Women and Men – Different but Equal: On the Impact of Identifier Style on Source Code Reading

Zohreh Sharafi[1], Zéphyrin Soh[1,2], Yann-Gaël Guéhéneuc[1], and Giuliano Antoniol[1]
[1]Ptidej Team and Soccer Lab, Département de Génie Informatique et Génie Logiciel
École Polytechnique de Montréal, Canada
[2] Department of Computer Engineering, UIT, University of Ngaoundéré, Cameroon
Email: {zohreh.sharafi,zephyrin.soh,yann-gael.gueheneuc,giuliano.antoniol}@polymtl.ca

***Abstract*—Program comprehension is preliminary to any program evolution task. Researchers agree that identifiers play an important role in code reading and program understanding activities. Yet, to the best of our knowledge, only one work investigated the impact of gender on the memorability of identifiers and thus, ultimately, on program comprehension. This paper reports the results of an experiment involving 15 male subjects and nine female subjects to study the impact of gender on the subjects' visual effort, required time, as well as accuracy to recall Camel Case versus Underscore identifiers in source code reading.**

We observe no statistically-significant difference in term of accuracy, required time, and effort. However, our data supports the conjecture that male and female subjects follow different comprehension strategies: female subjects seem to carefully weight all options and spend more time to rule out wrong answers while male subjects seem to quickly set their minds on some answers, possibly the wrong ones. Indeed, we found that the effort spent on wrong answers is significantly higher for female subjects and that there is an interaction between the effort that female subjects invested on wrong answers and their higher percentages of correct answers when compared to male subjects.

***Index Terms*—Gender issues, Identifier style, Eye-tracking study, Code reading.**

I. Introduction

Program evolution is often solely based on source code because program documentation is often not available or outdated. Furthermore, even source code comments are also not necessarily updated or rich enough to help developers in understanding a source code. In such a situation, the only reliable source of information is source code itself and program comprehension is completely dependent on identifiers. Consequently, several previous works investigated the role of identifiers and, possibly comments, in program comprehension [3], [9], [25]. These works agreed on the important role of identifiers in program comprehension, in code readability, and, more generally, in software evolution [3], [6], [9].

Some works also investigated the impact of the quality and style of identifiers on their recall by subjects and code readability [3], [25] . Developers often compose identifiers by concatenating terms, abbreviations, acronyms, and complete words. For example, Java developers often use English and the Camel Case (CC) convention to create identifiers. The CC convention is the practice of creating identifiers by concatenating terms with their first letter capitalized, giving the identifiers a camel-like look with flats and humps, *e.g.*, *absolutePath*. The CC convention is the de facto standard for Java. However, it is not common in other programming languages, such as in C, where developers use different conventions, in particular the use of underscore, *e.g.*, *absolute_path*.

The underlying conjecture in these works is that identifier style, *i.e.*, CC versus Underscore (US), affects memorability. Memorability [3], [25] is an important factors that impact the comprehension of identifiers. It includes being able to recall an identifier by providing some semantic information associated with it [18]. Lacking memorability may impair identifiers recall, code readability, and, ultimately, program comprehension. These previous works reported some conflicting results. In terms of accuracy, measured as the percentage of correctly recalled identifiers and thus memorability, Binkley *et al.* [3] reported that CC is more accurate than US while Sharif *et al.* [25] concluded that there is no difference.

Moreover, Lawrie *et al.* [21] investigated the impact of identifier makeup (length and structure) on memorability. They reported that informative identifiers are more important for female subjects than for male subjects. However, female subjects understand more from the abbreviations than male subjects. In software engineering, other works studied the impact of gender in problem solving and information processing activities [4], [12], [14], [22], [23]. As suggested by Burnett *et al.* [4], understanding differences between male and female subjects can reveal gender biases while benefiting both male subjects and female subjects, for example through the design of tools better adapted to each gender.

Although CC and US conventions are likely to be equivalent for literate developers, no previous work investigated the impact of gender on identifier style and identifier comprehension strategies. Consequently, in this paper, we report data from an experiment designed and performed to investigate identifier understanding and gender differences when using identifiers with the CC and US styles. We answer the following research question (RQ): **Does the developers' gender impact their effort, their required time, and as well as their ability to recall identifiers in source code reading?**

As underlined above, previous work, such as [3] and [25], reported contradictory findings on identifier styles. Thus, as a preliminary step, we seek to verify if, given our subject

population, there was a difference in identifier style preference between male subjects and female subjects.

In designing the experiment, we were inspired by the literature on the usefulness of eye-tracking systems to study the cognitive behavior and effort involved in solving some problem. An eye-tracking system provides information that are not available from traditional methods, including the subjects' patterns of visual attention during a task [5], [10], [19]. Indeed, an eye-tracking system is an unique tool to gather fine grain information on a subject's visual patterns. We use such a system to compute the subject visual effort spent on correct answers as well as on distractors. Furthermore, visual patterns can be represented as heatmaps to straightforwardly show the differences among subjects.

Our experiment consists of 24 subjects, nine female subjects and 15 male subjects. We administer to each subject three recall tasks, involving recalling the name of identifiers written in the CC or US style. We randomly assign the treatment (CC or US) to each subject. While subjects perform their tasks, we measure their effort as the amount of visual attention calculated from eye-tracker's data. We also measure the percentage of correctly recalled identifiers, and required time.

There are two main types of eye gaze data: eye fixation and saccade. A fixation is the stabilization of the eye on an object of interest for a period of time, whereas saccade are quick movements of eye from one fixation to another. It has been reported [10], [24] that the comprehension task occurs during fixation. Therefore, we use fixation's information for calculating the amount of visual attention used for measuring the visual effort of the subjects. Clearly, as the eye-tracking system allows us to identify the different regions and the time spent in different parts this also helps to highlight different task solving strategies.

The collected data supports the evidence that no difference exists between CC and US identifiers, even when considering gender. However, male and female subjects seem to follow different comprehension strategies where female subjects spend more visual effort on the incorrect choices (distractors) while answering the multiple choice questions to recall the identifiers. Female subjects put more visual effort (attention) than male subjects and analyze all distractors though the average time and accuracy of female subjects are not statistically different than that of the male subjects.

Moreover, to our surprise, we observe a statistically strong interaction between accuracy, effort spent on distractors, and correct answers when modelling the female subjects' accuracy. Also, we report that no correlation exists between visual effort on correct answers or distractors and the male subjects' accuracy. Data analysis via linear models supports the conjecture that it is not the time spent on distractor or correct answers that matters for female subjects but rather the complex and yet-unknown interaction between the two. Therefore, it is not how much visual effort is devoted to distractors or correct answers but rather some complex pondering of correct and wrong answers that matters to female subjects. We explain our findings using previous studies that reported that females are usually less confident than males while working with different programming environment [2], [4], [16].

Our findings support the evidence that female subjects analyse all the possible choices more precisely before choosing the correct answer. This finding is also supported by a preliminary analysis of subjects' heatmaps. An in-depth modelization and comparison of the heatmaps will be the subject to our future work. It is important to underline two points. First, the extra time spent by female subjects to analyze wrong alternatives do not statistically impact the overall time that they spend answering when compared to male subjects. It appears that male and female subjects follow two different strategies: female subjects carefully weight all options and spend more time to rule out wrong answers than male subjects. Second, though nine female subjects participate to the study, the population is not large enough and we cannot generalize our findings. Our findings at the moment should be considered more as a call for further studies concerning the role of gender in software engineering.

The paper is organized as follows: Section II describes the related works. Section III explains the design of the experiment. Section IV presents the analysis of the results and the discussion. Section V discusses the threats to the validity of our results. Section VI concludes and sketches future work.

II. Related Work

The approaches that are relevant to our research focus on identifier names, source code understanding and gender difference in problem solving activities.

Lawrie *et al.* [21] performed an experiment to investigate the impact of identifier length on source code understanding. They reported that full-word identifiers lead to better understanding in comparison with short identifiers.

Binkley *et al.* [3] analyzed the impact of identifier style on the speed and accuracy of source code modification tasks. Results implied that CC identifiers leads to higher accuracy but it takes more time in comparison with US.

Sharif *et al.* [25] carried on an eye-tracking experiment to analyze the effect of identifier style. Their results showed that there is no significant difference between CC and US styles with respect to the subjects' accuracy. Moreover, using US style leads to have higher efficiency and lower visual effort in their subjects in comparison with CC. They used eye-tracking data to calculate the visual effort.

Gender differences in problem solving activities has been investigated in different research in different domain. The research in Selectivity Hypothesis [22], [23] proposed that females pay more attention to details and use disparate, multiple cues for processing information in both simple and complex tasks. However, males tend to find the first cue and follow it to solve a problem. Males also use complex, comprehensive strategies only for complex problems.

Different studies investigate the impact of self-efficacy on strategies deployed by males and females to solve a problem. These studies explained that females has less self confidence towards their abilities to success than males [2], [28].

Hertzel *et al.* [16] reported that females do not feel confident if they do not have any task-specific experience but males do not hesitate to apply general knowledge for specific tasks.

Fisher *et al.* [12] conducted a study to compare male and female subjects' performance on program comprehension tasks. They proposed a model to link spatial cognition and program comprehension. Their result indicated that there is no differences between gender with respect to mental rotation, object and location memory. They explained that male and female subjects improve their skills and adapt them to be professional developers. Therefore, they are equivalent in these skills while working as a software developer. However, they concluded that females mostly used a bottom-up approach while males preferred to adapt a top-down approach.

Grigoreanu *et al.* [14] performed an experiment to investigate scripting strategies and explained how male and female subjects differently used these strategies. Male subjects used testing for all stages of debugging, including finding and fixing the bug and evaluating their fix. However, female subjects used testing only for finding bugs.

III. Experimental Design

The *goal* of our study is to investigate the relations between gender and subjects' visual effort, required time, as well as ability to recall identifiers in source code reading. The *quality focus* is identifiers memorability and thus program comprehension effort, which may depend on gender. The *perspective* is that of developers, who perform development or maintenance activities and need to understand a code fragment. It is also that of researchers to possibly find systematic bias that can be considered in the future empirical studies involving male and female subjects. The researchers could also use our findings to design methods, techniques, and tools better adapted to different developers or support different code reading and program understanding strategies. The *context* of this study consists of three program comprehension tasks involving 25 subjects (nine females) and two variants of three Java code fragments where identifiers have been coded following the CC style (first treatment) and the US style (second treatment). The experiment is conducted as not within-subjects design. An overview of our experiments is outlined in Table I.

A. Research Hypotheses

This study aims at answering our research question presented in Section I. This **RQ** leads us to formulate the following null hypotheses:

- Hα_{01}: there is no significant difference in the average accuracy between male and female subjects while reading the source code to recall identifiers.
- Hα_{02}: there is no significant difference in the amount of required time (speed) between male and female subjects while reading the source code to recall identifiers.
- Hα_{03}: there is no significant difference in the average visual effort between male and female subjects while reading the source code to recall identifiers.

TABLE II
List of identifiers used in the three Java programs.

	Code 1	**Code 2**	**Code 3**
Class name	Java2DFrame	DBTest	PrimeNumCalc
Method names	paint2DObjects drawString drawLine initComponents	executeQuery closeConnection	calcPrimeNums
Variable names	graphics2D roundRectangle sampleLine	dBConnection dBStatement dBResultSet dBDriver dBQuery dBURL	upperLimit nCounter innerLoop isPrimeNum

The above null hypotheses can be also detailed to account for both treatments: CC and US styles. Thus, we divided our analysis in two part. First, we verify if given our subjects any difference exists between the two styles (*i.e.*, CC versus US style). We anticipate no difference likely exists. Then, we study the impact of gender. In the following, for the sake of simplicity, we will refer to visual effort as effort.

B. Material

We use three small Java programs: a `2D graphical frame`, a `Database tester`, and a `prime number calculator`. We find these small programs in the Java Source Code Example Web page[1]. We adapt to the two identifier styles: CC and US, the two treatments. The lengths of the programs (in lines of code) are 30, 36, and 44, respectively.

We must use small Java programs to accurately and unambiguously quantify the visual effort. However, the source code size is similar to previous eye-tracking studies [26], [29]. Indeed, we can display a small Java program on a single screen and thus, for reading source codes, the subjects do not need to scroll down or traverse different pages. If the subjects had to scroll or traverse pages, it would have been difficult and error-prone to analyze the eye-trackers' data, especially if the subjects go back and forth between different pages.

Our experiment focuses on identifiers. We removed any comments from the source code. Thus, all the information about the source code is captured by its identifiers. We use two variants of the three programs: one in which all identifiers are written in CC style and another in which all identifiers follow the US style. We chose each identifier to contain only two or three terms. Table II shows the list of all identifiers used in our experiment for the CC program variants.

Figure 1 shows an example of source code and question.

C. Dependent, Independent, and Mitigating Variables

The subjects' gender (male or female) is the main independent variable along with the style of the identifiers (CC or US) while the dependent variables, summarized by Table I, are:

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

[1] http://www.javadb.com/

TABLE I
AN OVERVIEW OF THE EYE-TRACKING EXPERIMENT.

	Experiment
Goal	Study the impact of the gender, CC and US style on source code reading.
Independent variables	Gender: male (M) or female (F); identifier style: Camel Case (CC) or Underscore (US).
Dependent variables	Accuracy, Required time (Speed), and Visual Effort.
Mitigating variables	Study level and Style preference of subjects.

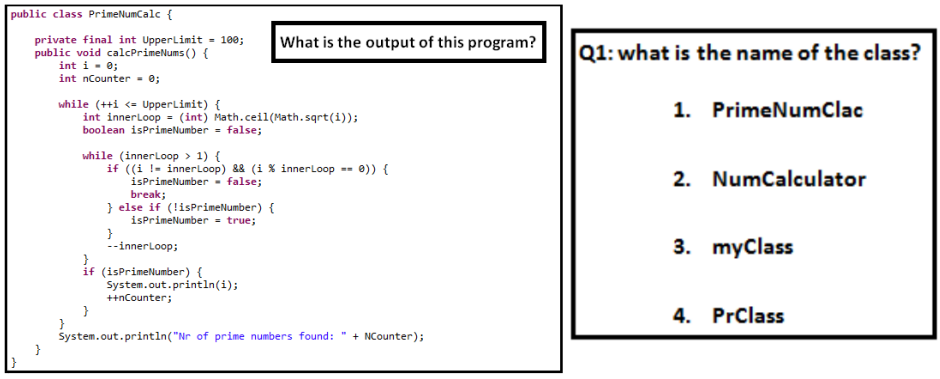

Fig. 1. (Left) source code stimulus; (right) question stimulus.

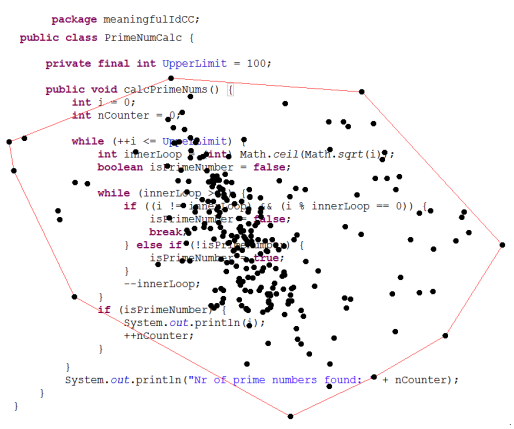

Fig. 2. A source code stimulus that contains a convex hull. The convex hull is shown by red lines with the fixations represented by black dots.

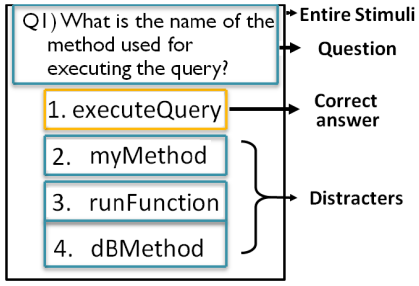

Fig. 3. A question stimulus that contains four areas of interest: entire stimulus, question, correct answer, distractors.

- **Required Time (Speed):** we measure this variable as the amount of time that each subject spends on the source code and question stimuli. We measure this variable using the eye-tracking system with each subject.
- **Effort**: we measure effort using the eye-tracker data. We consider effort as the amount of visual attention that subjects must spend to answer the question: less attention and less time means less effort.

In our experiment, we have two stimuli: the source code stimulus and the question stimulus that we differentiate. Thus, we use two different sets of metrics for effort calculation.

1) Source code stimulus: we calculate the convex hull of the fixations to compute the visual effort. A convex hull represents the smallest convex sets of fixations that contains all of a subject's fixations. Goldberg and Kotval [13] used this measure to evaluate the quality of the user interfaces.A smaller value for the 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 program. In Figure 2, we show the source code stimulus with the convex hull shown by red lines and the fixations shown by black dots.
2) Question stimulus: each question stimulus contains a multiple choice question. We collect data about fixations on the set of areas of interest (AOI) in the screen to compute the subjects' effort. An area of interest is a relevant element of the stimulus. We establish four AOI for the question stimulus as illustrated in Figure 3.
   - Entire stimulus: the question description and the four multiple choices.
   - Question: the question displayed at the top of the stimulus.
   - Correct answer: the choice that represents the correct answer.

TABLE III
METRICS FOR VISUAL EFFORT CALCULATION.

Numbers of eye fixations	
$FC(Q) = \sum_{a \in \text{ tasks, all answers}} f(a)$	(1)
$FR(correct) = \frac{\sum_{a \in \text{ correct answer}} f(a)}{\sum_{a \in \text{ correct answer} \cup \text{ distractors}} f(a)}$	(2)
$FR(distractors) = \frac{\sum_{a \in \text{ distractors}} f(a)}{\sum_{a \in \text{ correct answer} \cup \text{ distractors}} f(a)}$	(3)

- Distractors: the three other incorrect choices corresponding to the irrelevant AOIs.

We compute the visual effort using the metrics presented by Sharif *et al.* [25]. These metrics are based on the number of eye fixation. Table III shows their formula. A higher number of fixations indicates more effort to answer a question.

- Fixation Count on Question Stimulus $FC(Q)$: the total number of fixations on all six AOI for the entire stimulus. This stimulus is represented by the letter Q.
- Fixation Rate on Correct Identifier $FR(correct)$: the total number of fixations on the correct answer with respect to all four choices on the question stimulus.
- Fixation Rate on Distractors $FR(distractors)$: the total number of fixations on the incorrect choices with respect to all four choices on the question stimulus.

Mitigating variables are variables that might impact the effect of the independent variables on the dependent variables. In this experiment, we used a questionnaire to collect our two mitigating variables:

- Level of Study: values for this variable are B.Sc., M.Sc., and Ph.D.
- Style preferences: the values for this variable can be CC, US or None.

D. Subjects' Demography

The study participants are 24 volunteers, nine female subjects and 15 male subjects. The subjects were in B.Sc., M.Sc., and Ph.D. programs in the Department of Computer and Software Engineering at École Polytechnique de Montréal. We asked participants about their style preference. The 29% has no preferences (3 female subjects and 4 male subjects)while the other 71% preferred CC. We received the agreement from the Ethical Review Board of École Polytechnique de Montréal to perform and publish this study, which results are anonymous. The subject demographic is presented in Table IV.

E. Procedure

We conduct the experiment in a quiet, small room where the eye-tracking system is installed. We use a 27" LCD screen to show the stimuli while the subjects were seated approximately 70 cm away from the screen in a comfortable chair with arms and head rests. Before running the experiment, we briefly give a tutorial to explain the procedure of the experiment and the eye-tracking system (*e.g.*, how it works and what information is gathered by the eye-tracker). We also explain that the experiment consists of three pieces of source code and that subjects must answer five questions for each piece. We provide no explanation to subjects on the particular goal of the experiment. We also ask the subjects to fill a pre-experiment questionnaire to gather their basic information, such as gender and the level of study.

We ask each subject to read three different pieces of source code and recall the name of identifiers. We assign randomly each task to one of the two treatments (CC or US) in a way to avoid learning and hopefully maximize the possibility to observe a gender related difference.

For each subject, we first calibrate the eye-tracking system. Then, we present 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. No time limit is set but we asked subjects to answer the questions as soon as possible. The source code of each Java program is displayed. When subjects finish reading the source code, they press the "space" key to go to the next screen which contains one question and four choices, press space again, which displays a blank screen, and write down their answer to the question. To answer the question, *i.e.*, choose one of the four alternatives, subjects must recall the correct name of the identifier that performs a specific task in the program. Once a task finished and the answer given, subjects press the "space" key to go to the next Java program.

When subjects complete the three tasks, we ask them to answer the post-experiment questionnaire. The eye-tracking experiment took 25 minutes in average to complete.

TABLE IV
SUBJECTS' DEMOGRAPHY

Subjects' Demography				
Academic Background			**Gender**	
Ph.D.	M.Sc.	B.Sc	Male	Female
11	10	3	15	9

F. Eye-tracking System

FaceLAB from *Seeing Machine*[2] is a video-based remote eye-tracking system that we use for this experiment. It consists of two built-in cameras, one infrared pad, and one computer. By capturing subjects' head using facial features, including nose, eye-brows, and lips, FaceLAB tracks subjects' eye-movements. FaceLAB transmits eye-movements data to a data visualization tool, Gaze Tracker from *Eye Response*[3]. GazeTracker stores gaze the fixations and saccades associated with each image and displays all fixations in the foreground.

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

TABLE V
MAIN FEATURES OF THE EXPERIMENT.

Number of subjects (#)	24
Number of female subjects	9
Number of male subjects	15
Number of CC-related questions	147
Number of US-related questions	200
Total time of eye-tracking (hours)	11
Total number of fixations (#)	28,881

TABLE VI
SUBJECTS' CONTINGENCY TABLE FOR CC AND US IDENTIFIERS.

	Answers			
	Female		Male	
	Correct	Wrong	Correct	Wrong
CC	52	13	75	15
US	58	10	91	41

We use two screens 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 Java programs and the questions to the subjects.

G. Analysis Tool

When all the subjects completed their experiments, we used the Taupe system [8] to analyze the collected data. We defined the Areas Of Interest (AOI) on the stimuli and the subjects characteristics. Then, Taupe provided the results in CSV files that we exported to R [27] to apply statistic analyses.

IV. ANALYSIS AND RESULTS

In this section, we report hypotheses testing and discuss the results of our experiment. We use a range of analyses including the non-parametric, un-paired Wilcoxon test, logistic regression models [17], [30] as a proxy of correlation for dichotomous variables, linear regression models [7], [30], and contingency tables [30] to study and analyze the collected data. All data is available on-line[4].

A. Percentages of Correct Answers

Table V summarizes the collected data. Each subject, when answering a question, gave either a correct or a wrong answer: Table VI is the contingency table reporting for the two populations the numbers of correct and wrong answers for both CC and US identifiers. We use this data to perform a proportion test, to test whether the proportion of correct (wrong) answers in the overall population as well as in the two sub-populations (males and females) for the two treatments, CC and US, is the same.

Indeed, when we consider all subjects as part of the same population, we cannot reject the null hypothesis that CC and US correct answers have the same proportion. However, if we limit ourselves to the male population, we clearly see that male subjects prefer CC style vs. US style identifiers as they (roughly) gave three time more wrong answers when the code was written with US identifiers. This observation points to a lacking of training on US style coding style, *i.e.*, though male and female subjects have about the same training, male subjects seem less at ease with US identifiers.

[4]http://www.ptidej.net/download/experiments/icpc12b/

TABLE VII
BEST LOGISTIC REGRESSION ON THE EXPERIMENT POPULATION (AIC 370.01)

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.2884	0.2458	5.24	0.0000
Time	8.698e-06	3.459e-06	2.51	0.0119
GenderMale	-0.4631	0.2793	-1.66	0.0973

TABLE VIII
ALTERNATIVE LOGISTIC REGRESSION ON THE EXPERIMENT POPULATION (AIC 373.59)

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.3597	0.2433	5.59	0.0000
FR(Correct)	0.7987	0.3551	2.25	0.0245
GenderMale	-0.6086	0.2845	-2.14	0.0324

However, a more thorough analysis based on logistic regression sheds a different light. We use logistic regression analysis as a proxy for correlation when the response is a dichotomous variable as it is in the case of correct or wrong answers. Table VII and VIII report the variables retained by the logistic regression models built on the whole population of subjects along with the p-values and the AIC criterion. We observe that gender plays a marginal role in both models and that AIC value are close. We computed a logistic regression using R; the gender was coded as a factor with two levels "GenderMale" and "GenderFemale"; R just reports one of the two levels meaning that the model for a female is obtained by removing the "GenderMale" factor. In other words, in both models, the male coefficient is negative, *i.e.*, GenderMale lowers the probability and thus it impacts negatively the results of the models. Considering the best model using the AIC criterion in Table VII, although the time required to answer a question plays a role and as a single coefficient is statistically relevant, this time actually only marginally affects the probability as an increase in a unit of time would only marginally affect the probability because the coefficient is of the order of one over one million. If we consider the slightly worse model in Table VIII, we see that the effort spent on the correct answer, $FR(correct)$, increases the probability of a correct answer.

We do not make any claim on the validity of these models but we found their coefficient puzzling and, thus, we decided to study the two different sub-populations of male and female subjects independently, with the same type of analysis. We report the results of this analysis in Tables IX and X for female and male subjects, respectively. For both sub-populations, AIC criterion improves (more for the female sub-population than for the male one), supporting the heterogeneity of the population and further justifying the presence of gender in the models in Tables VII and VIII.

We can consider the two models in Tables IX and X as derived of the models in Tables VII and VIII. Yet, while for

TABLE IX
LOGISTIC REGRESSION ON THE FEMALE SUB-POPULATION (AIC 116.83)

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.9707	0.2884	3.37	0.0008
FR(Correct)	3.1580	1.3328	2.37	0.0178

TABLE X
LOGISTIC REGRESSION ON THE MALE SUB-POPULATION (AIC 244.25)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3637	0.2984	4.57	0.0000
US-style	-0.8522	0.3437	-2.48	0.0131
Time	8.996e-06	4.146e-06	2.17	0.0300

TABLE XI
VALUES FOR PERCENTAGES OF CORRECT ANSWERS AND REQUIRED TIME TO COMPARE MALE AND FEMALE SUBJECTS' ACCURACY AND SPEED.

	Accuracy %	**Required time (min)**
Male Subjects	0.745	5.94
Female Subjects	0.827	7.18

TABLE XII
WILCOXON P-VALUES (α = 0.05) FOR ACCURACY AND REQUIRED TIME.

Required Time		**Accuracy**
Source Code Stimulus	**Question Stimulus**	
0.2107	0.3472	0.3217

female subjects, their efforts on the correct answers help to explain the data (see Table IX); for the male subjects, only time plays a role and its effect is less strong than the coding style. Indeed, US-style has a much stronger effect in the model than time. Table VI reports that male subjects gave $75 + 91 = 166$ correct answers out of $75 + 91 + 15 + 41 = 222$, which means 75% correct answers. Thus, male subjects performed better than choosing a random choice because a purely random choice would have given a percentage of correct answers close to 25%. However, male subjects seem not to have invested too much time in pondering their answers. It is possible that they did not really-seriously performed their tasks, as even the US-style coefficient is smaller than the intercept. Yet, we cannot explain how it was possible for male subjects to achieve a percentage of correct answers of 75% on US-style identifiers without spending time to consider all the possible answers. We believe that other experiments must investigate this observation and attempt to bring some answers backed with evidences.

Table VI reports that female subjects gave $52 + 58 = 110$ correct answers out of $52 + 58 + 13 + 10 = 133$, which means 83% correct answers. Their errors on US-style identifiers was much lower than that of male subjects although male and female subjects have comparable curricula. When considering time in Tables VII and X, time plays no role for both male and female subjects.

B. Accuracy and Required Time (Speed)

In this subsection, we now model the subjects' accuracy, *i.e.*, precision [1]: the numbers of correct answers divided by the numbers of questions answered. In agreement with Table VI and the considerations in the previous sub-section, the values reported in Table XI show that female subjects performed 8% more accurately than male subjects but that they spent 18% more time than male subjects. However, Table XII reports that there is no significant differences between male and female subjects for their accuracy and the required time (speed) performing the tasks. Therefore, we cannot reject the null hypotheses $H\alpha_{01}$ and $H\alpha_{02}$.

The values in Table XI also mean that gender, for the current tasks and population, does neither significantly affect the subjects' accuracy nor required time by the subjects performing the tasks. These figures are also supported by the male and female subjects' Required time distributions not reported here for lack of space and because these distributions do not bring any additional insight.

As expected from the previous sub-section, male subjects' accuracy are not the same when comparing CC and US identifiers. The Wilcoxon test rejects the null hypothesis that the two sets have the same mean distribution with a confidence level of 0.09 (thus not a 95%-confidence level but a 90% one). We make a similar observation for time: the times spend by male subjects with CC and US identifiers are close (but at a 90%-confidence level, p-value: 0.06128). In conclusion, male subjects spent more time on US identifiers.

The male and female subjects' accuracy is the same with CC style identifiers but not with US style identifiers. The Wilcoxon test rejects the null hypothesis that the two sub-populations performed with the same accuracy with a p-value of 0.05376. Thus, strictly speaking, we cannot rule out the hypothesis at a 95%-confidence level but are quite close.

When considering the tradeoff between speed and accuracy, even if the overall difference is not significant, we observe a trend in the male sub-population as shown in Figure 4 while such a trend is not present for female subjects. We computed Figure 4 as follows: for each subject, we calculate the average total time to perform the experiment and the percentage of their correct answers (accuracy) and we draw a scatter-plot with a super-imposed trend line. In Figure 4, the horizontal axis represents the time and the vertical axis represents the accuracy. The trend for male subjects is just apparent because for both sub-populations (male and female subjects), the variable selection of the linear models only retain the intercepts. Thus, there is no correlation between time (or other measured variables) and accuracy. In conclusion, for male subjects, the data shows the apparent trend that they are more accurate if they spend more time, which is in agreement with the previous results of the following studies. Uwano *et al.* [29] observed that the longer a subject reads the code, the more efficiently s/he finds the defects in the source code. Moreover, Sharif *et al.* [26] reported that there is a correlation between scan time and defect detection time. However, our results are not statistically supported either male or female subjects.

C. Visual Effort

When considering the entire population, no significative linear model can explain accuracy with respect to visual effort

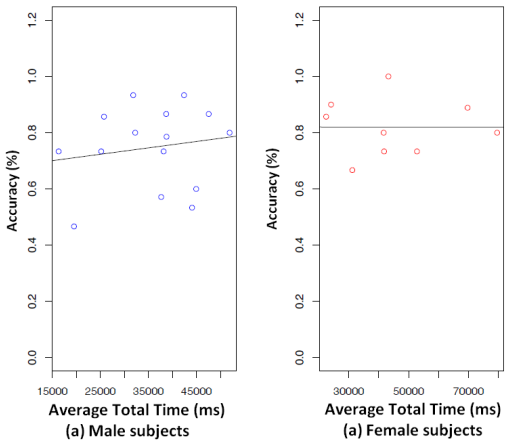

Fig. 4. Speed-Accuracy tradeoff for male and female subjects.

TABLE XIII
WILCOXON P-VALUES ($\alpha = 0.05$) FOR EACH VISUAL EFFORT MEASURE

Visual Effort Metrics			
FC(Q)	**FR(correct)**	**FR(distractor)**	**Convex hull**
0.7884	0.1737	0.006	0.7685

but we can gain insight by considering single variable and sub-populations. Table XIII reports the p-values for Wilcoxon tests comparing male versus female subjects on the different effort variables. We cannot observe statistical differences for the overall effort spent, $FC(Q)$, and the effort spent on the correct answers, $FR(correct)$. Interestingly, there is a statistically-significant difference for the effort spent on the wrong answers, the distractors, $FR(distractors)$. Moreover, in Table XIII, we observe that female subjects put more visual attention (effort) on distractors (irrelevant areas of interest) than male subjects. We can reject the null hypothesis H_{03}. Our rejection of the null hypothesis is also supported by Figure 5, which shows the box-plots of the three independent variables related to effort for male and female subjects. Figure 6 shows the heatmaps for a male and a female subject for one question. Heatmap is a color spectrum that represents the intensity of fixations. The heatmap for our female subject shows that her fixations are scattered through all choices while for our male subject, his fixations are focused on the correct choice.

We gain interesting insight again as in previous sub-sections when considering the male and female sub-populations independently. While we cannot model the accuracy of the male sub-population with a linear model and the dependent variables, we obtain a model with interacting variables for the female sub-population, shown in Table XV. The model is highly significant, it explains 84% of variability and support the conjecture that it is the complex interplay between the visual effort on correct answers and wrong answers that helps the female subjects to obtain a high accuracy. We will investigate the rationale for this observation (and lack thereof for male subjects) in future work. In particular, we have only nine female subjects and the model is likely over-fitted. Table

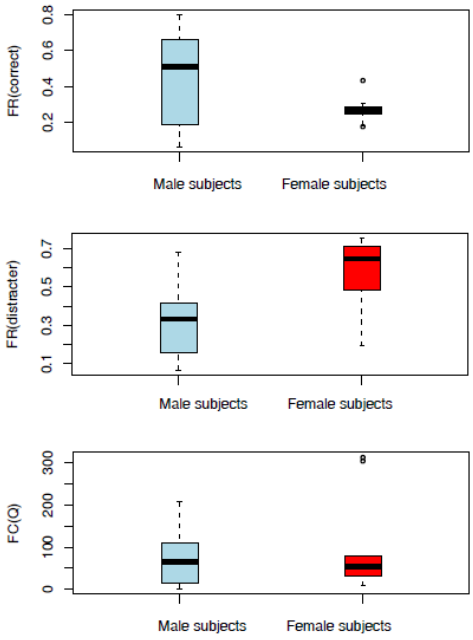

Fig. 5. Data distribution for visual effort.

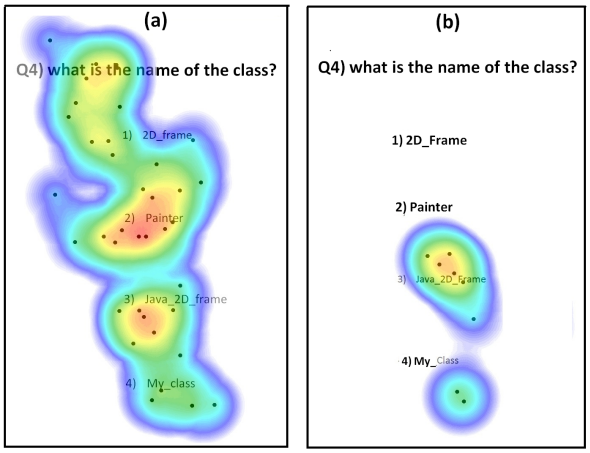

Fig. 6. Heatmap illustration of a female subject (a) and a male subject (b).

XIV reports a simplified model—not modelling interaction. This model explains less the variance as the adjusted R^2 is 45% compared to the 84% of the model in Table XV. Yet, the coefficients of the model in Table XIV are expected: the more the female subjects concentrate on distractors, the lower their accuracy while the more they concentrate on the correct answers, the better their accuracy.

V. THREATS TO VALIDITY

Different studies [2], [4], [16] report that females are usually less confident than males while using different programming environments. They report that self-efficacy—a person's confidence about her/his own competence in performing a specific task successfully—impacts the amount of effort that is spent and the strategies that are used to solve a problem. The root

TABLE XIV
EFFORT–ACCURACY MODEL FOR THE FEMALE SUB-POPULATION WITH NO INTERACTION (ADJUSTED R^2 OF 0.45)

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.7970	0.1470	5.42	0.0016
FR(distractor)	-0.5545	0.2124	-2.61	0.0401
FR(correct)	1.1711	0.5174	2.26	0.0642

TABLE XV
EFFORT–ACCURACY MODEL FOR THE FEMALE SUB-POPULATION (ADJUSTED R^2 OF 0.84)

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.9259	0.2984	6.45	0.0013
FR(distractor)	-2.8541	0.5969	-4.78	0.0050
FR(correct)	-4.3654	1.4377	-3.04	0.0289
FR(distractor):FR(correct)	10.5043	2.6753	3.93	0.0111

cause of differences in males' and females' behaviors and possibly their different abilities are likely to go back to their different roles in society and society evolution [11], [15]. Gathering seeds or harvesting require different sets of skills; accuracy may not be so relevant to harvest but picking seeds or identifying a poisonous or non-edible plant has always been critical for pre-industrial societies [11]. Over thousands of years of evolution, social training shaped and developed different skills and abilities in males and females [11], [15].

In our experiment, female subjects put more effort reading and analyzing distractors. Consequently, we could explain this observation by the lower level of self-efficacy perceived by the female subjects when compared to that of male subjects. Female subjects wanted to make sure that they find the correct answers precisely so they investigated all choices before making their decisions. In addition, there is no significant difference between the time that female or male subjects spent to perform the comprehension tasks. Thus, even though female subjects devote visual attention not only on the correct answers but also on distractors, they do not spend more time than male subjects. Nevertheless they are more effective as they obtained a better accuracy and a better precision.

One threat of validity is the fact that our experiment was designed by a woman and this may bias the experiment towards the female population. We believe this risk is somehow mitigated by the measured variables and the constraint on the code size that must fit into one screen. The risk is also intrinsic to any experiment of the past as we are not aware of experiments designed by mixed teams.

1) Internal validity: we consider three threats to the internal validity: maturation, instrumentation, and diffusion of the treatments. We mitigate the impact of maturation by assigning the different pieces of source code randomly to our subjects. This random ordering prevents fatigue effect concerning the code given at the end. The instrumentation threat is related to the equipment that we use in our study. We use a video-based eye-tracking system that does not involve any heavy goggle. Subjects can move their head easily without changing the calibration of camera. We also use a dentist chair to support subjects' neck to make them comfortable and avoid fatigue. To mitigate the possible diffusion of the treatments, we asked our subjects not to talk about the experiment with the other subjects.
2) Construct validity: we consider two threats to the construct validity: hypothesis guessing and apprehension. We did not inform the subjects about the precise goal of the experiment to avoid hypothesis guessing. We explained them the process of performing the experiment, the number of systems, and questions that they must answer. Regarding the apprehension threat, we explained how an eye-tracker system works and we assured them that there is no physical risk working with these instruments. We did not set a time limit, we asked subjects to answer the questions as soon as they can.
3) External validity: this threat is related to the generalization of our results. We used students as subjects in our study and we did not distinguish novices and experts. Kitchenham *et al.* [20] 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 and, so, are relatively close to the population of interest." In this paper, our subjects are graduate students with the good knowledge of Java. The number of our subjects appears to be low, we have 24 subjects, yet it is much more than any other studies. Sharif *et al.* [25] 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 use the non-parametric non-paired Wilcoxon statistical test to determine significance because we assume that our data is not normally distributed. To ensure the reliability of our measures, we chose well-documented measures from the previous work of Sharif *et al.* [25] and made sure that the eye-tracker is well calibrated for every subject before collecting data.

VI. CONCLUSION AND FUTURE WORK

We designed and performed an eye-tracking experiment to investigate the impact of gender on the performance of developers during code reading and program understanding activities. We also examined the effect of identifier style: camel case (CC) vs. underscore (US).

Our findings support the belief that CC and US styles do not impact the subjects' effort and program comprehension. We found no significant differences between identifier styles when considering accuracy and effort and gender. However, a more fine-grained analysis of the subjects' visual effort on correct and wrong answers revealed unexpected details. Indeed, the variability of this visual effort is significantly different between male and female subjects. While the time spent by male and female subjects is not significantly different, male and female subjects focused differently on the alternative

answers that we proposed them. Female subjects spent more effort on the wrong answers than male subjects, which can explain the higher accuracy of female subjects. Thus, female subjects seem to carefully weight all options and rule out wrong answers while male subjects seem to quickly set their minds on some answers, possibly the wrong ones.

First, we found a statistically strong interaction between accuracy, effort spent on distractors, and correct answers when modelling the female subjects' accuracy. Second, no correlation exists between visual effort on correct answers or distractors and the male subjects' accuracy. Consequently, for female subjects, it is not how much visual effort they spend on distractor or on correct answers but rather a mix of the two, *i.e.*, the complex pondering of correct and wrong answers, that describe best their accuracy. Yet, we must be careful because only nine female subjects participated to our study and, although this number is much higher than that of any previously reported studies, we cannot consider the population large enough to generalize. Our findings should be considered more as a hint to suggest further studies concerning the role of gender in software engineering rather than a general truth.

Therefore, in future work, we will (1) replicate our experiment with more female subjects, (2) conduct other eye-tracking experiments to (i) identify the specific strategy used by male subjects (if any), (ii) the factors affecting the male subjects' accuracy, (iii) the reasons why female subjects tend to focus more than male subjects on wrong answers, and (iv) the implication of our findings on the design of methods, techniques, and tools. We plan to perform an additional experiment without eye-tracking with more realistic tasks involving file switching and apply qualitative analyses and compare its results with the results of the quantitative analysis presented in this paper.

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 the Agence Universitaire de la Francophonie.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] L. Beckwith, M. Burnett, S. Wiedenbeck, C. Cook, S. Sorte, and M. Hastings, "Effectiveness of end-user debugging software features: Are there gender issues," in *In Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*. Press, 2005, pp. 869–878.

[3] D. Binkley, M. Davis, D. Lawrie, and C. Morrell, "To camelcase or under_score," in *IEEE 17th International Conference on Program Comprehension (ICPC)*. IEEE, 2009, pp. 158–167.

[4] M. Burnett, S. Fleming, S. Iqbal, G. Venolia, V. Rajaram, U. Farooq, V. Grigoreanu, and M. Czerwinski, "Gender differences and programming environments: Across programming populations," in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 2010, pp. 1–10.

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

[6] B. Caprile and P. Tonella, "Restructuring program identifier names," in *Proc. Int'l Conf. Software Maintenance (ICSM)*. IEEE Computer Society Press, 2000.

[7] S. Chatterjee and B. Price, *Regression Analysis by Example*. Wiley, 1977.

[8] B. De Smet, L. Lempereur, Z. Sharafi, Y.-G. Guéhéneuc, G. Antoniol, and N. Habra, "Taupe: Visualising and analysing eye-tracking data," 2011, submitted to Elsevier Editorial System for Science of Computer Programming.

[9] F. Deissenbock and M. Pizka, "Concise and consistent naming," in *Proceedings of the International Workshop on Program Comprehension*. IEEE CS Press, 2005, pp. 97 – 106.

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

[11] H. Fisher, *Anatomy of Love: A Natural History of Mating, Marriage, and Why We Stray*. The random house publishing group, 1994.

[12] M. Fisher, A. Cox, and L. Zhao, "Using sex differences to link spatial cognition and program comprehension," in *IEEE 22nd International Conference on Software Maintenance*. IEEE Computer Society, 2006, pp. 289–298.

[13] 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.

[14] V. Grigoreanu, J. Brundage, E. Bahna, M. Burnett, P. ElRif, and J. Snover, "Males and females script debugging strategies," *End-User Development*, pp. 205–224, 2009.

[15] M. Harris, *The cannibals and the kings: Origins of Cultures*. Vintage book edition, 1991.

[16] K. Hartzel, "How self-efficacy and gender issues affect software adoption and use," *Communications of the ACM*, vol. 46, no. 9, pp. 167–171, 2003.

[17] D. Hosmer and S. Lemeshow, *Applied Logistic Regression (2nd Edition)*. Wiley, 2000.

[18] D. M. Jones, 2003, the New C Standard (Identifiers) An Economic and Cultural Commentary.

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

[20] 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.

[21] D. Lawrie, C. Morrell, H. Feild, and D. Binkley, "Effective identifier names for comprehension and memory," *Innovations in Systems and Software Engineering*, vol. 3, no. 4, pp. 303–318, 2007.

[22] J. Meyers-Levy, *Gender Differences in Information Processing: A SelectivityInterpretation.* P. Cafferata and A. Tybout, (Eds) Cognitive and Affective Responses to Advertising. Lexington Books, 1989.

[23] E. O'Donnell and E. Johnson, "The effects of auditor gender and task complexity on information processing efficiency," *International Journal of Auditing*, vol. 5, no. 2, pp. 91–105, 2001.

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

[25] 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.

[26] 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.

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

[28] G. Torkzadeh and X. Koufteros, "Factorial validity of a computer self-efficacy scale and the impact of computer training," *Educational and Psychological Measurement*, vol. 54, no. 3, pp. 813–821, 1994.

[29] 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.

[30] W. N. Venables and B. D. Ripley, *Modern Applied Statistic with S-PLUS*. Springer, 1999.