* and *ease*. So, the refined goals can be obtained by replacing the term *understandability* (or *review*) with *speed to understand*, *effectiveness to understand* and *ease to understand* (or *speed to review*, *effectiveness to review* and *ease to review*).

B. Tasks

Before starting, each participant read and signed a letter of consent, adapted from [15]. Then, they saw a video with

[1] http://istar.rwth-aachen.de/
[2] http://istar.rwth-aachen.de/tiki-index.php?page=iStarQuickGuide

a small tutorial on *i**, covering all the model elements used in this evaluation. There were two versions of this video with exactly the same audio, but with the examples being presented in the standard *i** concrete syntax, or with the new *i** concrete syntax. Naturally, participants saw the video matching the concrete syntax they were about to use, in the evaluation.

Each participant in this study had to complete two tasks: *understanding* an *i** SR model from a Goods Acquisition domain and *reviewing* an *i** SR model from a Tolls System domain. In the *understanding* task, the participant had to analyse a correct *i** SR model and answer a question about it. In the *reviewing* task the participant had to analyse an incorrect *i** model and describe all the defects (s)he could identify. We deliberately introduced syntactic defects in the model. However, we have only informed the participants that their task was to find "defects", since describing explicitly the type of defects they should be looking for would have introduced a bias in the participants attention. This way, each participant was free to review the model using his best judgement, as a stakeholder new to *i** would.

In both tasks, the answers were recorded in audio, and we collected eye-tracking data while the participant was analysing each model. No eye-tracking feedback was visible to the participant, as this would be an unnecessary validity threat to the results. We also did not provided feedback on the extent to which participants were able to successfully complete the tasks, preventing possible contamination to subsequent tasks.

After each evaluation, participants filled in a NASA-TLX questionnaire [16, 17] to collect feedback on his perceptions with respect to the task he had just performed. In the end, each participant provided some basic demographic information.

C. Experimental material

As previously mentioned, the experimental material for this evaluation included a participant consent letter, two video tutorials (one on the standard *i** concrete syntax and another on the *new i** concrete syntax), two versions of the *i** SR model for each of the two tasks (*understanding* and *reviewing*), a NASA-TLX questionnaire, and a demographic questionnaire. To contrast the two alternative concrete syntaxes for *i**, we prepared two versions of each *i** SR model, one with the standard *i** notation and the other with the new *i** concrete syntax. The two versions of the model used for the understanding tasks are presented in Figs. 3a and 3b. The two versions of the model for reviewing tasks are presented in Figs. 3c and 3d. We were very conservative concerning readability. All the elements presented to participants, including textual labels, were comfortably readable in the 22 inch monitor used for conducting this experiment, so that readability would not be an issue. All figures share a common structure, with three Areas Of Interest (AOI): a **question** on top, a **language key** with the elements used in the model on the left side, and a **main area** with the model. For each task, we used a similar layout with both concrete syntaxes so that the only difference among them is the usage of a particular concrete syntax. For each task we annotated two sets of AOIs to analyse eye-tracking data. An AOI is classified as **relevant**, if it contains an element that belongs to the answer of the task, or **irrelevant**, otherwise. No textual descriptions of the scenarios under analysis were offered to the participants, so they had to answer our questions based only on the visual models.

Previous studies on *i** showed that its concrete syntax is semantically opaque, and that the alternative symbols used in this paper were easier to identify [11]. The assumption was that these symbols would improve the understandability of the model by non-experts. However, those stakeholders are likely to examine *i** models in a document or within an *i** editor. In both cases, it is common to have a language key available: it is considered good practice to add a language key in a document, and the editor's toolbar will also serve as a language key, in practice. As such, the presence of a language key is aligned to common practice.

All the materials used in this evaluation, can be found in the paper's companion site[3].

D. Participants

This evaluation was performed by 57 participants selected by convenience sampling. Most of them are students at different levels at our University. The main research question of this paper is whether improving the *i** concrete syntax has a real impact on its understandability by stakeholders other than Requirements Engineers. The latter may be specially trained to use this concrete syntax. Our target population is therefore non-experts, making our subjects better surrogates for this type of stakeholders than experienced RE practitioners. Students are often used as surrogates for practitioners in software engineering experiments [18, 19] and have shown to be a valid option in those experiments [20, 21].

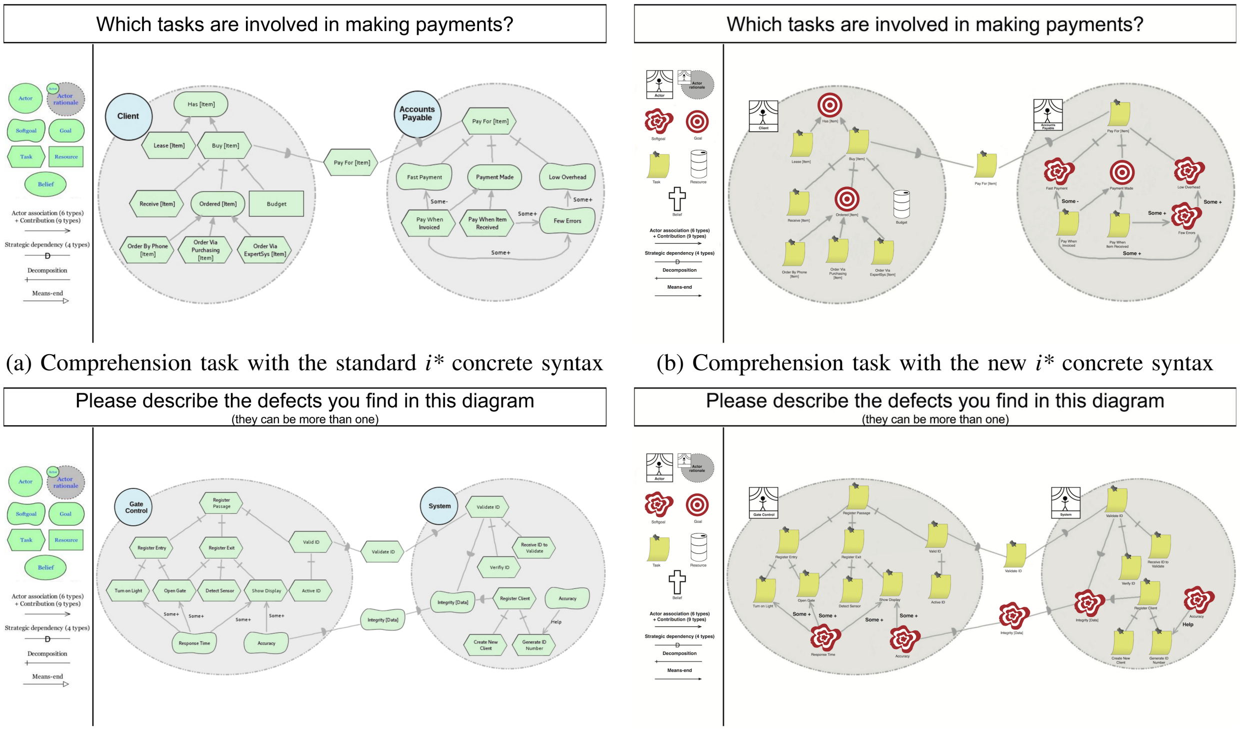
Of the total of participants, 27 were tested with the standard *i** concrete syntax, while the remaining 30 used the new concrete syntax. No participant was tested for both versions of the language, as a learning effect from one evaluation to the next could represent a confounding effect. Our goal was to have participants with a similar background testing both versions of the language. None of the participants who used the new concrete syntax participated in the evaluation of the standard concrete syntax, and *vice-versa*.

For each participant, we collected demographic data on previous *experience* with *i**, *age*, *highest completed level of education*, *current occupation* (student, researcher, or practitioner), *field of studies*, *gender*, *nationality*, and *usage of reading devices* (glasses, or contact lenses).

Regarding previous *experience* with *i**, in the group assigned to the standard concrete syntax, there was 1 participant who had used *i** in a professional context, 5 who learnt it in the context of a course, and the remaining 21 had no previous contact with *i**. For the new concrete syntax, 3 had learnt *i** in the context of a course and 27 did not know it.

Concerning participants *age* distribution, the assumption of normality is **not** reasonable as shown by a Shapiro-Wilk test

[3] https://sites.google.com/view/istarconcsyntaxexperiment/home



(a) Comprehension task with the standard *i** concrete syntax

(b) Comprehension task with the new *i** concrete syntax

(c) Review task with the standard *i** concrete syntax

(d) Review task with the new *i** concrete syntax

Fig. 3: Understand and review tasks proposed to participants

conducted on each of the participants groups ($p < 0.001$, in both cases), and confirmed by the visual inspection of boxplots, Q-Q plots and kernel density plots, omitted here for the sake of brevity. We then used the Welch t-test to test if there was a statistically significant difference between the age distribution in the two groups. The Welch t-test is robust when the group variances are unequal, and even if the sample sizes are unequal, as well as to departures from normality in the data. There was no statistically significant difference between the ages of participants using the standard $i*$ concrete syntax ($M = 26.38$, $SD = 5.933$) and those using the new concrete syntax ($M = 24.63$, $SD = 6.451$; $t(1) = 53.792$, $p = .295$).

With respect to the *highest completed level of education*, all participants had some university level training. For those tested with the standard concrete syntax, 5 completed high school, 13 had BSc degrees, 8 had an MSc degree and 1 a PhD degree. For those participating in the new concrete syntax evaluation, 7 completed high school, 18 had a BSc degree, 4 an MSc degree, and 1 a PhD degree. Concerning current occupation, the standard concrete syntax had 1 researcher, 3 practitioners, and the remainder were students; the new concrete syntax, had 1 researcher, 1 practitioner and the remainder were students.

Regarding *nationality*, 21 Portuguese, 4 Brazilians, 1 Croatian and 1 Spaniard used the standard concrete syntax, and 29 Portuguese and 1 Brazilian used the new concrete syntax.

Concerning the *field of studies*, for the standard concrete syntax, 25 were computer scientists, and 2 were industrial engineering. For the new concrete syntax, we had 22 computer scientists, 2 industrial engineers, 2 architects, 1 mechanical engineer, 1 manager, 1 civil engineer and 1 lawyer. For each concrete syntax, there were 4 female participants. The remainder were male. In terms of the usage of *reading devices*, 3 participants using the standard concrete syntax and 1 using the new concrete syntax had contact lenses while, in each group 4 and 8, respectively, wore glasses.

E. Hypotheses, parameters and variables

For each of the two high level goals, we define the null (H_0) and alternative hypotheses (H_1).

$H_{0Understand}$: Changing from a semantically opaque concrete syntax (standard *i**) to a more semantically transparent one (new *i**) does not influence *i** SR models *understandability*.
$H_{1Understand}$: Changing from a semantically opaque concrete syntax (standard *i**) to a more semantically transparent one (new *i**) influences *i** SR models *understandability*.

This hypothesis is further refined to cope with *accuracy*, *speed* and *effort*. For example:

$H_{0UnderstandAcc}$: Changing from a semantically opaque concrete syntax (standard *i**) to a more semantically transparent one (new *i**) does not influence *i** SR models *understanding accuracy*.
$H_{1UnderstandAcc}$: Changing from a semantically opaque concrete syntax (standard *i**) to a more semantically transparent one (new *i**) influences *i** SR models *understanding accuracy*.

And similarly for speed and ease of *understanding*. We follow the same approach and refine the *null* and the *alter-*

native hypotheses in the case of *review* into 3 sub-hypotheses, corresponding to *accuracy*, *speed* and *effort*. The independent variable is the *concrete syntax*, which may be *standard*, or *new*. The dependent variables are the same for both hypotheses, as well as their corresponding refined sub-hypotheses.

Assessing accuracy. The accuracy achieved by our participants is assessed by their responses with respect to their *precision*, and *recall*, using the following metrics:

- *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.

Higher values of *Precision*, *Recall*, and the *F-measure*, support the claim of a better accuracy.

Assessing speed. The speed achieved by our participants is assessed by several time-related indicators. We are interested not only in the overall response time, but also on the time it takes participants to provide valid answers. We will assess *speed* using the following metrics:

- *Duration* – the time taken by the participants to complete the task.
- *FirstDet* – First Detection; the time taken to accurately report the *first* response element; for the understanding task, this is the time for correctly reporting the first element that answers the question enunciated in the task; for the reviewing task, this is the time taken to report the first seeded defect in the model. If a participant does not correctly report at least one element, this metric will be treated as a missing value and removed from all further analysis procedures.
- *LastDet* – Last Detection; the time taken to accurately report the *last* response element; this is the dual for the *FirstDet* metric.

Lower values of these metrics support the claims of superiority of the corresponding concrete syntax with respect to its cognitive effectiveness in terms of improving the speed with which the models are understood and reviewed. While the overall *duration* addresses the time spent in the task, the other two metrics provide a detailed picture of the moment when the participant starts and ends providing valid feedback.

Assessing ease. The ease with which participants conduct their tasks is assessed by effort measures. Although time measures (as those we used for speed) are often used as proxies for effort, in the context of the *"Physics" of Notations* these are better matches for the speed component, which is likely to strongly correlate to ease. Instead, we focus our assessment in two information sources: the physical (*visual*) effort involved in exploring the model and the *perception of effort* reported by participants. The former is addressed with eye-tracking measurements, while the latter is assessed through a NASA-TLX questionnaire. We will consider the following metrics:

- *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). A fixation is a stabilisation of the eye on a part of the stimulus for a period of time between 200 and 300 ms.
- *FixIrrel* – Fixation Rate on Irrelevant elements; the fraction of number of fixations in an given AOI over the total number of fixations in the AOG.
- *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 irrelevant AOIs over the number of elements of the irrelevant AOIs.
- *TotSac* – total number of saccades while performing the task. A saccade is a sudden and quick eye-movement lasting between 40 to 50 ms.
- *Sac2Key* – number of saccades to the key AOI.
- *NASA-TLX score* – overall weighted score resulting from the application of the TLX questionnaire, covering perceived mental, physical and temporal demand, performance, effort and frustration for performing a task.

A higher number and duration of fixations is associated with a higher visual attention in a given set of AOIs (in this case, relevant *vs.* irrelevant model elements) [22–24]. For understating tasks, a higher Fixation Rate indicates higher efficiency associated with less effort to find the relevant AOIs [24–28]. As for reviewing tasks, a higher ratio indicates more visual effort to find defects [26, 29]. Regarding the Average Fixation Duration, a higher value indicates more time and attention devoted to AOIs [23, 25, 30], some state this ratio is correlated with cognitive processes [31, 32]. A higher number of saccades can be associated with a higher visual effort, meaning the participant may be somewhat "lost" in the model, making a more erratic model navigation [23, 28, 32, 33]. A higher number of saccades to the key can also be associated with difficulties with the concrete syntax. Concerning the NASA-TLX score, higher scores are associated with a higher perceived effort by the participants [17, 33]. Both for the eye-tracking and the NASA-TLX metrics, lower complexity will correspond to higher *ease* in performing the tasks.

F. Design

Data collection was performed in two different moments, one for each concrete syntax. Due to participants availability constraints, most of those using the standard *i** concrete syntax performed only either the understanding or the review task. Three of them performed both. As for the participants using the new *i** concrete syntax, they all performed both tasks. To reduce learning effects, for those performing 2 tasks, the relative order of those tasks changed from one participant to the next. Each participant worked only with one of the concrete syntaxes for *i**. We balanced the number of times each task was performed before, or after the other task.

The sequence experienced by each participant is illustrated in Table I, where each line represents a set of participants that followed a particular sequence of activities. T# refers to the task number and Back to the background questionnaire (demographic data). The tasks are encoded: concerning the first character, **U** stands for **u**nderstand, while **R** stands for **r**eview; the second character represents the particular concrete syntax used by that participant, where **S** stands for **S**tandard *i** concrete syntax and **N** stands for **N**ew *i** concrete syntax. There was no pre-defined sequence for ordering participants.

TABLE I: Experimental design

#Participants	Letter	Tutorial	T1	TLX	T2	TLX	Back
13	✓	✓	US	✓			✓
11	✓	✓	RS	✓			✓
2	✓	✓	US	✓	RS	✓	✓
1	✓	✓	RS	✓	US	✓	✓
15	✓	✓	UN	✓	RN	✓	✓
15	✓	✓	RN	✓	UN	✓	✓

The statistical analysis performed (Welch *t*-test) is robust concerning the different sample sizes, that is, a different number of participants performing each sequence.

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 him 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 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 sit 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 (the eye-tracker must point towards the subject's eyes, so the participant's height determines the ideal eye-tracker angle). The participant put on the headphones (equipped with a microphone), and the session started.

We asked each participant to watch a video tutorial 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, and 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 using the exact phrases and explanations present in the *i** wiki. At the end of this tutorial, we calibrated the eye-tracker, and started the evaluation session. Each participant was asked to perform a sequence of two 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 task. The task (and corresponding model) sequence varied from one participant to the next (discussed in Section III-F). Finally, each participant answered a short questionnaire about demographic information. 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 NASA-TLX sets of answers, one for each task, and the answers to the demographic questionnaire.

H. Analysis procedure

We collected descriptive statistics on our variables, namely the *mean*, *standard deviation*, *skewness* and *kurtosis*, to get an overview of their distribution. This was complemented with kernel density plots to help with the visual analysis of those distributions. Kernel density plots provide a more detailed picture of a distribution, when compared to boxplots, and are a better fit for comparing distributions in Software Engineering experimentation. This visual analysis was then complemented with Welch *t*-tests, which provide an alternative to the t-test, as they can robustly handle non-normal distributions, with different sample sizes and variances. Section V shows that the vast majority share these properties. A detailed discussion on the benefits of using kernel density plots *vs.* box plots, and using Welch *t*-test for comparing distributions in a robust way (as opposed to two samples t-test, or a non-parametric alternative to it, such as the Mann-Whitney U test) is in [34].

IV. Execution

A. Preparation

The data collection was carried out with a laptop connected to an external 22 inch, wide screen, full HD monitor, an EyeTribe eye-tracker[4], a set of headphones with a microphone, and an external mouse and keyboard. The experimenter controlled the session on the laptop, while the participant used the eye-tracker, headphones and microphone to perform the models' analysis, viewing the tasks in the external monitor. Each participant started by reading a consent information letter, then 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** language key (see Fig. 3d).

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 *excellent* calibrations (top levels of a 5 points ordinal scale) to proceed to the actual data collection.

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

B. Deviations

During the standard *i** concrete syntax experiment, we observed a technical problem with the software for audio capturing, leading to the exclusion of a total of 6 cases (4 Rev. + 2 Und.). This can be perceived in Table II, on sector for accuracy metrics where the number of participants is 12 for both tasks (instead of 14 and 16). Another situation in this same experiment, where we could not determine when the participant started viewing each of the models led to the partial exclusion of 1 case (Und.). In addition, a technical problem with the eye-tracker device led to the exclusion of 2 cases for the understanding task, one for each concrete syntax. This can be observed in Table II on the sectors for speed and visual effort metrics, where the total number for the understanding task is 10 and 29 instead of 12 and 30, respectively.

V. Analysis

A. Descriptive statistics

Table II presents the descriptive statistics for the metrics collected in our data analysis, (introduced in section III-E). For each metric we present 4 lines in the table. The first 2 refer to the understanding task, while the other 2 refer to the reviewing task. In the *Syntax* column we specify which of the syntaxes we are considering (*Stand.* represents the standard *i** concrete syntax, while *New* represents the new concrete syntax. We further present the mean, standard deviation, skewness, kurtosis, and the *p-value* for the Shapiro-Wilk normality test. The number of participants is not always the same for all metrics (as per section IV-B), and missing values were excluded from the analysis due to anomalies in the data collection process (e.g., a situation when no valid elements were detected, it made no sense to compute the corresponding metric; this is particularly noticeable in the review task).

The metrics are visually grouped to reflect the three components of cognitive effectiveness: *accuracy*, *speed* and *ease*. One of the most noteworthy features of our data set is that the shape of the distributions concerning variables related to *accuracy* and *speed* suggests that, in general, normality is **not** a reasonable assumption (*p-value* < 0.05). Several of the metrics concerning *ease* do have a distribution suggesting normality is a reasonable assumption (*p-value* ≥ 0.05). The variance of the distributions is not similar, for several of these variables. The visual inspection of boxplot diagrams, Q-Q plots and kernel density plots (ommitted here for the sake of brevity) further reinforced our assessment concerning data normality.

B. Data set preparation

In each session, we recorded without pausing the video and audio. The NASA-TLX questionnaire was answered directly online. During the data collection process, we took special care not to disturb, or distract, our participants. We manually collected 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. At 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.

TABLE II: Descriptive statistics

	Task	Syntax	#	Mean	S.D.	Skew	Kurt	S-W
Prec.	Und.	Stand.	12	.653	.325	-.638	-.115	.146
		New	30	.525	.291	.158	-1.075	.048
	Rev.	Stand.	12	.089	.215	2.363	4.881	.000
		New	30	.131	.271	2.135	3.955	.000
Recall	Und.	Stand.	12	.722	.446	-1.181	-.584	.000
		New	30	.678	.309	-.347	-1.172	.000
	Rev.	Stand.	12	.048	.111	2.055	2.640	.000
		New	30	.067	.190	4.390	21.296	.000
F-Meas.	Und.	Stand.	12	.615	.415	-.713	-1.241	.011
		New	30	.573	.280	-.175	-1.216	.026
	Rev.	Stand.	12	.061	.143	2.100	2.974	.000
		New	30	.079	.189	3.515	14.577	.000
Duration	Und.	Stand.	10	131.9	90.0	2.291	5.850	.001
		New	29	163.8	111.1	1.572	2.698	.001
	Rev.	Stand.	12	255.3	179.4	1.755	3.815	.014
		New	30	263.9	143.9	.734	-.177	.059
FirstDet	Und.	Stand.	10	120.2	103.0	2.138	5.340	.007
		New	28	106.1	75.9	2.015	4.969	.000
	Rev.	Stand.	2	174.5	137.9	-	-	-
		New	7	192.0	234.0	2.327	5.633	.002
LastDet	Und.	Stand.	10	126.3	104.2	2.037	4.941	.013
		New	28	113.6	78.8	1.906	4.385	.000
	Rev.	Stand.	2	217.0	89.1	-	-	-
		New	7	204.6	227.0	2.373	5.860	.001
RelFix	Und.	Stand.	10	.127	.158	.701	-1.614	.004
		New	29	.135	.095	1.081	.470	.005
	Rev.	Stand.	12	.086	.049	.257	.073	.984
		New	30	.026	.031	1.741	3.205	.001
IrrelFix	Und.	Stand.	10	.293	.201	.212	-1.305	.373
		New	29	.264	.259	.114	.-.898	.559
	Rev.	Stand.	12	.282	.100	-1.019	1.344	.376
		New	30	.370	.112	-.232	-.337	.573
AvRelDur	Und.	Stand.	10	174.0	213.2	.671	-1.464	.008
		New	29	327.7	153.7	.322	-.392	.832
	Rev.	Stand.	11	323.2	134.0	.800	.056	.184
		New	22	274.7	205.2	1.775	4.478	.003
AvIrrelDur	Und.	Stand.	10	305.1	98.0	.476	-.432	.724
		New	29	289.2	114.6	.385	-.732	.412
	Rev.	Stand.	12	238.0	57.3	.127	-1.976	.039
		New	30	292.7	103.4	1.795	5.056	.001
TotSac	Und.	Stand.	10	47.7	12.6	.139	-1.795	.035
		New	29	41.7	21.5	.266	-.587	.459
	Rev.	Stand.	12	460.0	324.9	1.653	3.314	.027
		New	30	458.8	256.5	.568	-.557	.128
Sac2Key	Und.	Stand.	10	165.4	28.7	-.959	.057	.174
		New	29	117.6	55.8	-.910	-.006	.008
	Rev.	Stand.	12	83.75	51.9	-.142	-.753	.795
		New	30	141.3	52.0	-1.587	2.724	.000
TLX	Und.	Stand.	16	47.7	12.6	.139	-1.795	.035
		New	29	41.7	21.5	.266	-.587	.459
	Rev.	Stand.	13	54.7	22.4	.027	-1.451	.469
		New	30	58.7	20.1	-.256	-.395	.895

C. Hypotheses testing

For testing our hypotheses, we used the Welch *t*-test, instead of the t test, as it is robust to deviations from the normal distribution, different sample sizes and variance in the samples, thus following the recommendations on data analysis for Software Engineering empirical evaluations [34] (which summarises best practices in statistical analysis on other domains).

RQ1: Does the adoption of a more semantically transparent concrete syntax improve the accuracy, speed and ease when performing understanding tasks on i SR models?* Table III summarises the Welch *t*-test results for the *Understand* task. There was a statistically significant difference (see Table II) between the total number of saccades (*TotSac*) made by participants while understanding the model represented with the standard concrete syntax ($M = 47.7$, $SD = 12.6$) and those using the new concrete syntax ($M = 41.7$, $SD = 21.5$; $t(1) = -3.247$, $p = .007$). This suggests a lower visual effort when using the new concrete syntax, in terms of saccades. A similar conclusion can be drawn concerning the number of saccades to the AOI where the language key to the concrete syntax was presented, with a statistically significant difference between the distribution with the standard concrete syntax ($M = 165.4$, $SD = 28.7$) and the new concrete syntax ($M = 117.6$, $SD = 55.8$; $t(1) = 3.469$, $p = .002$). Both variables are related to the *ease* component. We found no statistical evidence of differences concerning the remaining variables. Figs. 4a and 4b illustrate the heat maps representing the areas more frequently gazed during the understand tasks, with the standard and new concrete syntax, respectively.

TABLE III: Welch *t*-test scores for the *Understand* task

Metric	Statistic	df1	df2	Sig.
Precision	1.186	1	18.472	.251
Recall	.316	1	15.428	.756
F-Measure	.323	1	15.179	.751
Duration	-.907	1	19.254	.376
FirstDet	.397	1	12.669	.698
LastDet	.352	1	12.880	.730
FixRel	-.149	1	11.299	.884
FixIrrel	.437	1	11.155	.670
AvRelDur	.259	1	5.302	.806
AvIrrelDur	.422	1	18.198	.678
TotSac	-3.247	1	12.089	**.007**
Sac2Key	3.469	1	30.938	**.002**
NASA TLX	1.399	1	42.780	.243

RQ2: Does the adoption of a more semantically transparent concrete syntax improve the accuracy, speed and ease when performing reviewing tasks on i SR models?* Table IV summarises the Welch *t*-test results for the *Review* task. Again, there was a statistically significant difference in several variables concerning the *ease* component of cognitive effectiveness, when contrasting the number of relevant fixations, the number of irrelevant fixations, the average duration of irrelevant fixations and the number of saccades to the key. Some of the eye-tracking *ease* metrics suggest a lower complexity (i.e., an easier experience) when using the standard concrete syntax. Others, suggest the opposite. Specifically, the number of relevant fixations using the standard concrete syntax ($M = .86$, $SD = .49$) was higher than the one when using the new concrete syntax ($M = .026$, $SD = .031$; $t(1) = 3.935$, $p = .001$). The number of irrelevant fixations has raised from ($M = .282$, $SD = .100$) to ($M = .370$, $SD = .112$; $t(1) = -2.507$, $p = .020$). The average duration of fixations to irrelevant parts of the model was lower with the standard concrete syntax ($M = 238.0$, $SD = 57.3$) than with the new concrete syntax ($M = 292.7$, $SD = 103.4$; $t(1) = -2.178$, $p = .036$). Finally, the number of saccades to the language key was lower with the standard concrete syntax ($M = 83.8$, $SD = 59.1$) than with the new concrete syntax ($M = 141.3$, $SD = 52.0$; $t(1) = -3.244$, $p = .004$). We found no other statistically significant differences concerning the remaining variables. Figs. 4c and 4d illustrate the heat maps representing the areas more frequently gazed during the review tasks, with the standard and new concrete syntax, respectively.

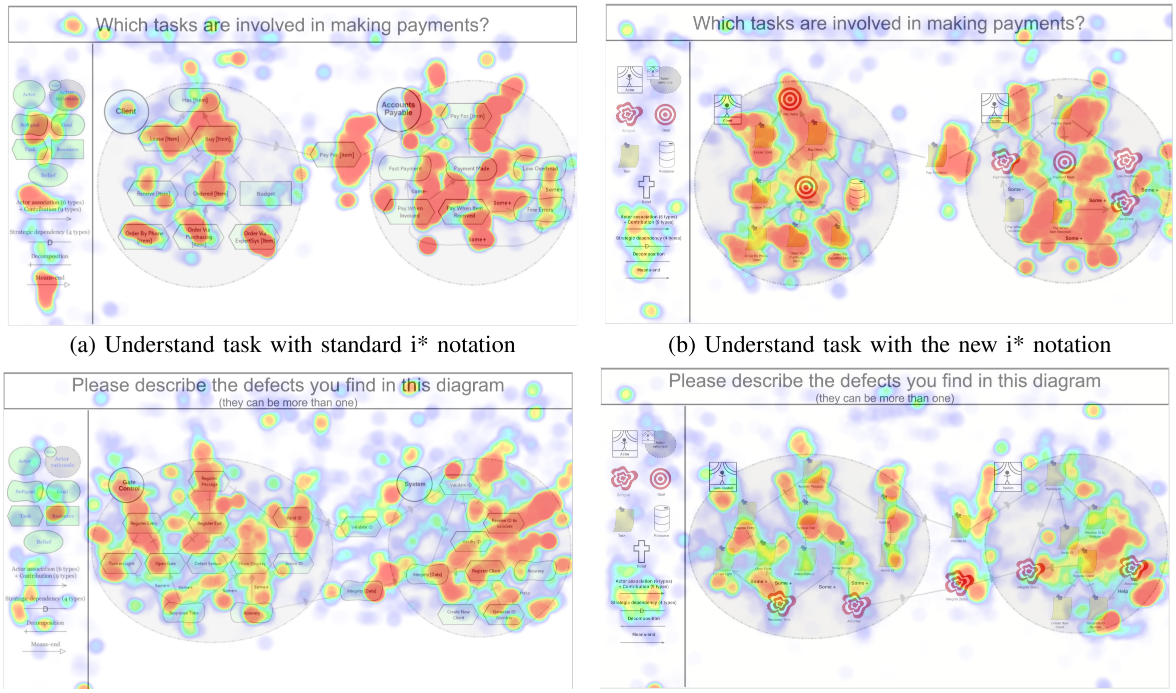
TABLE IV: Welch *t*-test scores for the *Review* task

Metric	Statistic	df1	df2	Sig.
Precision	.000	1	16.512	.983
Recall	.802	1	27.868	.378
F-Measure	.152	1	20.143	.701
Duration	-.148	1	16.967	.884
FirstDet	-.133	1	2.985	.903
LastDet	.117	1	5.181	.911
FixRel	3.935	1	14.769	**.001**
FixIrrel	-2.507	1	22.784	**.020**
AvRelDur	.813	1	28.519	.423
AvIrrelDur	-2.178	1	35.492	**.036**
TotSac	.011	1	16.775	.991
Sac2Key	-3.244	1	20.370	**.004**
NASA TLX	.296	1	20.851	.592

VI. Discussion

A. Evaluation of results and implications

RQ1: Does the adoption of a more semantically transparent concrete syntax improve the accuracy, speed and ease when performing understanding tasks on i SR models?* We found no evidence of improvements brought by the adoption of the new *i** concrete syntax, in terms of the accuracy and speed with which our participants performed their *understanding* task. The only statistically significant difference observed when contrasting the performance of users with each of the concrete syntaxes conveyed a greater visual effort while using the standard notation, observable through a higher number of saccades in general, and a higher number of saccades targeting the language key, on the left hand side of the screen. Both seem to convey a higher difficulty in using the standard *i** concrete syntax. That said, the level of success and overall time taken to perform the task are similar, regardless of the particular concrete syntax. Our interpretation is that, even if the particular concrete syntax created some extra difficulties, these must have not been significant. In fact, the self reported perception of the complexity of the task, through the NASA-TLX questionnaire, supports the interpretation that participants evaluated difficulty similarly, in both cases.



(a) Understand task with standard i* notation

(b) Understand task with the new i* notation

(c) Review task with standard i* notation

(d) Review task with the proposed i* notation

Fig. 4: Heat maps for the understand and review tasks in both notations

RQ2: Does the adoption of a more semantically transparent concrete syntax improve the accuracy, speed and ease when performing reviewing tasks on i SR models?* As with *RQ1*, we found no evidence of the benefits of the new concrete syntax, when compared to the standard, in terms of speed, or accuracy. Again, there were some differences in terms of visual effort. While the effort spent looking at the relevant parts of the model decreased, the effort on looking at irrelevant parts of the model increased, with the new notation. Similar to what we observed for *RQ1*, the feedback provided by the participants through the NASA-TLX questionnaire suggests that, if indeed there was an effort difference, the participants did not notice it.

B. Threats to validity

Conclusion validity. Although we have a reasonable number of participants, higher than most sample sizes reported in other eye tracking experiments (see [23]), sample size is a risk, as the results may not apply to larger populations. We plan to extend this study by performing replicas, and we facilitate independent replicas to independent teams, by sharing the materials used in this work.

Internal validity. The potential learning effect for participants from one task to the next was mitigated by assigning the tasks to participants in a way that those starting with the understanding task and those starting with the reviewing task were balanced. In addition, participants using one concrete syntax were **not** using the other one. We found no evidence of learning effects in the data. Finally, special care was taken to guarantee that all the materials produced were easily readable in the 22 inch monitor used for the experiment. We were limited by the technical specifications of the eye-tracker device, such as limitations in the external monitor dimensions and distance to the eye-tracker. The fonts and symbols used had to be big enough for easy visualisation by all participants. As such, the tested models are fragments of larger models. Notwithstanding, presenting only model fragments to focus the attention of the stakeholders is a common technique for improving communication with them. Moreover, our results show that the tasks were already challenging for our participants, with this model size. We need to resolve those technical limitations before the replication with bigger models.

External validity. Overall, our participants had little to no prior knowledge in *i**, making them good surrogates for non-expert stakeholders (our target population). Further research is needed to assess how these changes in concrete syntax would impact experienced Requirements Engineers. Also, the models used in our evaluation are neither representative of all possible alternative concrete syntaxes nor of all *i** SR models.

Construct validity. Since we have showed a video tutorial about *i**, and afterwards participants answered questions about *i** models, they might have felt that they were being evaluated. This may have caused an evaluation apprehension threat, where participants try to look better, which is confounded to the outcome of the experiment. To mitigate this threat, we have not informed the participants about what was being tested, i.e., their accuracy, speed and ease in the performed tasks.

C. Inferences

Inferences are discussed contrasting the results of the standard concrete syntax with the new *i** concrete syntax.

Similar speed and accuracy. Our results suggest that for *i** models of the complexity used in this evaluation there was no observable benefit in the *speed* and *accuracy* with which the participants were able to conduct their tasks. Two possible explanations for cancelling the effect of using symbols with a greater semantic transparency are: (1) the presence of a language key that facilitates the interpretation of the symbols in such a way that the higher semantic transparency of the new concrete syntax has no effect in the results, and (2) the results were mostly influenced with the difficulties of our participants with semantic aspects of the models rather than with syntactic ones. Concerning *ease*, we did find some indicators of eye-tracking suggesting different visual efforts. Further research is necessary to assess how consistently these results occur with other users, models and concrete syntaxes.

No deep overall impact of visual effort. The visual effort is lower for the new concrete syntax, as participants seem a bit more "lost" with the standard concrete syntax, making a more erratic model navigation (see the more scattered heat map footprint in Fig. 4). However, this was not perceived as a shortcoming by the participants. They were not even aware that they struggled more with the navigation, as suggested by their answers to the NASA-TLX questionnaire. Thus, although navigating in the standard concrete syntax models was visually harder, this had no practical impact in their overall performance. If the tasks were longer or in a higher number, though, the results could have been different, due to fatigue. This should be explored in subsequent studies.

Better symbol semantic transparency did not imply better model understanding. This is somewhat in line with the findings in [35], reporting that the application of the PoN theory is complex, often leading to sub-optimal concrete syntaxes proposals and evaluations. Even when the semantic transparency of the concrete syntax significantly improves, this does not necessarily translate into better performance when using the models, due to the context provided by the model, and, when available, the presence of a language key. Furthermore, semantic transparency is just one of the 9 PoN principles. Hence, we suggest that future studies consider syntactic improvements based on more than a single PoN principle. Plus, more realistic scenarios should be considered.

VII. Related work

Several studies were performed upon different modelling languages, particularly UML, BPMN, and some goal-oriented languages, such as *i** and KAOS. These studies aim at detecting problems concerning the languages' concrete syntax by using the PoN set of principles, and propose solutions to mitigate them. Moody et al. [5] propose several improvement recommendations for the concrete syntax of diagrams defined in UML 2.0 while Kouhen et al. [6] evaluate UML with a set of experiments and report on its lack of semantic transparency. Genon et al. [7] evaluate the cognitive effectiveness of the BPMN 2.0 concrete syntax, and Moody [8] identifies in BPMN serious issues that may hinder its usability and effectiveness in practice, particularly for communicating with end users. Regarding goal-oriented approaches, Moody et al. [10] analyse the cognitive effectiveness of *i**, Caire et al. [11] propose an approach to designing concrete syntaxes, demonstrated with *i**, that actively involves novice users in the process. Matulevičius et al. [9] evaluate how KAOS and Objectiver, its tool, help the modelling activity, offering recommendations for modellers, language designers and tool developers.

Störrle [36] studies the impact of the usage of good *vs.* bad diagram layouts on model comprehension tasks when using UML, in particular use cases, class, and activity diagrams. On a similar research line, Santos et al. [37] evaluate the effect of the layout guidelines on the *i** models understandability, by using eye-tracking. Other studies with eye-tracking, assessed the effort involved in the comprehension of software models like BPMN [38], ER [30], or TROPOS [39].

Albeit their importance, several PoN studies focused on the evaluation of individual symbols, and on the stakeholders' ability to correctly recognise them. Yet, software engineers use models. A significant difference from previous studies to this paper is that we perform our evaluation at the model level, rather than through isolated symbol recognition tasks.

VIII. Conclusion

We performed a quasi-experiment to compare the *accuracy*, *speed*, and *ease* of the standard *i** concrete syntax and an alternative *i** concrete syntax that resulted from the most successful symbol recognition evaluations for *i** [11]. A total of 57 participants performed understanding and reviewing tasks on *i** SR models. The data collected showed that the alternative concrete syntax had no significant impact in the *accuracy* and *speed* with which participants conducted their tasks. Increased semantic transparency alone did not lead to a better performance with the new *i** concrete syntax. The presence of a language key and the context provided by the model may have mitigated the effect of the increased semantic transparency of the new *i** symbols. For *ease*, we found some indicators of eye-tracking suggesting different visual efforts.

We only addressed one of the nine PoN principles in this study. Further studies should consider the various principles, the interactions among these, as well as their influence on the actual performance of practitioners in understanding and reviewing social goal models. It would be interesting to understand if the new concrete syntax has any drawback (e.g., in model construction) that hinders performance, or why the NASA-TLX questionnaire results do not support the visual effort clear in the heat map, or still, understand the fixation time on relevant/irrelevant AOIs and how they differ between the two groups of participants. Finally, it is necessary to assess how consistently our results occur with other users, models and concrete syntaxes. We plan to replicate the experiment in other contexts, and apply it to bigger and more complex models.

Acknowledgments

We thank NOVA LINCS UID/CEC/04516/2013 and FCT-MCTES SFRH/BD/108492/2015 for financial support.

References

[1] D. M. Fernández, S. Wagner, M. Kalinowski, P. Felderer, M. and· Mafra, Vetrò, T. Conte, M.-T. Christiansson, C. Greer D. Lassenius, T. Männistö, M. Nayabi, M. Oivo, B. Penzenstadler, P. Dietmar, R. Prikladnicki, G. Ruhe, A. Schekelmann, S. Sen, R. Spinola, A. Tuzcu, J. L. de la Vara, and R. Wieringa, "Naming the pain in requirements engineering – contemporary problems, causes, and effects in practice," *Empirical Software Engineering*, pp. 1–36, August 2016.

[2] D. Avison and G. Fitzgerald, *Information systems development: methodologies, techniques and tools*. McGraw Hill, 2003.

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

[4] J. H. Larkin and H. A. Simon, "Why a diagram is (sometimes) worth ten thousand words," *Cognitive science*, vol. 11, no. 1, pp. 65–100, 1987.

[5] D. Moody and J. van Hillegersberg, "Evaluating the visual syntax of uml: An analysis of the cognitive effectiveness of the uml family of diagrams," in *International Conference on Software Language Engineering*. Springer, 2008, pp. 16–34.

[6] A. El Kouhen, A. Gherbi, C. Dumoulin, and F. Khendek, "On the semantic transparency of visual notations: Experiments with uml," in *International SDL Forum*. Springer, 2015, pp. 122–137.

[7] N. Genon, P. Heymans, and D. Amyot, "Analysing the cognitive effectiveness of the bpmn 2.0 visual notation," in *Proceedings of the Third International Conference on Software Language Engineering*, ser. SLE'10. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 377–396.

[8] D. L. Moody, "Why a diagram is only sometimes worth a thousand words: An analysis of the bpmn 2.0 visual notation," Tech. Rep., 2011.

[9] R. Matulevičius and P. Heymans, "Visually effective goal models using kaos," in *International Conference on Conceptual Modeling*. Springer, 2007, pp. 265–275.

[10] D. L. Moody, P. Heymans, and R. Matulevičius, "Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation," *Requirements Engineering*, vol. 15, no. 2, pp. 141–175, 2010.

[11] 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.

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

[13] N. Genon, P. Caire, H. Toussaint, P. Heymans, and D. Moody, "Towards a more semantically transparent i* visual syntax," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2012, pp. 140–146.

[14] V. R. Basili and H. D. Rombach, "The TAME project: Towards improvement-oriented software environments," *IEEE Trans. Software Eng.*, vol. 14, no. 6, pp. 758–773, 1988. [Online]. Available: https://doi.org/10.1109/32.6156

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

[16] 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.

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

[18] D. I. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal, "A survey of controlled experiments in software engineering," *IEEE Transactions on Software Engineering*, vol. 31, no. 9, pp. 733–753, 2005.

[19] D. Falessi, N. Juristo, C. Wohlin, B. Turhan, J. Münch, A. Jedlitschka, and M. Oivo, "Empirical software engineering experts on the use of students and professionals in experiments," *Empirical Software Engineering*, vol. 23, no. 1, pp. 452–489, 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9523-3

[20] M. Höst, B. Regnell, and C. Wohlin, "Using students as subjects-a comparative study of students and professionals in lead-time impact assessment," *Empirical Software Engineering*, vol. 5, no. 3, pp. 201–214, 2000.

[21] T. Gorschek, P. Garre, S. Larsson, and C. Wohlin, "A model for technology transfer in practice," *IEEE Softw.*, vol. 23, no. 6, pp. 88–95, Nov. 2006. [Online]. Available: http://dx.doi.org/10.1109/MS.2006.147

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

[23] 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.

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

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

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

[27] B. Sharif, M. Falcone, and J. I. Maletic, “An eye-tracking study on the role of scan time in finding source code defects,” in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 381–384.

[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] B. Sharif, G. Jetty, J. Aponte, and E. Parra, “An empirical study assessing the effect of seeit 3d on comprehension,” in *Software Visualization (VISSOFT), 2013 First IEEE Working Conference on*. IEEE, 2013, pp. 1–10.

[30] 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.

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

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

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

[34] B. Kitchenham, L. Madeyski, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, “Robust statistical methods for empirical software engineering,” *Empirical Software Engineering*, vol. in press, p. 54, 2016.

[35] D. Van Der Linden and I. Hadar, “A systematic literature review of applications of the physics of notation,” *IEEE Transactions on Software Engineering*, 2018.

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

[37] M. Santos, C. Gralha, M. Goulão, J. Araújo, A. Moreira, and J. Cambeiro, “What is the impact of bad layout in the understandability of social goal models?” in *24th IEEE International Requirements Engineering Conference (RE’16)*. IEEE, 2016.

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

[39] 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.