A Comparison of Paper Sketch and Interactive Wireframe by Eye Movements Analysis, Survey, and Interview

Suzanne Kieffer
Université catholique de Louvain (UCLouvain),
Institute for Language and Communication (ILC)
suzanne.kieffer@uclouvain.be

Jean Vanderdonckt
Université catholique de Louvain (UCLouvain),
Louvain Research Inst. in Manage. and Org. (LouRIM)
jean.vanderdonckt@uclouvain.be

Abstract

Eye movement-based analyses have been extensively performed on graphical user interface designs, mainly on high-fidelity prototypes such as coded prototypes. However, practitioners usually initiate the development life cycle with low-fidelity prototypes, such as mock-ups or sketches. Since little or no eye movement analysis has been performed on the latter, would eye tracking transpose its benefits from high- to low-fidelity prototypes and produce different results? To bridge this gap, we performed an eye movement-based analysis that compares gaze point indexes, gaze event types and durations, fixation and saccade indexes produced by N=8 participants between two treatments, a paper prototype vs. a wireframe. The paper also reports a qualitative analysis based on the answers provided by these participants in a semi-directed interview and on a perceived usability questionnaire with 14 items. Due to its interactivity, the wireframe seems to foster a more exploratory approach to design (e.g., testing and navigating more extensively) than the paper prototype.

1. Introduction

Eye movement-based analysis has long been used for (re)designing (Bojko, 2006), generating (Cheng & Dey, 2019), and evaluating graphic user interface (GUI) prototypes based on several measures, such as eye movement locations and scan paths, and has been proven to be valid (Goldberg & Kotval, 1999). The lack of usability often results in more fixations and longer scan paths that cover larger areas. In these eye movement analyses, prototypes are primarily high-fidelity prototypes such as coded prototypes, which contrasts to the usual practice where practitioners start by designing low-fidelity prototypes such as sketches: based on user requirements, practitioners select simplified elements with limited functions, determine their rough layout into a low-fidelity prototype and iteratively test-and-refine the prototype (Kieffer, Rukonić, Kervyn de Meerendré, & Vanderdonckt, 2020). As soon as major user requirements are satisfied, they turn the low-fidelity prototype into a high-fidelity prototype by specifying the interaction flow and applying rules (*e.g.*, colors, fonts, typography, icons) from a style guide.

To bridge the gap between high-fidelity prototypes typically analyzed by eye tracking and low-fidelity prototypes typically serving to initiate iterative design, how does an eye movement analysis transpose to lower levels of fidelity in GUI prototyping? Both levels of fidelity, *i.e.* low and high, have their own advantages and shortcomings. If we want to benefit from the advantages brought by low-fidelity prototyping, could we perform eye-movement analyses on such prototypes as we do on high-fidelity ones? What do we win and what do we lose? In other words, are the results of an eye-movement analysis sensitive to the level of fidelity? There is little or no work performing eye-movement analysis on low-fidelity GUI prototypes to address these questions.

In order to address these questions, this paper makes the following contribution: it performs an eye movement-based analysis comparing two treatments: a low-fidelity paper sketch *vs.* a high-fidelity wireframe. A quantitative analysis compares the respective gaze point indexes, gaze event types and durations, fixation, and saccade indexes produced by N=8 participants in the two treatments, and a qualitative analysis compares the answers provided by these participants in a perceived usability questionnaire with 14 items adapted to these treatments. The remainder of this paper is structured as follows: Section 2 discusses previous eye-tracking work performed on GUI prototypes and reviews the salient advantages of low-fidelity *vs.* high-fidelity. Section 3 defines the methods, experimental protocol and design for the methodology we used in a controlled eye tracking experiment. Section 4 discusses the results obtained for both quantitative and qualitative analyses and discusses the limits of this experiment with respect to threats to validity. Section 5 concludes the paper and suggests some future avenues for this work.

URI: https://hdl.handle.net/10125/103128
978-0-9981331-6-4

Dimension	Paper sketch treatment	Wireframe treatment
Refinement	Lo-Fi: GUI elements sketched by hand	Hi-Fi: GUI elements close to their final L&F
Breadth	Lo-Fi: one feature	Lo-Fi: one feature
Depth	Hi-Fi: detailed feature	Hi-Fi: detailed feature
Interactivity	Lo-Fi: no dynamic aspects	Hi-Fi: point and click are simulated
Data model	Lo-Fi: incomplete	Hi-Fi: data are simulated

Table 1: LoF in both Treatments broken down according to McCurdy *et al.*'s five dimensions (McCurdy et al., 2006).

2. Related Work

2.1. Prototypes and their Level of Fidelity

The *Level of Fidelity* (LoF) of a GUI prototype corresponds to the precision with which it reproduces the realm of the desired GUI in the expected context of use. Coyette, Kieffer, and Vanderdonckt identify three levels: low, medium and high. *Low-fidelity* (Lo-Fi) prototypes capture the general information needed to obtain an overall understanding of the final GUI, removing all unnecessary details. The techniques for Lo-Fi prototyping include "paper and pencil" (Bailey & Konstan, 2003), "whiteboard/ blackboard"(Cherubini, Venolia, DeLine, & Ko, 2007), and "post-it" (Plimmer & Apperley, 2003). *Medium-Fidelity* (Me-Fi) prototypes give importance to content and interaction, while keeping other minor details such as typography or colors secondary. A typical example is Visio, where only the type, size, and GUI contents are specified graphically with a stencil. *High-fidelity* (Hi-Fi) prototypes look and feel as close as possible to the final GUI, are usable and almost complete in terms of functions and behavior. Designers produce an accurate image of the GUI through software such as Photoshop or code the GUI in a programming language. McCurdy et al. characterize the LoF of prototypes along five dimensions, each being Lo-Fi or Hi-Fi: (1) visual refinement ranging from sketches to pixel-precise prototypes; (2) breadth of functionality expressing how broad represented functions are; (3) depth of functionality expressing to what level of detail a feature is represented; (4) interactivity expressing how interactive GUI elements are; and (5) richness of the data model expressing how representative of the actual application domain data are. A combination of LoFs appears to make sense, provided that the Lo-Fi advantages are maintained (Petrie & Schneider, 2006). Hi-Fi elements can be integrated into the prototype while keeping a Lo-Fi for other parts without compromising the overall quality and avoiding early commitment. Table 1 compares the LoF of the paper sketch and the wireframe used in this study: most of the Lo-Fi in the paper sketch, respectively, and Hi-Fi in the wireframe allows us to define the paper sketch as Lo-Fi and the wireframe as Hi-Fi.

2.2. Level of Fidelity and GUI Design

Since the late 1980s, the question of which tool should be used for designing a GUI prototype has animated the scientific literature by investigating various factors such as media (Bailey & Konstan, 2003), software, process (Sangiorgi, Beuvens, & Vanderdonckt, 2012). Since these factors are numerous and inter-weaved, the LoF is at the heart of this debate. For example, Bailey *et al.* (Bailey & Konstan, 2003) advocate GUI prototyping with a dedicated software, as it allows for expressing more interaction features than a general-purpose authoring system, which in turn is better than a paper-and-pencil approach in this regard. However, Sefelin, Tscheligi, Tscheligi, and Giller report only insignificant differences between a sketch drawn on paper and a wireframe drawn via software regarding the number of both detected usability problems and suggestions for improvement. Therefore, paper sketches should be preferred when stakeholders are not experienced with prototyping software or do not want to invest their time in such a way, as paper sketches are much faster and cheaper to produce than wireframes. A wireframe should be recommended when the intended interaction is too dynamic to be illustrated on paper. Grip classifies prototyping tools (Van den Bergh, Sahni, Haesen, Luyten, & Coninx, 2011) to underline their ability to support GUI prototyping in various LoFs.

Switching from Lo- to Hi-Fi prototyping does not affect the perceived usability of the prototype (Wiklund, Thurrott, & Dumas, 1992). In fact, a similar proportion of participants detect individual usability problems in Lo-Fi and Hi-Fi prototypes, further strengthening the message that it is not necessary to wait for a Hi-Fi prototype to submit a design to usability testing (Virzi, Sokolov, & Karis, 1996). Further, despite differences in interaction style, usability testing and detected usability problems are independent of whether GUI prototypes are presented on paper or software (Walker, Takayama, & Landay, 2002). Moreover, a comparison between an early Lo-Fi and a final Hi-Fi GUI prototype showed no significant differences in perceived LoF (Uebelbacher, Sonderegger, & Sauer, 2013), the only difference being the beneficial presence of the experimenter in the Lo-Fi condition. However, Lo-Fi offers some advantages

compared to Hi-Fi regarding the content of user feedback (Sangiorgi et al., 2012): participants presented with a Lo-Fi prototype tend to focus on GUI elements and their behavior rather than on details such as color scheme or typography, irrelevant during early design.

These observations are not comparable with design diagrams (Yeung, Plimmer, Lobb, & Elliffe, 2008): the LoF of such diagram affects design performance and user perception. On the one hand, Lo-Fi diagrams were more effective for early design tasks, such as GUI early design, than the corresponding Hi-Fi prototypes. On the other hand, participants preferred higher-fidelity diagrams because they feel Lo-Fi versions are considered unprofessional, a contradictory reason to the observation that participants preferred Lo-Fi diagrams in the early design stage because they feel they are open, flexible, and unfinished (Rudd, Stern, & Isensee, 1996). By replication, Wohlin *et al.* (Wohlin et al., 2012) determined that a Hi-Fi prototype influences the ecological validity (Kieffer et al., 2020) with respect to its initial Lo-Fi version. Lo-Fi prototypes are an important source of inspiration by exposing experienced or non-expert practitioners to many prototypes.

2.3. Eye Tracking in Development Tasks

To the best of our knowledge, eye movement tracking has not been performed on various LoFs for GUI design. While knowledge about eye movements behavior on high-fidelity GUIs is abundant (Cheng & Dey, 2019; Goldberg & Kotval, 1999; R. J. K. Jacob, 1990; R. J. Jacob & Karn, 2003), little is known about them when the LoF decreases from high-fidelity. Yet, this knowledge is important for practitioners to apply eye tracking as early as they start designing GUI prototypes. Holmes and Zanker used an eye tracker to capture gaze fixations in a GUI prototype to confirm that this visual measure reliably represents the aesthetic quality of the GUI. Furthermore, Cheng and Dey exploited an interactive genetic algorithm to compute a GUI design fitness function based on the number of fixations, the fixation duration, and the first fixation on a target GUI element. On the basis of these measures, they inferred end users' preferences and adapted their GUI.

In contrast, eye movement tracking has been performed on various software artifacts, such as programming code, configurations, models to support software engineering and programming tasks, and high-fidelity GUI designs. For instance, iTrace (Shaffer et al., 2015) is an Eclipse plug-in that incorporates eye tracking in an Integrated Development Environment (IDE) to record the developer's eye movements while developing and as such, that allows to study eye movement patterns of software engineers manipulating software artifacts (*e.g.*, Java code, text/HTML/XML documents, diagrams). Therefore, iTrace has a wide range of applications, provided that the development task remains in the realm of the IDE. Since such an IDE typically generates multiple graphical and textual views of programs, these views need to be coordinated with the main program understanding. To analyze this coordination, data mining techniques detect high-frequency patterns from eye movements. Different visual patterns were found among the participants based on their programming experience, familiarity with the IDE and debugging performance (Hejmady & Narayanan, 2012). Similarly, Sharif, Falcone, and Maletic confirmed that the longer time a developers spends in initially scanning the software code, the quicker they find the defect based on the scan time. Program comprehension (Bednarik & Tukiainen, 2006) and code summarization tasks (Abid, Maletic, & Sharif, 2019) can be also approached by eye movement analysis. Approaches from research on natural-language text reading is also applicable to source code, however not without review (Busjahn, Bednarik, & Schulte, 2014). Among all these tasks, scanning and visual search are the most frequently investigated tasks. For example, Cutrell and Guan investigated how changing the presentation of search results influence their perception: adding information to the contextual snippet improved performance for the search task, but not for a navigation task. More specifically, for GUI tasks, Bojko proved that a simple redesign of the web page resulted in fewer fixations and dense areas on the heat maps when participants had to perform a series of search tasks. A visual feature-based attention prediction model for GUI elements of a web site is elaborated by Vidyapu, Vedula, and Bhattacharya based on a multi-class Support Vector Machine (MSVM) to learn using the visual features and their associated attention.

3. Methodology

3.1. Methods

To answer our research questions, we conducted a comparative study between a Lo-Fi and a Hi-Fi GUI prototype with $N=8$ participants involving eye tracking, survey, and interview. Eye tracking data allowed us to compare eye movements between treatments, survey data to assess perceived usability and LoF in each treatment, and interview data to compare the attitude of participants toward each GUI prototype (Tullis & Albert, 2013). Specifically, we asked participants to perform a guideline-based review of both a Lo-Fi paper sketch and a Hi-Fi wireframe (Table 1), referred hereafter to as "paper" and "wireframe" Treatments respectively.

3.2. Participants

We recruited eight participants (8 women) between 21 and 23 years of age (M=22) through a registered student mailing list, who had previous experience with GUI design and evaluation, as they had successfully completed a 5 ECTS course about GUI design and evaluation. No compensation was offered.

3.3. Stimuli and Experimental Design

To control the learning effect bias with the experimental task, we produced a GUI design for two imaginary mobile applications: one for shopping clothes and the other for tracking basketball teams, hereafter referred to as "fashion" and "basketball" Themes respectively. We chose them as they do not require any domain expertise and are rather popular. The "fashion" design mimics an electronic commerce application and is more interactive than the "basketball" design, which is information-oriented and focuses on visual searching. Each GUI design was produced on both a Lo-Fi paper sketch and a Hi-Fi wireframe drawn with Pencil Project V3.1.0, an open source multiplatform GUI prototyping tool to create mockups. Each GUI design had to include between three and five different screens. Transitions between screens had to be implemented in the wireframe. We initiated a call for design satisfying these requirements, scored each received prototype and kept the two pairs of prototypes having the highest score for both GUI prototypes. Furthermore, to control the order effect bias between treatments, we used a counterbalanced design (Campbell & Stanley, 1963) with a 2×2 Latin square arrangement between Treatments ("paper" *vs.* "wireframe") and Themes ("basketball" *vs.* "fashion"). Participants were randomized to a given experimental sequence (1^{st} treatment & 1^{st} theme ; 2^{nd} treatment & 2^{nd} theme), each treatment and each theme occurring in each position, first or second.

3.4. Apparatus

The scanned paper prototypes and the wireframe prototypes were displayed on a Dell Precision series 75100 (2.7 GHz Intel Core i7 processor, 8 Go 1600 MHz DDR3 RAM) computer equipped with a Tobii Pro X3-120 eye-tracker for browsing the stimuli with a screen resolution set to 2048 pixels × 1152 pixels and a timestamp of 8 msec. We used a separate laptop (MacBook Air, 1.8 GHz Intel Core i5 processor, 8 Go 1600 MHz DDR3 RAM, Intel HD Graphics 6000 graphic card, 13-inch screen) for filling out the forms.

3.5. Questionnaire and Interview Guide

We used a 14-item questionnaire to trigger an evaluation intention among participants, to compare the perceived usability between each treatment (items 1-9) and to compare the perceived LoF between treatment according to McCurdy et al. dimensions (items 10-14). Items 1-9 correspond to the nine research-based usability guidelines (Leavitt & Shneiderman, 2006) focused on GUI and scored the highest in terms of relative importance (5 on a scale of 1-5, 1 referring to less important guidelines and 5 to the most important guidelines). The nine items are: (1) Create a positive first impression of your Site, (2) Place important items consistently, (3) Place important items at top center, (4) Eliminate horizontal scrolling, (5) Use clear category labels, (6) Use clear link labels, (7) Make action sequences clear, (8) Organize information clearly, (9) Facilitate scanning. Items 10-14 correspond to the five prototype dimensions defined by McCurdy et al.. The five additional items are: (10) The prototype is visually refined, (11) The functionality represented within the prototype is broad, (12) Any one feature or sequence represented is detailed, (13) The elements captured and represented to the user by the prototype are interactive, (14) The data employed by the prototype are representative of the domain data. For each treatment, participants scored their level of agreement with each item on a 5-point Likert scale (1=strongly disagree to 5=strongly agree).

We conducted a semi-directive interview to ask participants about their experience with prototype selection. These self-reported data allowed us to compare the attitudes of the participants towards each prototype (Tullis & Albert, 2013). The interview guide involved three questions: (1) What strategies did you adopt to analyze the prototypes? (2) Which difficulties did you encounter during the analysis? (3) Which prototype was the easiest to analyse?

3.6. Procedure

Each participant carried out the experiment in our usability laboratory, according to the following steps:

1. *Consent form*: each participant was welcomed, signed a GDPR-compliant consent form, and completed a questionnaire on their background.
2. *Instructions*: each participant was instructed to read the written instructions and the questionnaire and was free to ask any question if needed.
3. *First analysis*: the first treatment was administrated and each participant was required to analyze the corresponding prototype displayed on the PC equipped with the eye tracker and to fill in

the questionnaire on another laptop. Participants were allowed to freely fill out the questionnaire while analyzing the GUI prototype.

4. *First interview*: the experimenter interviewed each participant regarding their eye paths on the prototype according to the interview guide.
5. *Second analysis*: the second treatment was administered similarly to the first one.
6. *Second interview*: the participant was interviewed a second time regarding the second prototype.

3.7. Data Management

In the data collection file, we encoded the independent variables as follows:

1. TREATMENT, a qualitative nominal variable that takes "paper" or "wireframe" as modalities, representing the GUI designed in Lo-Fi and Hi-Fi, respectively.
2. THEME, a qualitative nominal variable that takes the "basketball" or "fashion" as modalities, representing the GUI designed for the two aforementioned mobile applications.

4. Results and Discussion

We assigned an index to the participants as follows: P1, P2, P3 and P4 to the participants who performed the experimental tasks with "paper+basketball" and "wireframe+fashion" combination, and P5, P6, P7 and P8 to the participants who performed the experimental tasks with "paper+fashion" and "wireframe+basketball" combination. This indexing strategy is not to be mistaken for the random assignment of participants to a given treatment order. We exported the eye tracking data from the eye tracking software into tab-separated values (TSV) files, one for each session, and then converted them into MS Excel for further computation. Eye tracking data collection includes the following measures:

1. GAZE POINT INDEX, an instantaneous measurement of gaze events with a timestamp.
2. GAZE EVENT TYPE, a qualitative nominal variable with a value of "fixation" or "saccade".
3. GAZE EVENT DURATION, a non-null natural number measuring the duration of gaze points in milliseconds.
4. FIXATION INDEX, a non-null natural number measuring multiple gaze points with their spatial (x, y) coordinates, along with their starting and ending times. Fixations are usually understood as relatively stationary eye positions over an element lasting from 50 to 600 msec. *Fixation indexes are constructions, outputs of a mathematical algorithm that translates the sequence of raw gaze points into an associated sequence of fixations.*
5. SACCADE INDEX, a positive natural number that measures the type of eye movement used to move the fovea from one Point of Interest (POI) to another, based on the average duration of a saccade lasting from 20 to 40 msec.

We recorded in a single MS Excel file the perceived usability scores (1-5 ordinal scale) assigned by the participants to each of the 14 questionnaire items for each treatment ("paper" vs. "wireframe"). We transcribed the interviews into separate MS word files.

4.1. Eye-tracking Data

Analysis for Participants. From the length of the boxes and whiskers in the box plots displaying participants' gaze points (Fig. 1, index at the top and duration at the bottom), we conclude that there is an important inter-individual variability in both fixations and saccades. To confirm this observation, we computed the following series of inferential statistics tests on the fixation index and the duration of the gaze event per participant. We computed a Shapiro-Wilk test and a D'Agostino's K^2 test, a goodness-of-fit measure of departure from normality. Regarding both fixation index and gaze event duration, neither all participants taken together nor any participant taken in isolation are sample from a normal distribution, as both Shapiro-Wilk and D'Agostino tests are negative. Second, since distributions are not normal, we computed a Kruskal-Wallis H test and found a significant difference between all participants in both fixation index (H=1063.60, N=12849, *df*=7, α=.05, $p\leq.001^{***}$) and gaze event duration (H=313.98, N=12849, *df*=7, α=.05, $p\leq.001^{***}$). Although a significant Kruskal–Wallis test indicates that at least one sample stochastically dominates the other, it identifies neither where this stochastic dominance occurs nor for how many pairs of groups the stochastic dominance is obtained. A Nemenyi test determines which participants are significantly different. Regarding the fixation index per participant, most participants are significantly different from each other in pairs, except P1 *vs.* P3, P1 *vs.* P7, P2 *vs.* P7, P3 *vs.* P7, P4 *vs.* P6 (*e.g.*, $P1$ and $P3$: R=227.38, q=2.47, p=.65). Similarly regarding the duration of the gaze event, except P2 *vs.* P3, P2 *vs.* P4, P2 *vs.* P5, P3 *vs.* P4, P3 *vs.* P5, P4 *vs.* P5, P6 *vs.* P7 (*e.g.*, $P2$ and $P3$: R=231.69, q=2.57, p=.61).

Paper vs. Wireframe. We used the same procedure as above to compare the treatments "paper" and "wireframe" (Fig. 2, index on the left and duration on the right). We found that fixation index and gaze event

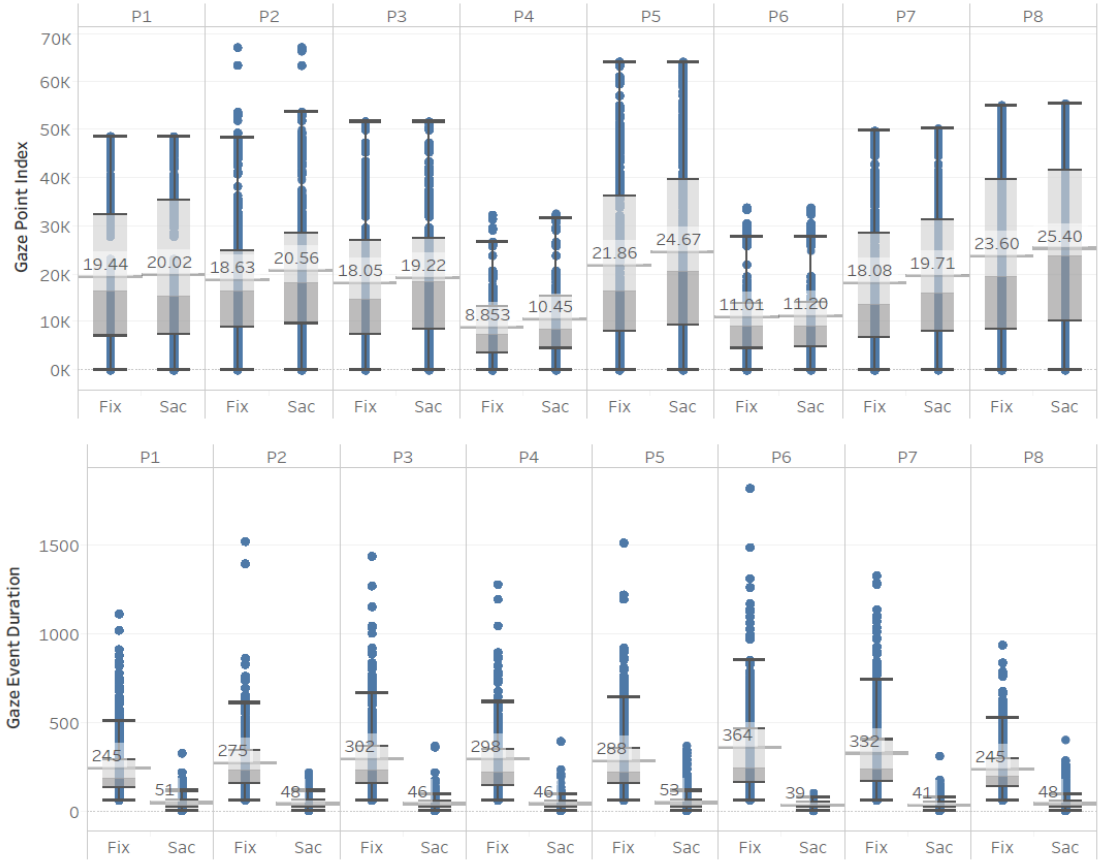

Figure 1: Number of Gaze Point Index (top) and duration of gaze events in milliseconds (bottom) for each Gaze Event Type (fixations *vs.* saccades) broken down by Participant.

duration per treatment are not sampled from a normal distribution. A Kruskal-Wallis H test returned a very highly significant difference between "paper" and "wireframe" in both fixation index (H=322.79, N=12849, df=1, α=.05, $p\leq$.001***) and gaze event duration (H=212.71, N=12849, df=1, α=.05, $p\leq$.001***). The participants used more fixations in the "paper" (Mdn=460) than in the "wireframe" (Mdn=350) and used shorter gaze events in the "paper" (Mdn=158) than in the "wireframe" (Mdn=175).

Basketball vs. Fashion. We used the same procedure as above to compare the themes of "basketball" and "fashion" (Fig. 3, index on the left and duration on the right). We found that fixation index and gaze event duration per treatment are not sampled from a normal distribution. A Kruskal-Wallis H test did not return significant differences between "basketball" and "fashion" neither in fixation index (H=2.67, N=12849, df=1, α=.05, p=.10) nor in gaze event duration (H=0.49, N=12849, df=1, α=.05, p=.48), which was also confirmed by a Nemenyi test. Participants performed the same way on both themes.

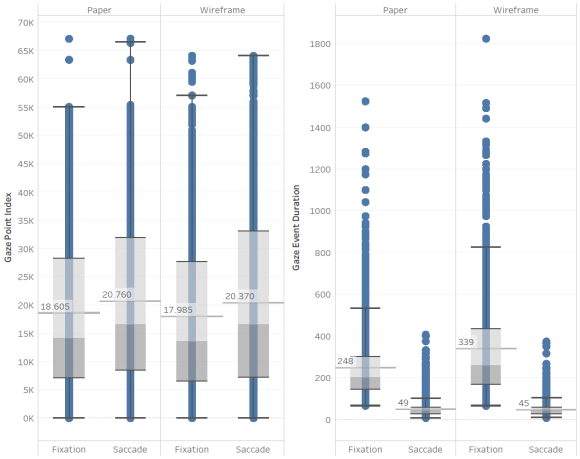

Figure 2: Number of Gaze Point Index (left) and duration of gaze events in milliseconds (right) for each Gaze Event Type (fixations *vs.* saccades) broken down by Treatment.

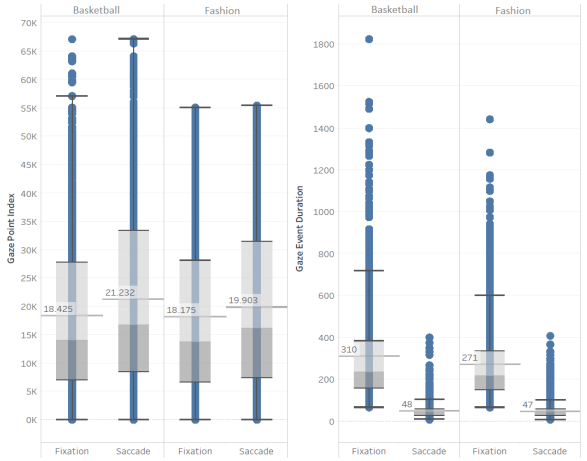

Figure 3: Number of Gaze Point Index (left) and duration of gaze events in milliseconds (right) for each Gaze Event Type (fixations *vs.* saccades) broken down by Theme.

4.2. Questionnaire and Interview

Fig. 4 shows a divergent stacked bar chart showing the participants' responses to the perceived usability questionnaire by treatment ("paper" *vs.* "wireframe"). First, the internal consistency among the participants is more reliable in "paper" (Cronbach's α=0.91, which indicates a very good level of reliability (Cortina, 1993), and Guttman's reliability δ=0.99 with 2,000 iterations) than in "wireframe" (Cronbach's α=0.59, which is a little below the generally acceptable range, and Guttman's reliability =0.97 with 2,000 iterations). Overall, the participants were quite reliable in both treatments (Cronbach's α=0.83 and Guttman's reliability =0.97 also with 2,000 iterations). The answers provided by the participants in "wireframe" (M=3.73, SD=0.84) are significantly different (a Wilcoxon signed rank test for two pairs of samples returns T=858.5, *Z-score*=1.97, p=.027*) from those provided in "paper" (M=3.48, SD=1.082). Overall, the average score is higher with a more contracted standard deviation for the "wireframe", the disagreement is more important in "paper" (16/112=23%) than in the "wireframe" (12/112=11%) and the agreement is stronger for the "wireframe" (82/112=73%), while the neutral value remains comparable (15 *vs.* 18). This suggests that the wireframes were perceived better than the paper version. Second, considering the answers to McCurdy et al. dimensions ($Q10 - Q14$), we notice the following major differences. $Q10$="The prototype is visually refined" is obviously less well estimated in the "paper" treatment than in the "wireframe" counterpart, which is normal since its LoF is lower. The results of $Q13$="The elements captured and represented in the prototype are interactive" and $Q14$="The prototype uses data that are representative of real-world data" are at the expense of "paper" since the fidelity of the interaction is at its lowest level: only a few comments are given to perceive the behavior as opposed to a partial simulation for the "wireframe". This suggests that a wireframe, even with its simplest "Point and click" form, is better appreciated. The remaining two questions (($Q11 - Q12$) do not vary much from one treatment to another, as the corresponding LoF does not influence the answers. For example, $Q11$="The prototype covers a wide range of functionalities" does not vary much between treatments as the behavior is materialized on a static media.

Third, with respect to research-based usability guidelines (Leavitt & Shneiderman, 2006) ($Q1 - Q9$), most of the responses are in favor of the "wireframe", except for $Q8$="The prototype organizes the information clearly", $Q6$="The prototype uses clear link labels" and $Q5$="The prototype uses clear category labels". This might be due to the fact that visual elements, such as interactive components, links, or wireframe labels, do not make the affordance clear enough. This is consistent with the testing pattern observed for the wireframe in which participants point and click, as opposed to the paper prototype in which participants expect these elements to remain static without any visual details distracting them from the contents to analyze. Another explanation might lie in the presence of manuscript annotations on the paper prototype, and the absence of such annotations on the wireframe. The participants $P5$ and $P6$ used back and forth eye movements between these annotations and the components of the mobile application in the "paper" treatment (Fig. 5). Fig. 6 shows the heat maps for the "wireframe" treatment. In the absence of such annotations, participants $P5$ and $P1$ did not perform any such eye movements in the "wireframe" treatment. Overall, all these elements seem to argue in favor of the "wireframe" treatment. However, the analysis of the interviews allows us to put forward three counter-arguments in favour of "paper" treatment:

1. Half the participants (4/8) found the paper prototype easier to analyze as the manuscript annotations allowed them to better understand the purpose of the prototype, 2 in 8 found the analysis equally easy in both treatments, and only 2 in 8 found the wireframe prototype easier to analyze.
2. Half the participants (4/8) found that the paper prototype was more efficient to make future "clickable" components pop out, while most participants (6/8) clicked on every component of the wireframe for not overlooking them.
3. While all participants understood the purpose of the prototype in the "paper" treatment, only half did so in the "wireframe" treatment.

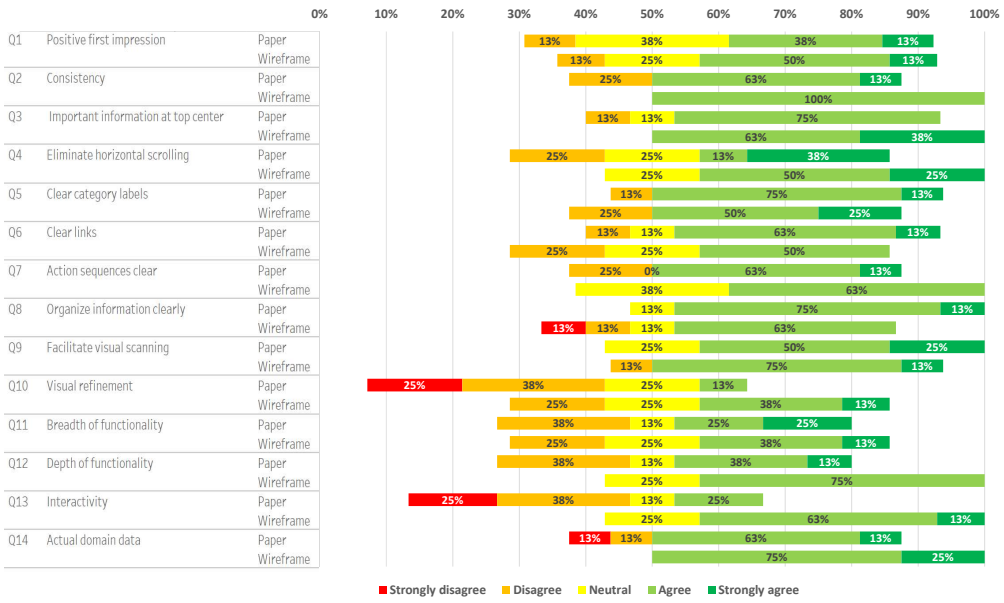

Figure 4: Perceived usability scores for each question of the questionnaire broken down by Treatment.

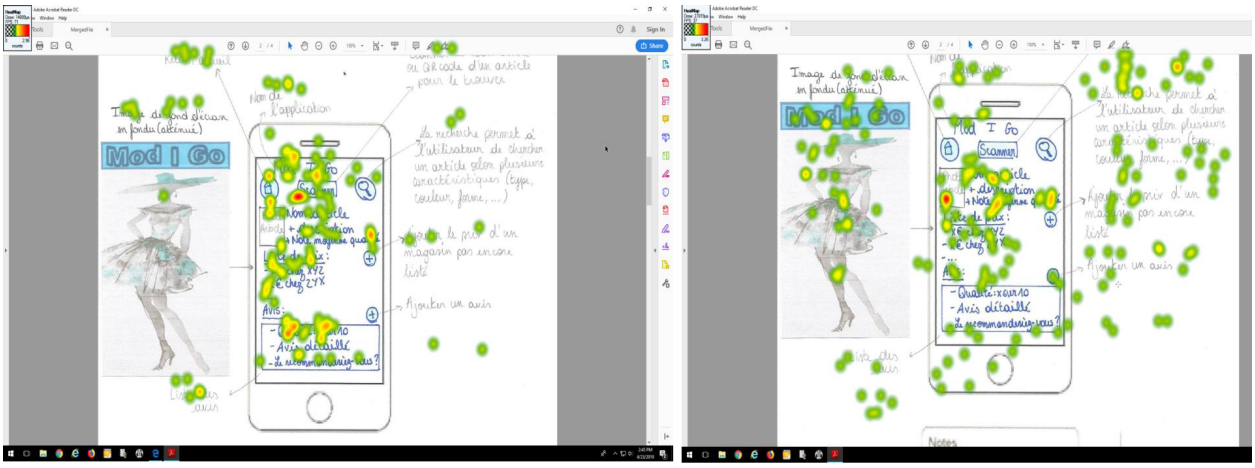

Figure 5: Heat maps in "paper+fashion" GUI prototype for participants P5 (left) and P6 (right).

In turn, the back and forth eye movements observed in the "paper" treatment (Fig. 5) combined with an increased understanding of the purpose of the prototype as highlighted by the interviews align with an increased number of fixations combined with shorter gaze points in the "paper" treatment.

4.3. Threats to Validity

A threat to the statistical conclusion lies in the small sample size (8 participants). However, we compensated for this relative weakness by increasing the duration of the experimental task (1 hour per individual session), which allowed us to collect about 400,000 eye-tracking raw data (391,252 entries in our clean dataset). In addition, although we recruited participants from a homogeneous population (8 women with the same previous experience with GUI design), we observed an important inter-individual variability between participants. However, our analysis of eye movements in GUI prototypes identified the common ground between them: *e.g.* participants oscillated between two main regions (comments *vs.* prototype) in "paper", and were able to keep their locus on the prototype only in " wireframe", thus optimizing their feedback. In addition, we used counterbalancing to control the threats to internal validity such as history, maturation, testing, instrumentation, regression, selection and mortality (Campbell & Stanley, 1963). Finally, external validity concerns whether the situation captured in the experimental setting corresponds to the natural situation toward which researchers wish to generalize (Kieffer, 2017). Accordingly, we cannot generalize our findings

Figure 6: Heat maps in the "wireframe" treatment for participants $P1$ (left) and $P5$ (right).

to experienced designers, as we recruited participants from a sample of novice female designers, or to any type of prototype evaluation, as we asked participants to perform a guideline-based review. To increase the generalizability of our findings, we should repeat this experiment with a sample of experienced designers, including men and women, and with other evaluation methods to confirm that the results hold for a broader population and for a broader scope of evaluations.

5. Conclusion

In this paper, we analyzed the eye movements performed by N=8 participants on two GUI prototypes for two themes in two treatments: paper *vs.* wireframe. We compared the results obtained for these two themes and two treatments to suggest the following findings: the wireframe version provided better overall satisfaction compared to the paper version, attracted more attention to essential elements, and the discussion was more oriented toward high-level design elements as opposed to low-level details. Expression about dynamic aspects is more important than that of a paper prototype without falling into the trap of excessive description. Participants acting on the wireframe were able to keep their locus of attention on the prototype itself and their locus of control on it while interacting, thus maintaining consistency between what they are looking at and what they are commenting on. Due to its interactivity, the wireframe prototype fosters a more exploratory approach to design (*e.g.*, testing various widgets such as push buttons, navigating more extensively and investigating alternative behaviors) than the paper version. Low-fidelity prototypes are important for several reasons: designers typically start by sketching such horizontal low-fidelity GUI prototypes (as opposed to vertical high-fidelity GUI prototypes) that are usually subject to eye tracking studies in the literature), the amount of usability problems detected on low-fidelity prototypes remains the same as for high-fidelity, a low-fidelity prototype can always gracefully evolve towards a prototype with a higher LoF if it starts from a wireframe rather than from a paper prototype. This transition can be supported by existing software that offers multiple levels of fidelity (Coyette et al., 2007; Suleri, Sermuga Pandian, Shishkovets, & Jarke, 2019) and the transition between them, some up to code generation (Pandian, Suleri, & Jarke, 2020).

By investigating McCurdy's dimensions on a paper sketch *vs.* an interactive wireframe, this paper identified that these two treatments actually span more largely on these dimensions and do not represent a clear-cut partition of these dimensions. Paper-and-pencil prototypes are not all necessarily performed in Lo-Fi. For example, a paper sketch can range from Lo-Fi (*e.g.* when no interaction is prototyped) to Hi-Fi (*e.g.* when interactivity is simulated). Similarly, the wireframe performed by a wireframe tool can range from Lo-Fi (*e.g.* when no interaction is prototyped) to Hi-Fi (*e.g.* when interactivity is simulated). To better cover this spanning, future work will replicate the study with treatments covering other overlapping dimensions.

Acknowledgements

The authors acknowledge the support from the Institute for Language and Communication (ILC).

References

Abid, N. J., Maletic, J. I., & Sharif, B. (2019). Using developer eye movements to externalize the mental model used in code summarization tasks. In *Proc. of ETRA '19* (pp. 13:1–13:9). New York: ACM.

Bailey, B. P., & Konstan, J. A. (2003). Are informal tools better?: Comparing demais, pencil and paper, and authorware for early multimedia design. In *Proc. of CHI '03* (pp. 313–320). New York: ACM.

Bednarik, R., & Tukiainen, M. (2006). An eye-tracking methodology for characterizing program comp. processes. In *Proc. of ETRA '06*.

Bojko, A. (2006). Using eye tracking to compare web page designs: A case study. *J. of Usability Studies*.

Busjahn, T., Bednarik, R., & Schulte, C. (2014). What influences dwell time during source code reading?: Analysis of element type and frequency as factors. In *Proc. of ETRA '14* (pp. 335–338). NY: ACM.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.*

Chicago: Rand McNally Company.

Cheng, S., & Dey, A. K. (2019). I see, you design: user interface intelligent design system with eye tracking and interactive genetic algorithm. *CCF Trans. Perv. Comput. Int.*, *1*(3), 224–236.

Cherubini, M., Venolia, G., DeLine, R., & Ko, A. J. (2007). Let's go to the whiteboard: How and why software dev. use drawings. In *Proc. of CHI '07.*

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Coyette, A., Kieffer, S., & Vanderdonckt, J. (2007). Multi-fidelity prototyping of user interfaces. In *Proc. of Interact '07* (pp. 150–164). Berlin.

Cutrell, E., & Guan, Z. (2007). What are you looking for?: An eye-tracking study of information usage in web search. In *Proc. of CHI '07* (pp. 407–416).

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *IJIE*, 631–645.

Hejmady, P., & Narayanan, N. H. (2012). Visual attention patterns during program debugging with an ide. In *Proc. of ETRA '12* (pp. 197–200). NY: ACM.

Holmes, T., & Zanker, J. M. (2012). Using an oculomotor signature as an indicator of aesthetic preference. *i-Perception*, *3*(7), 426-439.

Jacob, R. J., & Karn, K. S. (2003). Commentary on section 4 - eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (p. 573 - 605).

Jacob, R. J. K. (1990). What you look at is what you get: Eye movement-based interaction techniques. In *Proc. of CHI '90* (p. 11–18). New York: ACM.

Kieffer, S. (2017). ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *AIS Trans. on HCI*, *9*(2), 149–172.

Kieffer, S., Rukonić, L., Kervyn de Meerendré, V., & Vanderdonckt, J. (2020). A process reference model for UX. In *Computer vision, imaging and computer graphics theory and applications.*

Leavitt, M. O., & Shneiderman, B. (2006). *Research-based web design and usability guidelines*. U.S. Dept. of Health and Human Serv.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., & Vera, A. (2006). Breaking the fidelity barrier: An examination of our current characterization of prototypes and an example of a mixed-fidelity success. In *Proc. of CHI '06* (pp. 1233–1242).

Pandian, V. P. S., Suleri, S., & Jarke, M. (2020). Syn: Synthetic dataset for training ui element detector from lo-fi sketches. In *Proc. of IUI '20* (p. 79–80).

Petrie, J. N., & Schneider, K. A. (2006). Mixed-fidelity prototyping of uis. In *Proc. of DSVIS '06.*

Plimmer, B., & Apperley, M. D. (2003). Software for students to sketch interface designs. In *Proc. of INTERACT '03* (pp. 73–80). IOS Press.

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity proto. debate. *Inter.*, *3*(1), 76–85.

Sangiorgi, U. B., Beuvens, F., & Vanderdonckt, J. (2012). User interface design by collaborative sketching. In *Proc. of DIS '12* (pp. 378–387). ACM.

Sefelin, R., Tscheligi, M., Tscheligi, M., & Giller, V. (2003). Paper prototyping - what is it good for?: A comparison of paper- and computer-based low-fidelity prototyping. In *Proc. of CHI EA '03.*

Shaffer, T. R., Wise, J. L., Walters, B. M., Müller, S. C., Falcone, M., & Sharif, B. (2015). itrace: Enabling eye tracking on software artifacts within the ide to support software engineering tasks. In *Proc. of ESEC/FSE 2015* (pp. 954–957).

Sharif, B., Falcone, M., & Maletic, J. I. (2012). An eye-tracking study on the role of scan time in finding source code defects. In *Proc. ETRA '12.*

Suleri, S., Sermuga Pandian, V. P., Shishkovets, S., & Jarke, M. (2019). Eve: A sketch-based software prototyping workbench. In *Proc. of CHI EA '09.*

Tullis, T., & Albert, B. (2013). *Measuring the user experience: collecting, analysing, and presenting usability metrics*. Elsevier.

Uebelbacher, A., Sonderegger, A., & Sauer, J. (2013, 01). Effects of Perceived Prototype Fidelity in Usability Testing under Different Conditions of Observer Presence. *Int. with Comp.*, *25*(1), 91-101.

Van den Bergh, J., Sahni, D., Haesen, M., Luyten, K., & Coninx, K. (2011). Grip: Get better results from interactive prototypes. In *Proc. of EICS '11.*

Vidyapu, S., Vedula, V. S., & Bhattacharya, S. (n.d.). Quantitative visual attention prediction on webpage images using multiclass svm. In *Proc. of ETRA '19* (pp. 90:1–90:9). New York: ACM.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. In *Proc. of CHI '96.*

Walker, M., Takayama, L., & Landay, J. A. (2002). High-fidelity or low-fidelity, paper or computer? choosing attributes when testing web prototypes. *Proc. of HFES '02*, *46*(5), 661-665.

Wiklund, M. E., Thurrott, C., & Dumas, J. S. (1992). Does the fidelity of software prototypes affect the perception of usability? *Proc. of HFES '92*, *36*(4).

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Berlin.

Yeung, L., Plimmer, B., Lobb, B., & Elliffe, D. (2008). Effect of fidelity in diagram presentation. In *Proc. of BCS-HCI '08* (pp. 35–44).