An Empirical Study on the Efficiency of Graphical vs. Textual Representations in Requirements Comprehension

Zohreh Sharafi[1][2], Alessandro Marchetto[3], Angelo Susi[3], Giuliano Antoniol[2], Yann-Gaël Guéhéneuc[1]
[1] Ptidej Team, DGIGL, Polytechnique Montréal, Canada
[2] Soccer Lab., DGIGL, Polytechnique Montréal, Canada
[3] Fondazione Bruno Kessler - FBK, Trento, Italy
zohreh.sharafi@polymtl.ca, {marchetto, susi}@fbk.eu, antoniol@ieee.org, yann-gael.gueheneuc@polymtl.ca

Abstract—**Graphical representations are used to visualise, specify, and document software artifacts in all stages of software development process. In contrast with text, graphical representations are presented in two-dimensional form, which seems easy to process. However, few empirical studies investigated the efficiency of graphical representations vs. textual ones in modelling and presenting software requirements. Therefore, in this paper, we report the results of an eye-tracking experiment involving 28 participants to study the impact of structured textual vs. graphical representations on subjects' efficiency while performing requirement comprehension tasks. We measure subjects' efficiency in terms of the percentage of correct answers (accuracy) and of the time and effort spend to perform the tasks.**

We observe no statistically-significant difference in term of accuracy. However, our subjects spent more time and effort while working with the graphical representation although this extra time and effort does not affect accuracy. Our findings challenge the general assumption that graphical representations are more efficient than the textual ones at least in the case of developers not familiar with the graphical representation. Indeed, our results emphasise that training can significantly improve the efficiency of our subjects working with graphical representations. Moreover, by comparing the visual paths of our subjects, we observe that the spatial structure of the graphical representation leads our subjects to follow two different strategies (top-down vs. bottom-up) and subsequently this hierarchical structure helps developers to ease the difficulty of model comprehension tasks.

Index Terms—**Graphical representation, Textual representation, Eye-tracking study, Visual path.**

I. Introduction

"A picture is worth a thousand words": graphical information play a vital role in presenting software artifacts along the entire software life cycle from inception and requirement through deployment, maintenance and retirement [1]. Graphical representations (*e.g.*, UML diagrams) are effective tools to (1) promote a quick understanding of data, (2) facilitate data processing and data comparison, and (3) enhance the communication process between end-users and developers.

Moreover, graphical representations do not heavily dependent on natural languages, thus they could mitigate the language barrier problem [2]. Conventional wisdom assumes that graphical representations carry information more effectively to non-technical people than textual one [1]. Based on all these reasons, it can be expected that developers prefer to use graphical representations in contrast with structured textual ones to understand the software under study. Yet, to the best of our knowledge, only a handful of studies investigated the effectiveness of graphical vs. textual representations or the developers' preferences (*i.e.*, textual vs. graphical representations) in program understanding tasks [2], [3], [4].

This paper reports the results of an empirical investigation conducted to quantify the effect of graphical vs. structured textual representations in requirements comprehension tasks. We focus on requirements because they play an important role in any software projects by providing the goals for the development of the system. Hence, understanding requirements is mandatory in any successful software project or task. Given a requirements understanding task and two requirement representations (a graphical vs. a textual), we investigate answer accuracy, the the adopted strategy, the time spent, the effort, and preferences. In the following, for sake of simplicity we will refer to structured textual representation simply as textual documentation. Our study aims at answering the following research questions:

RQ1: Does the type of requirement representations (graphical vs. textual) impact the developers' effort, time, and answer accuracy in requirements comprehension tasks?

RQ2: Does the structure of the representations lead developers to use specific task-solving strategies (top-down vs. bottom-up) during requirements comprehension tasks?

RQ3: Given a graphical and textual representation of a requirements comprehension task, is there any preferred representation by the subjects?

The experiment leverages the availability of a modern eye-tracking system, a tool useful and used by other researchers to study the cognitive process involved in any problem solving activities, including model comprehension. An eye-tracking system provides information that is not available from traditional methods [5], [6], [7], including the exact location and duration of where the subject is looking (eye-gaze data). The information about eye-gaze data helps not only to compute subjects' visual effort while reading requirements but also to display subjects' patterns of eye-movement (visual path). In this

978-1-4673-3092-3/13 © 2013 IEEE

paper, we also investigate visual paths to identify different task-solving strategies used by subjects to perform a requirements comprehension tasks.

The experiment compares a TROPOS [8] graphical representation with a structured textual representation. There exist several textual and graphical representations to describe requirements, including GLR[1] and many others [9]. We chose TROPOS because it is based on the i* modelling framework and proposes goal-oriented modelling techniques. Also, TROPOS includes both a graphical and a textual notations. The 28 subjects (12 female and 16 male subject) answer a set of requirement understanding questions on (1) a textual TROPOS model, (2) a graphical TROPOS model, and (3) a model with both textual and graphical TROPOS representations.

While subjects perform their tasks, the eye-tracking system captures and records eye movements. The answer for the requirement comprehension questions as well as collected eye-movement data are then used to calculate answer accuracy, overall time, and visual effort. Each subject was asked to fill a pre-experiment questionnaire (*e.g.*, attained degree, English proficiency, mother language) and a post-experiment questionnaire in which each subject expresses her or his preference between the representations.

No subject has previous experience with TROPOS models and that TROPOS formalism was introduced in a short hands-on session of about 20 min. Despite this, we do not observe a significance difference between the accuracy obtained when the graphical representation was used vs. the textual representation. However, the time and effort spent on the graphical representations is substantially higher, with medium Cohen-d effect size, than the time spent on textual representation.

Surprisingly, the subjects' preference is largely in favor of the graphical representation, even though it requires a higher effort. However, our subjects spend significantly less time and less effort while working with the third model compared to the two models. This finding points at the value of educating users of graphical representations because the efficiency of our subjects improves after performing the comprehension tasks and learning the TROPOS formalism.

Our results imply that the further a subject's native language is far from English, the more time she spends to perform requirements comprehension tasks. This result is the same for graphical and textual representations, which means that the graphical representation did not reduce the language barrier.

We also investigate the visual paths to analyse different task-solving strategies used by subjects. We use a novel approach called ScanMatch [10] to compare the different visual paths of our subjects while working on textual and graphical representations. We find that subjects use either bottom-up or top-down strategies while using the graphical representation. We believe that it is the horizontal structure of the graphical representation that leads subjects to follow the specific strategy to understand the model and answer the questions. We conjecture that TROPOS structure help subjects and ease the understanding tasks. More empirical investigations are needed to verify if a different layout with possibly non-hierarchical models may impair subjects performance.

The paper is organised as follows: In Section II, we provide the necessary background to this paper. Section III describes the related work. Section IV explains the design of the experiment. Section V presents the analysis of the results following by discussions. Section VI describes the results of the study of subjects' visual paths. Section VII discusses the threats to the validity of our results. Section VIII concludes and sketches future work.

II. Background

In this section, we present the necessary background to this paper: TROPOS and eye-tracking.

A. TROPOS

TROPOS is a goal-based oriented modeling approach to visualise requirement through actor and goal diagrams [8]. TROPOS defines five basic concepts including:

- Actor: it represents a position (role) or an agent that can be a human stakeholder or an artificial agent (software and hardware system).
- Goal: it is an interest of an actor that can be composed of several sub-goals. For each actor, there should be at least one goal.
- Task: it is a particular course of actions that can be executed to satisfy a goal.
- Resource: it is a physical (*e.g.*, printer) or informational (*e.g.*, notes, web-sites) entity that could be used/produced by actors through different tasks to realise goals.
- Social dependency (between two actors): one actor depends on another actor to achieve a goal, execute a task, or deliver a resource.

TROPOS models are visualised through actor and goal diagrams. An actor diagram is a graph whose nodes represent actors while the vertices show the dependencies between pairs of actors. A goal diagram is dedicated to an individual actor and represents the actor's main goals (their decomposition into sub-goals), the tasks and the resources to achieve the goals.

In addition, a TROPOS model can be represented using graphical or textual formats. In graphical format, each concept is shown using a unique element *e.g.*, the actor, goal, task, and resource are represented by a circle, ellipse, hexagon, and rectangle respectively. Moreover, all elements are shown based on three levels of abstraction including the goal level (high level) at the top, the task level in the middle, and the resource level at the bottom that contains goal and task and resource elements respectively. In textual format, each sentences starts with a word mentioning the name of the concepts that is explained (*e.g.*, "Goal: health emergency management."). In this paper, we use the goal diagram in both graphical and textual format.

[1] http://www.cs.toronto.edu/km/GRL/

B. Eye-Tracking

Eye-trackers are designed based on human visual capability to provide additional insight on subject's focus of visual attention to reason about their underlying cognitive processes [11]. Visual attention is the selection of a specific Area of Interest (AOI) visually from the entire visual domain (stimulus) that can trigger the mental processes required for task solving.

An eye-tracking system provides two types of gaze data: eye fixations and saccades. A fixation is the stabilisation of the eye on an object of interest for a period of time, whereas saccades are quick movements of the eyes from one fixation to another. Previous studies [6], [11] show that the comprehension task occurs mainly during eye fixation. Therefore, we use fixation data to compute the amount of visual attention used for measuring the subjects' visual effort. We also consider a visual path as a list of visited AOIs sorted chronologically and use subjects' paths to identify our subjects' task-solving strategies.

In our proposed experiment, we use FaceLAB from *Seeing-Machine*[2], which is a video-based remote eye-tracking system. It has two built-in cameras, one infrared pad, and one computer. FaceLAB tracks subjects' eye-movements by capturing subjects' head using facial features, including nose, eye-brows, and lips. FaceLAB sends eye-movement data to a data visualisation tool, Gaze Tracker from *Eye Response*[3]. Gaze Tracker stores eye-movement data including fixations and saccades associated with each image and can display all fixations on top of the image.

We use two 27" LCD monitors for our experiment: the first one is used by the experimenter to set up and run the experiments while monitoring the quality of the eye-tracking data. We use the second one (screen resolution is 1920×1080) for displaying the models and the questions to the subjects.

III. Related Work

In recent years graphical representations have received increasing attention. However, only a few works addressed the comparison of textual vs. graphical representations.

Ottensooser *et al.* [3] reported significant improvement in understanding of business processes when subjects work with textual representations. They also underlined that subjects must learn how to understand and use the specific graphical notations before starting to use them.

Somervell *et al.* [12] were interested to investigate, when subjects are working on dual-tasks situations by sharing their attentions between different tasks, which combination of graphical and textual representations were more efficient. They considered three different criteria including: facilitation of information monitoring, awareness of information, and introduction of distraction. As a result, they provided a list of guidelines on the use of a combination of textual and graphical representations to improve subjects' efficiency.

Razali *et al.* [13] compared UML-based graphical formal specification vs. a purely textual formal specification in understanding a software specification. They reported that a combination of semi-formal and formal notations improves the subjects' accuracy in comprehension tasks.

Recently, Heijstek [2] *et al.* reported findings pointing to the fact that neither textual nor graphical representations were significantly effective for understanding a software architecture. They also reported that the more experienced subjects, mostly, preferred textual representations.

This paper stems from the belief that there is a need to investigate the difference (if any) between graphical and textual requirement representations. The work presented in this paper is complementary to previous work, because we investigate the impact of textual and graphical representations not only on subjects' effectiveness but also on the strategies that they use to read and understand the requirements representations.

IV. Experimental Design

The *goal* of our study is to investigate the relations between the type of requirement representations (graphical vs. textual) and subjects' visual effort, required time, as well as accuracy in understanding requirements. The *quality focus* is the efficiency of the textual representation compared to that of the graphical representation in requirements comprehension. The *perspective* is that of developers who must understand a software systems to perform development or maintenance tasks. It is also that of researchers who could use our findings to design methods, techniques, and tools to support representations better tailored to comprehension tasks. The *context* of this study consists of three requirements comprehension tasks involving 28 subjects. The experiment is conducted as a within-subject design where the order of treatments depends on the subject's assigned group.

A. Research Hypotheses

To answer the research questions presented in Section I, we propose several null hypotheses. We formulate RQ1 null hypotheses as follows:

- Hα_{11}: There is no significant difference in the average accuracy of the subjects' answers when performing the requirement understanding tasks with graphical and textual representations.
- Hα_{12}: There is no significant difference in the average task time when performing the requirement understanding tasks with graphical and textual representations.
- Hα_{13}: There is no significant difference in the average visual effort when performing the requirement understanding tasks with graphical and textual representations.

Research question RQ2 deals with the subjects' problem solving strategies; we formulate the following null hypothesis:

- Hα_{21}: Despite the structure of the representations, subjects will not use specific task solving strategies while working with the representation to answer the comprehension questions.

Finally, for research question RQ3, we formulate the following null hypothesis:

[2] http://www.seeingmachines.com/

[3] http://www.eyeresponse.com/

TABLE I
DESIGN GROUPS FOR ASSIGNING MODELS TO DIFFERENT SUBJECTS.

1	Session 1: Model A in graphical representation. Session 2: Model B in textual representation. Session 3: Model C in both graphical and textual representation.
2	Session 1: Model A in textual representation. Session 2: Model B in graphical representation. Session 3: Model C in both graphical and textual representation.
3	Session 1: Model B in graphical representation. Session 2: Model A in textual representation. Session 3: Model C in both graphical and textual representation.
4	Session 1: Model B in textual representation. Session 2: Model A in graphical representation. Session 3: Model C in both graphical and textual representation.

- $H\alpha_{31}$: Between graphical and textual representation, there is no preferred representation by subjects.

B. Material

We randomly assign our subjects to the four groups presented in Table I. Each subject works on three different sessions. In each session, each subject works with one treatment: graphical representation, textual representation, and a mixed model, including both graphical and textual representations. Three objects are used: Model A, Model B, and Model C. These objects are extracted from a real industrial project, documenting a hospital information system. These models are presented using the TROPOS graphical and textual representations.

Model A and B are presented in either graphical or textual form for each subject while model C is presented in both graphical and textual form at the same time. Therefore, our subjects can choose between the graphical or textual representations or use both of them while working with model C. Figure 1 and Figure 2 show examples of our graphical and textual representations along with different AOIs. In these figures, only portions of the stimuli are presented.

For Model A and Model B, our graphical representations contain 17 and 18 elements while the number of lines for the textual representations are 19 and 22 for Model A and Model B respectively. The graphical representation of Model C contains 13 elements and its textual representation contains 15 lines of text (font size = 16).

The number of elements and lines is one limitation of the experiment using eye-tracking systems. There is a need to accurately and unambiguously quantify the visual effort by precisely identifying AOIs and thus precisely locating elements of the presented requirements. Yet, the number of lines for the textual representations in our experiment is similar to the source code size of previous eye-tracking studies [14], [15]. Moreover, the number of elements that are presented in our graphical representations are in the range of recommended number of elements for effective program comprehension [16].

In each session, the subjects answer six comprehension questions. These questions were divided into two main categories, bottom-up and top-down as shown in Table II. In top-down questions, we ask our subjects to find the resources or the tasks to fulfill a goal. For these questions, our subjects must first find the top-level goal (high-level of abstraction) and then refine the abstraction to find the required tasks or resources (lower-level of abstraction). In bottom-up questions, we ask our subjects to find the goals that use specific resources. For these questions, our subjects must first find the resources then, going up in the representation, identify the goal.

TABLE II
TWO GROUPS FOR COMPREHENSION QUESTIONS.

Top-down Group	Bottom-up Group
1) What are the resources that used in the realisation of the Goal X? 2) What are the tasks that used in the realisation of the Goal X?	1) What is the usage of resource Y? 2) Resource Y is used in task Z, is it correct or not?

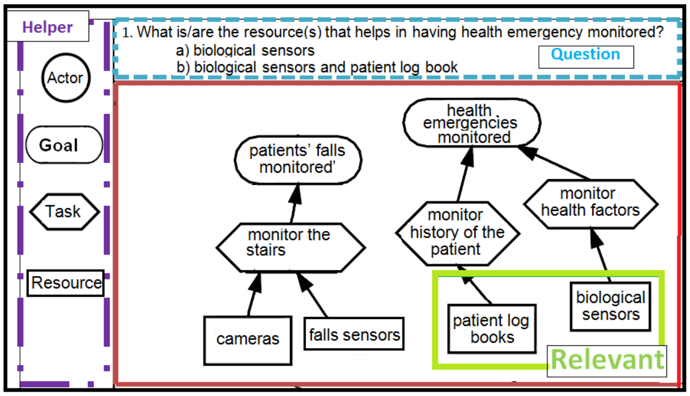

1. What is/are the resource(s) that helps in having health emergency monitored?
a) biological sensors
b) biological sensors and patient log book

Fig. 1. Portion of the graphical stimulus that contains five AOIs: 1. Model area contains 2. Model relevant and 3. Model irrelevant areas; 4. The question area and 5. The help area.

C. Dependent, Independent, and Mitigating Variables

The type of representation (graphical vs. textual) is the independent variable, *i.e.*, treatment, for RQ1 and RQ2. For RQ3, we focus on Model C (the mixed model) presenting both treatments to understand subjects' preferences.

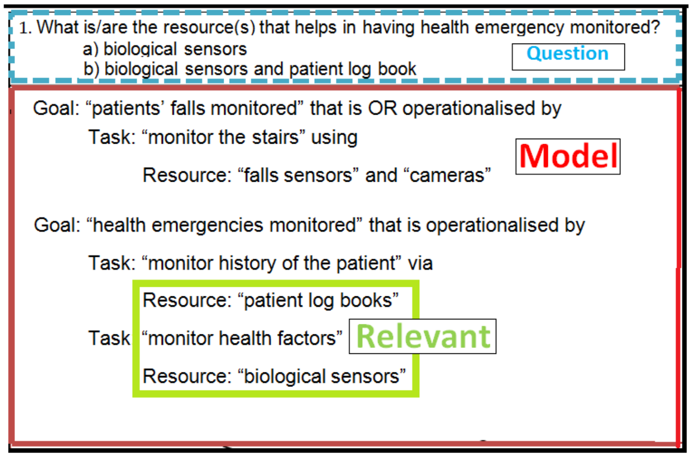
The dependent variables are chosen as follows:

Accuracy: we quantify and measure this variable by the percentage of correct answers given by a subject in the multiple choice questions.

Time: we measure this variable as the amount of time that each subject spends on each model stimulus. We measure this variable using the eye-tracking system.

Effort: we consider effort as the amount of visual attention that subjects spend to answer the question. We assume that less attention and less time means less effort.

In our experiment, we have two treatments/stimuli: the graphical representation and the textual representation. We collect data about fixations on the set of areas of interest (AOI) in each stimulus to compute the subjects' effort. We establish five AOIs for graphical stimulus and four AOIs for textual stimulus as illustrated in Figure 1 and Figure 2, respectively. Some areas cannot be clearly identified in the figures as areas may overlap/intersect, *e.g.*, the model area and the question area. In addition, some areas are parts of other areas, *e.g.*, Model area contains Model relevant and irrelevant areas.



1. What is/are the resource(s) that helps in having health emergency monitored?
a) biological sensors
b) biological sensors and patient log book

Goal: "patients' falls monitored" that is OR operationalised by
Task: "monitor the stairs" using
Resource: "falls sensors" and "cameras"

Goal: "health emergencies monitored" that is operationalised by
Task: "monitor history of the patient" via
Resource: "patient log books"
Task "monitor health factors"
Resource: "biological sensors"

Fig. 2. Portion of the textual stimulus that contains four AOIs: 1. Model area contains 2. Model relevant and 3. Model irrelevant areas; 4. The question area.

Thus, we defined:

- Model area (M): it includes all model elements.
- Model relevant area (Re): it includes all model elements that are related to the correct answer.
- Model irrelevant are (Ir): it includes all model elements that are *not* related to the correct answer.
- Question area (Q): it includes the question and multiple choices that appear at the top of the screen.
- Help area (H): it includes help items for the graphical model displaying TROPOS graphical elements.

We use two metrics for calculating the visual effort: the Average Fixation Duration (AFD) and the surface of the AOI convex hull.

The Average Fixation Duration (AFD): it is correlated with cognitive functions that take place during a problem-solving task [17]. Let $ET(F_i)$ and $ST(F_i)$ represent the end time and start time for fixation F_i and n the total number of fixations in a given AOI. Longer fixations mean that subjects are spending more time analysing and interpreting the model elements to build their internal mental model. Notations that require shorter fixations are thus more efficient than the ones with longer fixations [18]. The metrics AFD is thus:

$$AFD(AOI) = \frac{\sum_{i=1}^{n}(ET(F_i) - ST(F_i))\text{AOI}}{n} \quad (1)$$

The surface of convex hull: a convex hull represents the smallest convex sets of fixations that contains all of a subject's fixations. Goldberg *et al.* [17] proposed and used this measure to evaluate the quality of user interfaces. A smaller value for the (surface of) convex hull indicates that the fixations are close from one another and, thus, that the subject made less effort to find the usable parts of the stimulus in-hand.

Mitigating variables may impact the effect of the independent variables on the dependent variables. In this experiment, we used a questionnaire to collect data about the following mitigating variables: (1) level of study: values for this variable are B.Sc., M.Sc., and Ph.D; (2) level of experience in object-oriented modelling; (3) level of UML knowledge; (4) English language proficiency; and (5) linguistic distance. We provide further explanation about the two last characteristics in the following.

TABLE III
LANGUAGE GROUPING AND DISTANCE. LANGUAGE IS SELF-REPORTED BY OUR SUBJECTS, THE SCORE IS REPORTED IN [19] WHILE THE DISTANCE IS $1/score$ AS REPORTED IN [20]

Language	n	Score	Family	Distance
English	0	-	Indo-European	0.00
French	8	2.5	Indo-European	0.40
Farsi	14	2.00	Indo-European	0.50
Bulgarian	1	2	Indo-European	0.50
Bengali	2	1.75	Indo-European	0.57
Mandarin	2	1.50	Sino-Tibetan	0.67
Arabic	1	1.5	Afroasiatic	0.67

English language proficiency: we ask our subjects to provide a self-assessment (very poor, poor, satisfactory, good, very good) of their English language proficiency. 26% of our subjects evaluate their English language proficiency as satisfactory while 40% and 34% evaluate it as good and very good respectively.

Linguistic distance: Chiswick *et al.* [20] proposed a measure, called Linguistic distance, to find out how difficult for someone who knows language A to learn language B. They assigned each language a score [19] and stated that "if it is more difficult to learn language B1, than it is to learn language B2, it can be said that language B1 is more 'distant' from A than language B2.".

Heijstek *et al.* [2] adopted this measure to investigate whether the difference between their subjects' native language and English can impact their efficiency and accuracy while understanding a model. In this paper, we are also interested to investigate if the distance of our subjects' native language from English can impact our findings. In Table III, we provide a list of the languages that are encountered in the experiment and their associated linguistic distances.

D. Subjects' Demography

The study participants are 28 volunteers, 12 female subjects and 16 male subjects. The subjects are two B.Sc., 11 M.Sc., and 15 Ph.D. students from computer and software engineering and science departments of the Montreal area.

Before the experiment, we inform our participants that the experiment has three sessions and that each session was allotted about 10 minutes and that they are free to leave at any time without incurring any penalty. Collected information is anonymous. We validate the response forms to make sure that participants correctly followed the experiment procedure.

E. Procedure

We use a quiet room to perform the experiment. A 27" LCD screen is used to display the stimuli while the subjects are seated approximately 70 cm away from the screen in a comfortable chair with arms and head rests. Before running the experiment, we give a tutorial to explain TROPOS modelling concepts and its elements during about 20 minutes. Then, we

TABLE IV
MAIN FEATURES OF THE EXPERIMENT.

Collected data	
Number of subjects (#)	28
Total number of questions	504
Number of Text-related questions	168
Number of Graphical-related questions	168
Number of Mixed questions	168
Total time of eye-tracking (hours)	2.85
Total number of fixations (#)	50,652

briefly explain how the eye-tracking system works and what information is gathered by the tool.

For each subject, we first calibrate the eye-tracking system. Then, we start the first session by presenting the first screen, which describes to the subject how to perform the tasks and complete the experiment. When subjects begin a task, we start collecting data. We do not give any time limit to the subjects. We display a representation and a question at the same time; therefore, subjects always have access to the representation to answer the questions. When subjects finish and find an answer, they press the “space" key to go to the next blank screen, and write down their answer to the question, *i.e.*, choose one of the two alternatives. Once a task is finished and the answer given, subjects press the “space" key to go to the next question. When subjects complete the three tasks, we ask them to answer the post-experiment questionnaire.

V. ANALYSIS AND RESULTS

In this section, we report hypotheses testing and discuss the results of our experiment. Table IV summarises the collected data. A replication package is available upon request.

In the analysis of our data, we made no assumption and applied non-parametric, non-paired tests to determine significance differences. We use Taupe [21] to analyse the collected data. Taupe provides the results about fixations and time for each AOI as well as AFDs and convex hull sizes in CSV files that we export to R [22] to perform statistic analyses.

A. Percentages of Correct Answers (Accuracy)

Each subject, when answering a question, chose either a correct or a wrong answer. Table V is the contingency table reporting the number and the percentages of correct and wrong answers for textual, graphical, and mixed representations. We test our hypotheses, $H\alpha_{11}$, to find any potential advantage of graphical vs. textual representations. After applying two-tailed Wilcoxon with (α = 0.05), the p-value reports that there is no significant difference between textual and graphical representations regarding accuracy. We cannot reject the null hypothesis $H\alpha_{11}$. Our results concur with the findings of Heijstek *et al.* [2], who reported that neither graphical nor textual representation had a significant effect on correct answers.

B. Time

We investigate the second hypothesis, $H\alpha_{12}$, which examines the effect of representation type on the time that subjects spend on each stimulus to read a question, analyse the model, and answer the questions.

TABLE V
SUBJECTS' CONTINGENCY TABLE FOR TEXTUAL, GRAPHICAL, AND MIXED REPRESENTATIONS.

Answers					
Textual		Graphical		Mixed	
Correct	Wrong	Correct	Wrong	Correct	Wrong
164	4	165	3	162	6
97%	3%	98%	2%	96%	4%

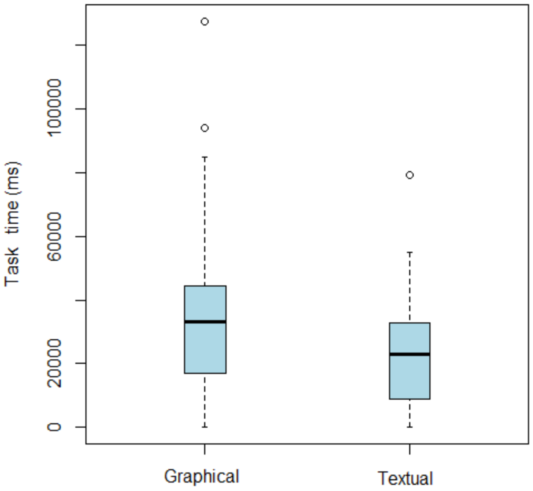


Fig. 3. Descriptive statistics for task time for graphical and textual representations.

Figure 3 presents the distribution of the task-time dependent variable for graphical and textual representations while Table VI shows the average amount of time that our subjects spent on graphical and textual representations for Model A, B, and C, separately. On average, our subjects spent 47% more time (10,498 ms) on graphical representation than textual representation. There is a significant difference between graphical and textual representations (p-value = 1.487e-05, Cohen-d: 0.54 (medium effect)) although this extra time does not affect the accuracy. We can reject the null hypothesis $H\alpha_{12}$.

Moreover, eye-tracking gives us the ability to compute the time that is spent on different parts of a representation separately. Therefore, by considering the set of AOIs that are presented in Section IV-C, we separately compute the percentage of time that our subjects spent on different AOIs of textual and graphical representations. Figure 4 shows the percentages of time that our subjects spent on different parts of representations. When we compare the amount of time on different parts, the results shows that our subjects spend more time on both, relevant (p-value = 0.0016 $<$ 0.05, Cohen-d: 1.0 (large effect)) and irrelevant (p-value = 0.0022 $<$ 0.05, Cohen-d: 1.0 (large effect)) parts of the presentation when working with the graphical presentation compared to the textual model. The time spent on the helper part was negligible.

TABLE VI
AVERAGE TIME AND EFFORT SPENT BY SUBJECTS ON MODELS A, B, AND C WHILE PERFORMING THE REQUIREMENTS UNDERSTANDING TASKS.

		Model A	**Model B**	**Model C**
Average time (ms) (Standard deviation)	**Graphical**	36,925.5 (23,258.8)	32,856.3 (18,471.60)	12,195.47 (1276.39)
	Textual	23,346.5 (14,570.77)	22,340.05 (15,406.64)	4,445.18 (751.80)
Average Effort (Standard deviation)	**Graphical**	72.96 (29.05)	65.89 (27.56)	36.30 (26.29)
	Textual	33.9 (37.88)	35.7 (48.26)	16.61 (15.39)

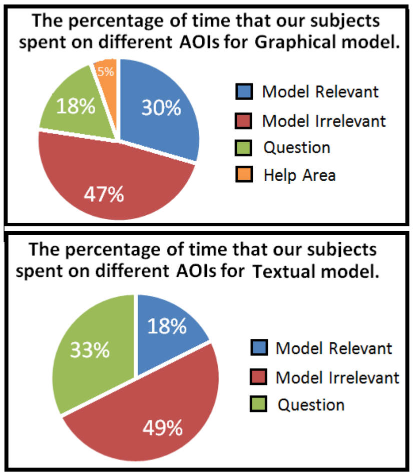


Fig. 4. The percentage of time that our subjects spend on different part of graphical and textual representations.

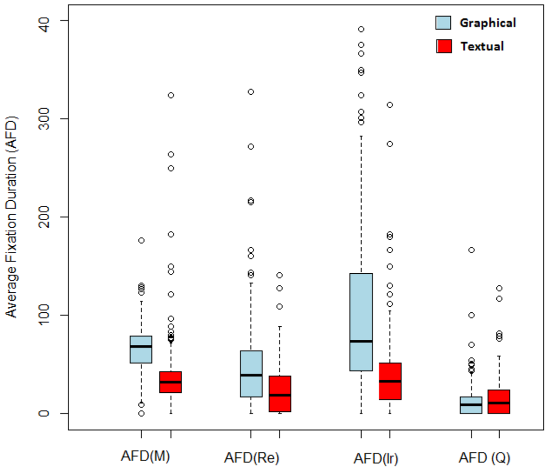


Fig. 5. Descriptive statistics for AFD for graphical and textual models.

TABLE VII
TWO-TAILED WILCOXON P-VALUE (α = 0.05) AND COHEN-D FOR THE AVERAGE FIXATION DURATION (AFD) METRICS OF DIFFERENT AOIS.

Variables	**p-value**	**Cohen-d**
AFD(M)	$<$ 2.2e-16	0.64
AFD(Re)	$<$ 9.328e-07	0.50
AFD(Ir)	$<$ 6.787e-11	0.83
AFD(Q)	0.07163	–

C. Visual Effort

We analyse the third hypothesis, $H\alpha_{13}$, which examines the effect of a representation type on the effort that subjects spend to perform model comprehension tasks. We compare the value of our subjects' AFDs for graphical and textual representations while considering different AOIs including Model area (AFD(M)), Question area (AFD(Q)), Model relevant (AFD(Re)) and Model Irrelevant (AFD(Ir)) areas. Figure 5 presents the distribution of the AFD dependent variable for graphical and textual representations. The results of applying Wilcoxon test as presented in Table VII show that there is a significant difference between graphical and textual representation while comparing AFD for Model area (AFD(M)) and Model relevant (AFD(Re)) and Model Irrelevant (AFD(Ir)) areas. These values show that our subjects spent more visual effort while looking and analysing the irrelevant parts of the model to find the relevant elements and relevant parts to find the correct answer with the graphical representations.

Our results show that our subjects have longer fixations while working with the graphical representations, which means they are spending more effort analysing and interpreting the elements of graphical representations. As expected, there is no significant differences between the amount of visual effort for Question part (AFD(Q)) because the question part is the same for both representations.

Moreover, we use the non-parametric Wilcoxon test to compare the surface of the convex hull of our subjects' fixations on Model AOI for textual and graphical representation. We use Taupe [21] to compute the surface of convex hulls. As expected, using the textual representation significantly decreases the value of convex hull (α = 0.05, p-value = 0.01). This result shows that the fixations are close from one another when our subjects work with textual representation, thus, our subjects put less effort to explore the whole model to find the relevant parts of the stimulus to answer the question.

D. Impact of the Mitigating Variables

English language proficiency: there is no interaction between subjects' English proficiency self-assessment and accuracy, time, and effort.

Linguistic distance: we find that the language distance is significantly correlated with time (p-value $<$ 0.05). This result shows that the further a subject's native language is from English, the more time she spends to find the answer. This result is independent from the type of representation, which implies that the graphical representation could not help our non-native subjects to overcome the language barrier. This result is in agreement with the work of Heijstek *et al.* [2].

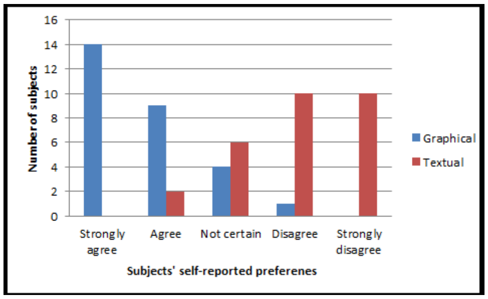


Fig. 6. Subjects' self reported preferences about working with graphical vs. textual representations.

Experience: to evaluate the impact of experience, we consider study level, level of UML and of object-oriented modelling, and the number of years of modelling experience. None of these values significantly interact with representation types to have an effect on accuracy, time, and effort. Our subjects are students so their prior modelling knowledge is rather homogeneous. We also design our tasks such that solving them does not require industrial experience. Our results show that the impact of representation types and the task cognitive complexity is not hidden by the subjects' experience.

E. Representation Preference

In our post-questionnaire, we ask our subjects their preferences for answering comprehension question using graphical vs. textual representations. 82% of our subjects prefer to work with a graphical representation as shown in Figure 6. Moreover, in the third session, when we provide subjects with a mixed model (Model C), consisting of both graphical and textual representations, for 96% of the questions, our subjects started with the graphical representation to find the answer.

Yet, our subjects spend significantly more time (p-value < 0.05, Cohen-d: 1.0) and effort (p-value < 0.05, Cohen-d: 1.02) on graphical representations (see Table VI). Therefore, we reject $H\alpha_{31}$ and answer RQ3 as follows: our subjects' preferences is largely in favor of graphical representation. We conclude that subjects' preference is not related to the real effort and time spent.

We apply Kruskal-Wallis rank sum test on three sets of time and visual effort to see if our subjects spend equal amount of time and effort on the three models. The result of the test (p-value < 0.05) for both time and effort confirm that our subjects spend more time and effort on Model A and B compared to Model C. This result confirms that, although none of our subjects is familiar or use TROPOS before, they learn TROPOS through performing the experiment and this can significantly improve their efficiency. This finding emphasise that (in agreement with previous works [23]) developers must learn how to use a graphical representation before using it.

VI. Visual Path Analysis

We perform the following steps to answer RQ2: first, we consider a new set of AOIs including goal, task, and resource areas representing different levels of abstraction of TROPOS representations and also the Question and Help areas, as shown in Figure 7.

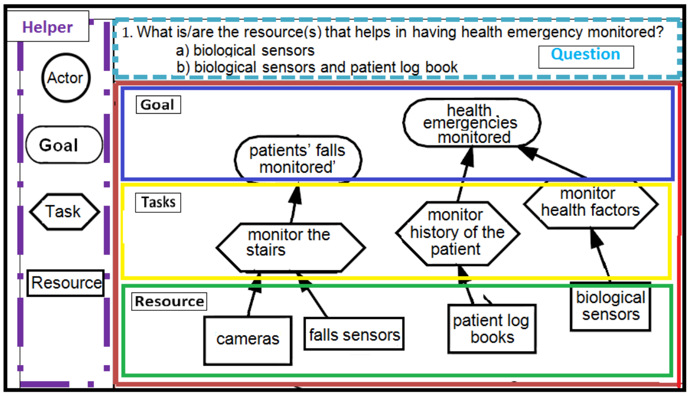

Helper
Actor
Goal
Task
Resource
1. What is/are the resource(s) that helps in having health emergency monitored?
a) biological sensors
b) biological sensors and patient log book
Question
Goal
health emergencies monitored
patients' falls monitored'
Tasks
monitor health factors
monitor history of the patient
monitor the stairs
Resource
biological sensors
patient log books
cameras
falls sensors

Fig. 7. A set of five AOIs including goal level, task level, resource level, question and helper areas for TROPOS graphical representation.

Second, we use ScanMatch [10] to compute and show different visual paths for each subject working on all questions. ScanMatch assigns a character to represent each AOI. The question AOI, goal AOI, task AOI, resource AOI and helper AOI are represented by B, C, D, E, and F respectively. To display the visual path, a small letter is attached to the capital letter to make it easy to read the sequence. Therefore, if a subject goes from goal AOI to task AOI and then to resource AOI, the sequence for her visual path is cCdDeE.

Third, using ScanMatch, we compare the subjects' visual paths when working with six different questions of the Model A and Model B, presented in graphical representations. ScanMatch calculates the similarity values to show the similarity of two visual paths temporally and spatially. If two visual paths are identical, the score is 1. If they do not have any relationship, the value will be 0. We calculate the similarities for all subjects pair-wise. For example, if subjects 1, 2, and 3 are working on Q1 of Model A, we compare and compute the visual paths of subjects 1, 2, and 3 then, we calculate the similarity value of subject 1 vs. subject 2, subject 1 vs. subject 3, and subject 2 vs. subject 3. We perform this procedure for all six questions for both models A and B and obtain, for each question, a list of similarity values for each pair of subjects.

Fourth, based on our top-down and bottom-up questions presented in Table II, we expect to detect two different strategies. Our subjects answer six different questions for each model. For each question, we have a list of similarity values for the visual paths of all pairs of subjects. We perform the two-tailed version of the unpaired Wilcoxon test, using the Bonferroni correction, on these lists. Then, we compute Cliff's delta to indicate the right categorisation with the highest effect size. Based on this result, our questions are divided into two groups: first group consists of Q1, Q2, Q4, and Q5 while the second group consists of Q3 and Q6. The result of the statistical test shows that our subjects use different strategies to answer questions presented in the first and second group with the first group containing top-down questions and the second group consisting of bottom-up questions.

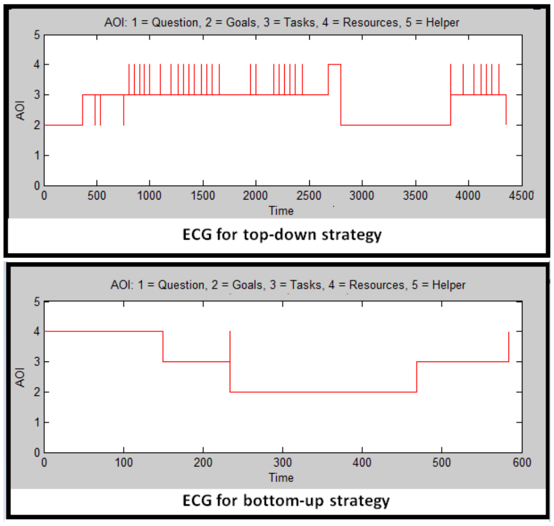


Fig. 8. The STG graph which shows the navigation sequence of AOIs for a subject using top-down and bottom-up strategies.

Finally, to visually show the different task-solving strategies used by our subjects, we draw the structure transition graphs (STG) proposed by Sim *et al.* [24] for different subjects working on two different groups of questions: top-down and bottom-up. Sim *et al.* [24] use the STG to visualise the progress and the rhythm of program comprehension while asking subjects to modify a system. They depict STG to show that developers attempts to go back and forth within files that are in neighboring layers. They conclude that the STGs show patterns in developers' behaviour. We adopte this method to show how our subjects navigate through different AOIs (different levels of abstractions) to answer the questions for graphical representations.

As shown in Figure 8, the x-axis in the STG is time while the y-axis shows the different AOIs. A STG shows the navigation sequence between AOIs for subjects while looking and analysing the graphical model to answer two different types of questions: bottom-up and top-down. As shown in the image at the top, the subjects started from goal part (level 2) going through task part (level 3) to reach the resource part (level 4) to answer a question of top-down group. The second image at the bottom shows the bottom-up navigation that started from resource part (level 4), going through task part (level 3), and finally reaches goal part (level 2) to answer a question of group B.

Sometimes our subjects read the question from the paper in hand. Therefore, they do not visit the question part (level 1). We believe that fast traversing between two different AOIs, which appears as a straight vertical line in the STG, can be considered as a saccade and can be removed. The vertical line between AOI 1 and AOI 2 means that the subject was looking at AOI 1 then suddenly looked at AOI 2 for less that 0.01 ms and looked back at AOI 1. Because the subjects do not spent enough time on AOI 2, we can not consider the vertical lines as a complete transition between two AOIs.

Our findings reject $H\alpha_{21}$ and confirm that the hierarchical structure of the TROPOS graphical representation leads our subjects to follow a specific strategy, either top-down or bottom-up, to answer the question.

VII. Threats to Validity

In the following, we discuss a number of factors that may have influenced our results.

1) Internal validity: we randomly assign our subjects to one of the four sessions of experiment and also change the order of representations for each subject to mitigate the impact of maturation. We also prevent fatigue effect concerning the models given at the end using random ordering. To mitigate the possible diffusion of the treatments, we ask our subjects not to talk about the experiment with the other subjects.

Regarding the instrumentation threat that is related to the equipment used in our study, we use a video-based eye-tracking system that does not have any heavy goggle. Our subjects could move their heads easily without changing the calibration of the camera. Another potential threat is that subjects might behave differently and being under stress because we record them using the eye-tracking camera. However, we explain to subjects that eye-tracker does not provide any video.

2) Construct validity: We do not inform the subjects about the precise goal of the experiment to avoid hypothesis guessing. We explain them the process of performing the experiment, the number of session, and questions that they must answer. We do not set a time limit, we ask subjects to answer the questions as soon as they can.

3) External validity: this threat is related to the generalization of our results. We use students as subjects, our subjects are graduate, Masters, and Ph.D students (except for two bachelor student who are currently enrolled in their 4^{th} year) with good knowledge of object-oriented modeling. We do not distinguish novices and experts. Kitchenham *et al.* [25] mentioned that "using students as subjects is not a major issue as long as you are interested in evaluating the use of a technique by novice or non-expert software engineers. Students are the next generation of software professionals so, are relatively close to the population of interest." We have 28 subjects, which is much more than some previous eye-tracking studies. Sharif *et al.* [26] had 15 subjects and they mentioned that eye-tracking studies usually have about the same number of subjects.

4) Conclusion validity: to address conclusion validity, we do not consider any assumption regarding the distribution and normally of our data, therefore, we use non-parametric, non-paired statistical tests to determine significant differences. Moreover, we choose well-documented measures from the previous works [18], [26] to ensure the reliability of our measures. Finally, we make sure that the eye-tracker is well calibrated for every subject before collecting data.

VIII. Conclusion and Future Work

We designed and performed an eye-tracking experiment to investigate the impact of textual vs. graphical representations on subjects' efficiency while performing requirements comprehension tasks. We also examined the effect of graphical representation structure on subjects' strategies.

We found no statistically-significant differences between representation types when considering accuracy. However, our subjects spent more time and effort while working with the TROPOS graphical representations. Hence, although our subjects mostly preferred to use the graphical representation, they performed the requirements comprehension tasks more efficiently while working with the TROPOS structured textual representation. In addition, our subjects performed significantly better regarding time and effort while working with the mixed model after they worked with the two first models and learnt TROPOS. This result implies that the formalism of a graphical representation must be learnt by users and that training is required before the benefits of a graphical representation can materialise. In addition, the subjects who performed more efficiently with both graphical and textual representations had a native language close to English, which implies that the graphical representation could not help non-native English speakers to improve their efficiency.

When we compared the subjects' visual paths, we observed that they followed two different strategies: top-down and bottom-up. Our findings suggest further studies concerning the impact of representation structure (layout) on developers' strategies and performance. In future work, we will also replicate our study with more subjects and different representations. Moreover, we plan to perform an additional experiment without eye-tracking with more realistic tasks and more complex models.

Acknowledgment

The authors would like to thank the participants of the study as this work would not be possible without their collaboration. This work has been partially supported by the NSERC Research Chairs on Software Cost-effective Change, Evolution and on Software Patterns and Patterns of Software, and by Fonds de recherche du Québec (FQRNT).

References

[1] 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.

[2] W. Heijstek, T. Kuhne, and M. R. V. Chaudron, "Experimental analysis of textual and graphical representations for software architecture design," in *Proceedings of the 2011 International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 167–176. [Online]. Available: http://dx.doi.org/10.1109/ESEM.2011.25

[3] A. Ottensooser, A. Fekete, H. A. Reijers, J. Mendling, and C. Menictas, "Making sense of business process descriptions: An experimental comparison of graphical and textual notations," *Journal of Systems and Software*, vol. 85, no. 3, pp. 596–606, Mar. 2012.

[4] S. Yusuf, H. Kagdi, and J. I. Maletic, "Assessing the comprehension of uml class diagrams via eye tracking," in *Proceedings of the 15th IEEE International Conference on Program Comprehension*, ser. ICPC '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 113–122. [Online]. Available: http://dx.doi.org/10.1109/ICPC.2007.10

[5] A. Calitz, M. Pretorius, and D. Greunen, "The evaluation of information visualisation techniques using eye tracking," 2009.

[6] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer-Verlag New York Inc, 2007.

[7] J. N. Kara Pernice, "Eyetracking methodology: How to conduct and evaluate usability studies using eyetracking," `http://www.useit.com/eyetracking/methodology`, August 2009.

[8] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos, "Tropos: An agent-oriented software development methodology," *Autonomous Agents and Multi-Agent Systems*, vol. 8, no. 3, pp. 203–236, 2004.

[9] P. Laurent, P. Mader, J. Cleland-Huang, and A. Steele, "A taxonomy and visual notation for modeling globally distributed requirements engineering projects," in *Global Software Engineering (ICGSE), 2010 5th IEEE International Conference on*. IEEE, 2010, pp. 35–44.

[10] F. Cristino, S. Mathot, J. Theeuwes, and I. D. Gilchrist, "Scanmatch: A novel method for comparing fixation sequences." *Behaviour Research Method*, vol. 42, pp. 692–700, 2010.

[11] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[12] J. Somervell, C. M. Chewar, and D. S. Mccrickard, "Evaluating graphical vs. textual secondary displays for information notification," in *Proceedings of the ACM Southeast Conference, Raleigh NC*, 2002, pp. 153–160.

[13] C. F. Snook and R. Harrison, "Experimental comparison of the comprehensibility of a uml-based formal specification versus a textual one," in *Proceedings of 11 th International Conference on Evaluation and Assessment in Software Engineering (EASE*, 2007, pp. 955–971.

[14] 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*, ser. ETRA '12. New York, NY, USA: ACM, 2012, pp. 381–384.

[15] H. Uwano, M. Nakamura, A. Monden, and K.-i. Matsumoto, "Analyzing individual performance of source code review using reviewers' eye movement," in *Proceedings of the 2006 symposium on Eye tracking research & applications*, ser. ETRA '06. New York, NY, USA: ACM, 2006, pp. 133–140.

[16] S. Ambler, *The Elements of UML (TM) 2.0 Style*. Cambridge University Press, 2005.

[17] 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.

[18] G. Cepeda Porras and Y. 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.

[19] L. Hart-Gonzalez and S. Lindemann, "Expected achievement in speaking proficiency," School of Language Studies, Foreign Services Institute, Department of State, Tech. Rep., 1993.

[20] B. Chiswick and P. Miller, "Linguistic distance: A quantitative measure of the distance between english and other languages," *Journal of Multilingual and Multicultural Development*, vol. 26, no. 1, pp. 1–11, 2005.

[21] 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*, 2012.

[22] R. Team *et al.*, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, no. 01/19, 2010.

[23] K. Siau, "Informational and computational equivalence in comparing information modeling methods," *Journal of Database Management (JDM)*, vol. 15, no. 1, pp. 73–86, 2004.

[24] S. E. Sim, S. Ratanotayanon, and L. Cotran, "Structure transition graphs: An ecg for program comprehension?" in *ICPC*. IEEE Computer Society, 2009, pp. 303–304.

[25] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Trans. Softw. Eng.*, vol. 28, no. 8, pp. 721–734, Aug. 2002.

[26] B. Sharif and J. Maletic, "An eye tracking study on camelcase and under_score identifier styles," in *IEEE 18th International Conference on Program Comprehension (ICPC)*. IEEE, 2010, pp. 196–205.