

**Louvain School of Management**

# **Evaluating the Visual Design of Graphical User Interfaces on Mobile Devices: A Family of Experiments**

Author(s): Zoé Dubois  
Supervisor(s): Pr. Jean Vanderdonckt (UCLouvain)  
Academic year 2021-2022  
Dissertation for the master of Business Engineering  
Master [120] : Business Engineering  
Daytime schedule



# ACKNOWLEDGEMENTS

I would like to express my gratitude to the people who helped me in the realisation of this thesis and made it possible.

First of all, I sincerely thank Professor Jean Vanderdonckt who provided me with knowledge, guidance, suggestions and encouragement throughout my thesis. I would like to thank him for the precious time he gave me to answer my questions and to share his experience, advice and perspective.

Secondly, I would like to acknowledge Nicolas Burny for his help and availability, especially concerning the experimental part. His contribution to the deployment and completion of the experiments was essential. Furthermore, I am grateful to have been able to discuss the results of my experiments and their interpretation with him.

Finally, I would like to thank my family and friends for their continuous support, availability and advice throughout the realisation of this thesis.



# ABSTRACT

The growing expansion of e-commerce practices has challenged traditional companies in the retail industry. In order to increase or capture new market parts, companies have been encouraged to strengthen their online presence. Therefore, the visual platforms for user interaction, called Graphical User Interfaces (GUIs) benefit from a particular attention from companies. Indeed, the aesthetics and the visual complexity of these interfaces significantly impact the users and represent a mean to establish a competitive advantage. Our research focuses on the measures of visual design on mobile devices, as this kind of device has become increasingly popular. Moreover, few discussions exist in the literature regarding this subject and stay poorly developed. On the one hand, this thesis aims to cover the gap that exists in the literature. On the other hand, it aims to find measures in order to assess the aesthetics and visual complexity of mobile devices through a Systematic Literature Review (SLR) and a family of experiments. We found that the most promising model of prediction captures 68% of the beauty ratings variability through eight measures presented by Aalto Interface Metrics (AIM). Furthermore, no significant aesthetic difference was observed among the popular and less popular websites. We also observed that four variables possess a higher mean among the e-commerce websites category compared to the other categories.

# Contents

|          |                                                                                 |           |
|----------|---------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                             | <b>1</b>  |
| 1.1      | Context of the problem . . . . .                                                | 1         |
| 1.2      | Objective of the thesis . . . . .                                               | 4         |
| 1.2.1    | Mission statement . . . . .                                                     | 4         |
| 1.2.2    | Research question . . . . .                                                     | 4         |
| 1.2.3    | Working hypotheses . . . . .                                                    | 6         |
| 1.2.4    | Overall methodology . . . . .                                                   | 7         |
| 1.3      | Outline of the thesis . . . . .                                                 | 8         |
| <b>2</b> | <b>Related work</b>                                                             | <b>9</b>  |
| 2.1      | Graphical User Interfaces . . . . .                                             | 9         |
| 2.1.1    | Mobile User Interfaces . . . . .                                                | 9         |
| 2.2      | Aesthetics . . . . .                                                            | 10        |
| 2.2.1    | Concept evolution and definitions . . . . .                                     | 10        |
| 2.2.2    | Importance of aesthetics . . . . .                                              | 10        |
| 2.2.3    | Measuring aesthetics . . . . .                                                  | 11        |
| 2.2.4    | Visual complexity . . . . .                                                     | 12        |
| <b>3</b> | <b>Systematic Literature Review</b>                                             | <b>13</b> |
| 3.1      | Method . . . . .                                                                | 13        |
| 3.2      | Results and discussion . . . . .                                                | 16        |
| 3.2.1    | Query 1 . . . . .                                                               | 16        |
| 3.2.1.1  | Measuring the visual impressions and the complexity of<br>mobile apps . . . . . | 16        |
| 3.2.1.2  | Enumerating several complexity factors . . . . .                                | 19        |
| 3.2.1.3  | Assessing the visual design of Android apps . . . . .                           | 19        |
| 3.2.1.4  | Measuring the simplicity of smartphones . . . . .                               | 20        |
| 3.2.2    | Query 2 . . . . .                                                               | 20        |
| 3.2.3    | Query 3 . . . . .                                                               | 20        |
| 3.2.4    | Query 4 . . . . .                                                               | 21        |

|          |                                                                              |           |
|----------|------------------------------------------------------------------------------|-----------|
| 3.2.4.1  | Detecting the aesthetic defects of mobile interfaces with measures . . . . . | 21        |
| 3.2.5    | Query 5 . . . . .                                                            | 25        |
| 3.2.5.1  | Measuring the complexity of mobile applications . . . . .                    | 25        |
| 3.3      | Conclusion of the Systematic Literature Review . . . . .                     | 28        |
| <b>4</b> | <b>Experiment 1 - SimilarWeb / QUESTIM</b>                                   | <b>29</b> |
| 4.1      | Experimental method . . . . .                                                | 29        |
| 4.2      | Results and discussion . . . . .                                             | 32        |
| <b>5</b> | <b>Experiment 2 - Alexa ranking / QUESTIM</b>                                | <b>36</b> |
| 5.1      | Experimental method . . . . .                                                | 37        |
| 5.2      | Results and discussion . . . . .                                             | 38        |
| 5.2.1    | Descriptive statistics . . . . .                                             | 38        |
| 5.2.2    | Inferential statistics . . . . .                                             | 42        |
| <b>6</b> | <b>Experiment 3 - SimilarWeb / AIM / Ergonomic criteria</b>                  | <b>46</b> |
| 6.1      | Experimental method . . . . .                                                | 47        |
| 6.2      | Results and discussion . . . . .                                             | 51        |
| 6.2.1    | Descriptive statistics . . . . .                                             | 51        |
| 6.2.2    | Inferential statistics . . . . .                                             | 54        |
| <b>7</b> | <b>Conclusion and perspectives</b>                                           | <b>60</b> |
| 7.1      | Summary of the contributions . . . . .                                       | 60        |
| 7.2      | Benefits and limits . . . . .                                                | 62        |
| 7.3      | Future work . . . . .                                                        | 63        |
| 7.3.1    | Short term . . . . .                                                         | 63        |
| 7.3.2    | Midterm . . . . .                                                            | 64        |
|          | <b>Appendices</b>                                                            | <b>71</b> |
| <b>A</b> | <b>Aesthetics measures: additional formulas</b>                              | <b>72</b> |
| A.1      | Regularity . . . . .                                                         | 72        |
| A.2      | Homogeneity . . . . .                                                        | 73        |
| A.3      | Balance . . . . .                                                            | 73        |
| A.4      | Cohesion . . . . .                                                           | 74        |
| A.5      | Unity . . . . .                                                              | 74        |
| A.6      | Element smallness . . . . .                                                  | 75        |
| A.7      | Misalignment . . . . .                                                       | 76        |
| A.8      | Imbalance . . . . .                                                          | 76        |
| A.9      | Colour complexity . . . . .                                                  | 77        |

|                                                                                                                                         |           |
|-----------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>B Experiment 1</b>                                                                                                                   | <b>78</b> |
| B.1 Shapiro-Wilk normality tests of the QUESTIM measures of SimilarWeb . . .                                                            | 78        |
| <b>C Experiment 2</b>                                                                                                                   | <b>79</b> |
| C.1 Box plot of the average beauty rating . . . . .                                                                                     | 79        |
| C.2 Descriptive statistics of the average beauty rating . . . . .                                                                       | 80        |
| C.3 Shapiro-Wilk normality tests of the QUESTIM measures of Alexa . . . . .                                                             | 80        |
| C.4 Principal Component Analysis with the QUESTIM measures . . . . .                                                                    | 81        |
| C.5 Linear regression of the average beauty rating as a function of all QUESTIM<br>measures . . . . .                                   | 82        |
| C.6 VIF - linear regression of the average beauty rating as a function of all the<br>QUESTIM measures . . . . .                         | 83        |
| C.7 VIF - linear regression of the average beauty rating as a function of 14<br>unfiltered QUESTIM measures . . . . .                   | 83        |
| C.8 Stepwise regression using 14 QUESTIM measures to predict the average<br>beauty rating . . . . .                                     | 84        |
| C.9 Final linear regression of the average beauty rating using the QUESTIM<br>measures . . . . .                                        | 85        |
| C.10 Analysis of the variables among the first half compared to the second half<br>of the Alexa ranking sites . . . . .                 | 86        |
| <b>D Experiment 3</b>                                                                                                                   | <b>88</b> |
| D.1 Box plot of the average beauty and complexity rating . . . . .                                                                      | 88        |
| D.2 Descriptive statistics of the average beauty and complexity rating . . . . .                                                        | 89        |
| D.3 Shapiro-Wilk normality tests of the ergonomic criteria . . . . .                                                                    | 89        |
| D.4 Principal Component Analysis with the ergonomic criteria . . . . .                                                                  | 90        |
| D.5 Descriptive statistics of the AIM measures . . . . .                                                                                | 91        |
| D.6 Shapiro-Wilk normality tests of the AIM measures . . . . .                                                                          | 94        |
| D.7 Linear regression of the average beauty rating as a function of all AIM<br>measures . . . . .                                       | 97        |
| D.8 VIF - linear regression of the average beauty rating as a function of all AIM<br>measures . . . . .                                 | 98        |
| D.9 Remaining AIM measures after having filtered the highly correlated vari-<br>ables (average beauty rating prediction) . . . . .      | 99        |
| D.10 Stepwise regression using the 27 AIM measures to predict the average<br>beauty rating . . . . .                                    | 100       |
| D.11 Final linear regression of the average beauty rating using the AIM measures                                                        | 101       |
| D.12 Remaining AIM measures after having filtered the highly correlated vari-<br>ables (average complexity rating prediction) . . . . . | 102       |

|                                                                                                                            |     |
|----------------------------------------------------------------------------------------------------------------------------|-----|
| D.13 Stepwise regression using the 25 AIM measures to predict the average complexity rating . . . . .                      | 103 |
| D.14 Final linear regression of the complexity average rating using the AIM measures . . . . .                             | 104 |
| D.15 VIF - linear regression of the average beauty rating as a function of the ergonomic criteria . . . . .                | 104 |
| D.16 Stepwise regression using the ergonomic criteria to predict the average beauty rating . . . . .                       | 105 |
| D.17 VIF - linear regression of the average complexity rating as a function of the ergonomic criteria . . . . .            | 105 |
| D.18 Stepwise regression using the ergonomic criteria to predict the average complexity rating . . . . .                   | 106 |
| D.19 Analysis of the variables among the e-commerce and shopping category compared to the rest of the categories . . . . . | 106 |
| D.20 Means of some variables among the e-commerce and shopping category compared to the rest of the categories . . . . .   | 108 |

# List of Figures

|     |                                                                         |    |
|-----|-------------------------------------------------------------------------|----|
| 4.1 | Zone segmentation on QUESTIM . . . . .                                  | 30 |
| 4.2 | Box plot of the QUESTIM measures for SimilarWeb websites . . . . .      | 33 |
| 4.3 | Pair plot of the QUESTIM measures for SimilarWeb websites . . . . .     | 35 |
| 5.1 | Visual summary of the inferential statistics (experiment 2) . . . . .   | 38 |
| 5.2 | Box plot of the QUESTIM measures for Alexa websites . . . . .           | 39 |
| 5.3 | Pair plot of the QUESTIM measures for Alexa websites . . . . .          | 41 |
| 5.4 | Scatter plot of the average beauty rating per popularity rank . . . . . | 45 |
| 6.1 | Visual summary of the inferential statistics (experiment 3) . . . . .   | 50 |
| 6.2 | Box plot of the ergonomic criteria . . . . .                            | 52 |
| 6.3 | Pair plot of the ergonomic criteria . . . . .                           | 53 |

# List of Tables

|     |                                                                                                     |    |
|-----|-----------------------------------------------------------------------------------------------------|----|
| 3.1 | Searching strings and options of the libraries . . . . .                                            | 15 |
| 4.1 | Descriptive statistics of the QUESTIM measures . . . . .                                            | 34 |
| 5.1 | Descriptive statistics of the QUESTIM measures . . . . .                                            | 40 |
| 5.2 | Correlations of the QUESTIM measures with the average beauty rating . .                             | 42 |
| 6.1 | Descriptive statistics of the ergonomic criteria . . . . .                                          | 51 |
| 6.2 | Correlations of the ergonomic criteria with the average beauty and com-<br>plexity rating . . . . . | 53 |

# Chapter 1

## Introduction

### 1.1 Context of the problem

Over the years, Internet has become a widely used system in our daily lives. The number of active Internet users is growing and was estimated at more than 4.5 billion at the beginning of 2021. It represents more than half the world population. Among these numbers, more than 95% of the Internet users are also mobile Internet users (Johnson, 2021a). Furthermore, mobile devices are increasingly used to access the Internet and represent more than 50% of the Internet traffic (Clement, 2021). They also tend to be preferred to software applications given the fact that mobile applications (apps) are becoming more accessible, versatile, user-friendly and sophisticated (Bessghaier et al., 2021). This growing importance of mobiles is expected to progress during the coming years as they are more and more available and affordable (Ceci, 2021).

As the Internet and mobile phones have become increasingly popular, a large number of websites and applications are evolving in a very competitive market. Indeed, on the World Wide Web, users have the opportunity to quickly switch from one website to another without great effort (Robins and Holmes, 2008). The app market, which also faces a lot of competition, is saturated. Moreover, users are choosing the apps in a more and more selective way. In order to cope with this growing competition, companies tend to focus on visual details design since pleasing interfaces can provide a significant competitive advantage (Riegler and Holzmann, 2018).

In addition, the retail industry has encountered major changes over the last decades. As previously mentioned, the growing importance of the Internet has enabled the development of e-commerce. Electronic commerce is defined as the transactions performed using electronic means (Niranjanamurthy et al., 2013). Furthermore, this developing trend was reinforced by the on-going situation of COVID-19 that led to a significant increase of

virtual practices. As a matter of fact, in 2020, two billion buyers purchased a product or a service on the Internet (Coppola, 2021; Lund et al., 2021).

What is more, e-commerce has become a disruption that changed the way the goods are exchanged. Indeed, from the customer’s perspective, it has allowed to facilitate the price comparison between competitors, to give a simple and instantaneous access to a large possibility of choices, etc. Regarding the sellers, this has enabled them to ease the customization of content and has brought new possibilities to capture market intelligence by collecting data, among other things (Niranjanamurthy et al., 2013). This new expanding trend challenges the traditional retail industry and forces companies to carefully review their strategy and reinforce their online presence.

In conjunction with the increase in mobile usage and e-commerce practices, mobile transactions such as m-commerce are also rising in popularity (Johnson, 2021b). According to Niranjanamurthy et al (2013), m-commerce is defined as the operation of financial and promotional activities thanks to wireless devices.

Given the growing importance of e-commerce and m-commerce, it has become essential for companies to pay attention to the online interaction that they can potentially have with their current or future clients. Besides, the “space of interaction between users and an information system” (Zen, 2017, p.23) is named a user interface. The aesthetics of this interface appears to be one of the most essential features. Indeed, it is one of the most noticeable features of the interfaces at first glance and has a significant impact on the user experience (Reinecke et al., 2013). Companies are increasingly aware of this variable and adapt their interfaces accordingly.

As described by Lavier and Tractinsky (2004), the visual appeal <sup>1</sup> has an influence on enjoyment and is positively correlated with users’ satisfaction (Choi and Lee, 2012). Interface aesthetics impact users’ credibility, loyalty (Bessghaier et al., 2021) acceptability, comprehensibility, productivity (D. C. L. Ngo et al., 2003) and their trust perception (Robins and Holmes, 2008). The aesthetics is also linked to the post-usage perception of interface usability. Systems perceived as more usable are more likely to be bought or used (Tractinsky et al., 2000). It is therefore not surprising that this aspect has become a major issue for companies working in the e-commerce sector.

Besides, the ISO 25010 standard aims at introducing a scope to assess the quality of a system. Amongst other things, it is characterized by its usability that is composed of the user interface aesthetics, which is itself defined as the capability of the interface to provide pleasing and satisfying interaction (ISO 25000, 2021).

---

<sup>1</sup>From now on, the “aesthetics”, “visual appeal” and “beauty” terms will be used as synonyms since they are treated as such in some studies and are all related to the visual attractiveness of a user interface (Lindgaard et al., 2006; Tuch et al., 2012)

In the literature, studies used to mainly focus on the evaluation of the usability of interfaces (D. C. L. Ngo et al., 2000) while neglecting the aesthetics aspect (Lavie and Tractinsky, 2004). According to Ngo et al. (2001), aesthetics theories were considered as primitive in the early 21<sup>st</sup> century. Nevertheless, the literature has tried to find, through the years, a trade-off between discussions on usability and visual appeal. Scientific community started to pay more attention to the role of aesthetics and its effect on users (Lavie and Tractinsky, 2004).

The aesthetics remains a rather complex concept and it is important to explore its underlying notions (Miniukovich and De Angeli, 2014a). While visual appeal on desktop interfaces starts to grow and attract attention (Lavie and Tractinsky, 2004), very few initiatives and proposals relate to the visual quality assessment of Mobile User Interfaces (MUIs). In the literature, no general agreement regarding the mobile interfaces evaluation has been settled and only few assessment tools have been elaborated (Soui et al., 2017). We also noticed a lack of measures to evaluate the complexity of MUIs without user participation (Riegler and Holzmann, 2018). The originality of this thesis lies in the fact that we will study the aesthetics measures of websites on mobile devices while the subject remains little discussed in the literature.

To cope with the limited number of measures available for mobile devices, it is possible to use pre-existing measures or to create new ones. In this way, the usage of the measures created for computers can be extended to mobiles by leaving them as it is or by making some adjustments to them. The validation step, in this new application context, remains essential as desktop design methods cannot directly be implemented into mobile devices (Riegler and Holzmann, 2018).

It is necessary to be aware that mobiles present challenges that are different from those encountered with computers. Therefore, aesthetics evaluation measures and methods developed for desktops must be treated with precaution when implemented for mobile devices. First of all, smartphones have a limited screen size. This means that available space must be optimized to provide an ease of use as well as an aesthetically pleasing interface. Then, smartphones are used with fingers which replace the mouse and keyboard of a computer. It sometimes leads to small adaptations of the displayed elements. Thirdly, mobile phones are used in a different context than computers. Indeed, smartphones can be used on motion and, as a matter of fact, receive less attention. Stimuli from the external environment in which the user is moving can disrupt the smartphone usage and prevent the user from fully focusing on the screen. For instance, a user in the street may be disturbed by sounds and visual elements such as changing light conditions. Consequently, users are expected to be able to perform tasks efficiently and with a minimum of cognitive load. Finally, the use of smartphones is limited in time due to the battery capacity (Chittaro, 2011; Choi and Lee, 2012).

In addition to validating mobile measures, it is essential to provide an efficient and automatic means of assessing the interfaces aesthetics to designers (Riegler and Holzmann, 2018). Indeed, evaluating the visual appeal of user interfaces requires time and incurs expenses (Robins and Holmes, 2008). More especially, methods that involve users to assess certain aspects related to an interface are time-consuming and expensive. Human intervention in the evaluation process also makes it more error prone (Soui et al., 2017). Moreover, automatic computation eases the comparison between various interfaces (Riegler and Holzmann, 2018). The challenges related to the automation of aesthetic measures are even more important for start-ups, smaller companies and freelancers. Unlike large companies, they have budget constraints and limited resources. They cannot outsource the websites or applications design or afford extensive training sessions (Miniukovich and De Angeli, 2015). For all of these reasons, it is necessary to assess the interface aesthetics using automatic measures which are, by definition, computed based on the features of interfaces (Miniukovich and De Angeli, 2014b).

In view of the growing importance of mobiles in our daily lives and the influence of aesthetics on perceptions, impressions and users' behaviors, it is essential for companies to focus on the visual design of their web pages or applications. Therefore, designers need an efficient and automatic means to assess the visual quality of their interfaces without spending many resources. Nevertheless, the literature regarding the evaluation of Graphical User Interfaces (GUIs) on mobile devices is not widespread and further investigations are required.

## 1.2 Objective of the thesis

### 1.2.1 Mission statement

The main mission of this thesis is to gather the visual design measures that assess the visual aesthetics of Graphical User Interfaces on mobile devices and to confront them with human judgement. Indeed, the aesthetics of interfaces is a complex subject and it is interesting to understand the factors that make it up.

### 1.2.2 Research question

The main objective of our research (see Section 1.2.1) is translated into the following research question:

**Which are the measures to assess the perceived aesthetics of a website in general and of electronic commerce websites in particular, on mobile devices?**

As it will be explained below (see Section 1.3), we will study this question thanks to two methods. We will explore the related literature through a Systematic Literature Review. On the other side, we will conduct a family of experiments to extend the measures usage from desktops to mobile devices and to develop models predicting user ratings. Indeed, we will confront the aesthetics measures with the human judgement regarding the visual appeal of the interfaces.

While the first experiment will explore the aesthetic measures, offered by the Quality Estimator using Measures (QUESTIM) application, the second experiment will look at the predictive power of these measures, with respect to the aesthetics perceived by the users (research question 1.1). Secondly, we will investigate the link between aesthetics and popularity through research question 1.2. In the literature, the impact of aesthetics on usability has been clearly established (Tractinsky et al., 2000). Instead of conducting an additional experiment on the subject, we decided to look at aesthetics from the perspective of the website popularity.

- Research Question 1.1: Do the QUESTIM measures satisfactorily predict the perceived aesthetics of electronic commerce websites on mobile devices?
  - Hypothesis 1.1: All QUESTIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.
- Research Question 1.2: Are popular electronic commerce websites perceived as aesthetic on mobile devices?
  - Hypothesis 1.2.1: The popular electronic commerce websites are perceived by users as aesthetic on mobile devices.
  - Hypothesis 1.2.2: The popular electronic commerce websites are aesthetic (under QUESTIM measures) on mobile devices.

In the third experiment, we will confront the aesthetics measures with the human judgement thanks to the research question 1.3. We will collect the users aesthetics and complexity ratings for mobile interfaces. Indeed, we will explore the visual complexity concept as this high level notion is rather complex and essential. This is one of the most noticeable features at first glance. This also impacts the user’s initial impression regarding the interface (Reinecke et al., 2013). Moreover, complexity is linked with the aesthetics concept and seems to influence the perceived visual appeal (Miniukovich and De Angeli, 2014b). Secondly, in the research question 1.4, we will analyse the aesthetics and the complexity through the different categories of websites. It was shown, in the literature, that page design metrics differ according to the website category (Ivory et al., 2001).

- Research Question 1.3: Do the Aalto Interface Metrics (AIM) measures and the ergonomic criteria satisfactorily predict the perceived aesthetics and complexity of

electronic commerce websites on mobile devices?

- Hypothesis 1.3.1: All AIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.
  - Hypothesis 1.3.2: All AIM measures predict the perceived complexity of electronic commerce websites on mobile devices.
  - Hypothesis 1.3.3: All ergonomic criteria predict the perceived aesthetics of electronic commerce websites on mobile devices.
  - Hypothesis 1.3.4: All ergonomic criteria predict the perceived complexity of electronic commerce websites on mobile devices.
- Research Question 1.4: Compared to other sites, how are the complexity, the aesthetics and the ergonomics of the electronic commerce websites perceived and evaluated on mobile devices?
    - Hypothesis 1.4.1: Electronic commerce websites are perceived as aesthetic by users compared with the other websites on mobile devices.
    - Hypothesis 1.4.2: Electronic commerce websites are not perceived as complex by users compared with the other websites on mobile devices.
    - Hypothesis 1.4.3: The electronic commerce websites are more aesthetic (under AIM) than other websites on mobile devices.
    - Hypothesis 1.4.4: The electronic commerce websites are more ergonomic (under ergonomic criteria) than other websites on mobile devices.

### 1.2.3 Working hypotheses

Our thesis is part of the research on the visual design of Graphic User Interfaces. The scope of our research was narrowed through successive choices.

Firstly, we will focus on the aesthetics evaluation of mobile devices interfaces given the fact that the subject has not received much attention in the literature. Moreover, this type of device is recent and increasingly popular. Specifically, the visual appeal evaluation will be deployed on website screenshots while not interacting dynamically with the system. We also aim to focus on automatic aesthetics measures as much as possible.

Secondly, we will especially target websites, a particular type of Graphical User Interfaces in our family of experiments. The Systematic Literature Review, which aims at identifying and collecting aesthetics measures on mobile devices, will more generally focus on mobile interfaces as the literature is already primitive on the subject.

Then, on our experiments will focus on e-commerce websites. Indeed, design measures might vary among website categories. Furthermore, rating predictions accuracy might increase if categories are defined (Ivory et al., 2001). We chose the e-commerce category given that this activity is expanding and benefits from a particular interest from many companies.

To evaluate the aesthetics of mobile websites based on homepage screenshots, we will use different sets of measures and criteria. Firstly, 16 aesthetics measures will be semi-automatically computed on Quality Estimator using Measures<sup>2</sup> (QUESTIM), an online application. Aesthetics evaluation will also be performed thanks to Alto Interface Metrics<sup>3</sup> (AIM) which automatically evaluates the design of Graphical User Interfaces through a large set of measures (Oulasvirta et al., 2018). Moreover, eight ergonomic criteria will be computed thanks to human evaluation. These non-automatic measures were, nevertheless, subject to validation (Scapin et al., 2001). Furthermore, they will be assessed by students of the human-computer interface course delivered by Jean Vanderdonckt (Vanderdonckt, 2021) after having received detailed explanations on the subject. It should be emphasized that these criteria do not entirely focus on the visual aspects of an interface as they cover different features of the ergonomics. The ergonomics consists in diverse branches such Human-Computer Interaction, User Experience Design and Visual Design (Zen, 2017). However, it is interesting to explore the impact of these ergonomic criteria on the users' judgement, given that we are concerned by the underlying factor of aesthetics and complexity. For instance, consistency is related to the information identification ease. Evaluating the impact of this criterion on the visual judgement is relevant.

### 1.2.4 Overall methodology

The research process is composed of three different steps that can be iterative. First, the research purpose is determined and some research questions are linked to it. The implementation phase includes establishing a research design and collecting data. Then, analysis and evaluation steps aim to analyse and interpret the obtained results (Gotteland et al., 2012).

The critical literature review, regarding a research question, examines the state of knowledge in the literature and provides a critical analysis. We will adopt a systematic review methodology. This method aims to locate, analyse and summarise the studies related to a research question by using a pre-defined strategy (Saunders et al., 2009).

In addition, we will deploy a quantitative methodology through a family of experiments. Indeed, in order to answer our research question (see Section 1.2.2), we will

---

<sup>2</sup>See <http://questimapp.appspot.com/>

<sup>3</sup>See <https://interfacemetrics.aalto.fi/>

quantitatively test hypotheses (Gotteland et al., 2012, Jacquemin, 2017). For the sake of clarity, the study design of the experiments will be detailed in the chapter dedicated to each experiment (see Sections 4.1, 5.1, 6.1). Thus, the type of data involved will be explained, as well as how it will be collected and analysed, etc. (Jacquemin, 2017). In common with the three experiments conducted, we can attest that the data are primary and quantitative. Indeed, primary data, unlike secondary data, are not pre-existing and need to be collected (Gotteland et al., 2012).

### 1.3 Outline of the thesis

This master thesis is organised into different chapters. The first chapter is dedicated to the introduction of our research. The organisational, physical and statistical context in which our research is implemented is explained. Therefore, we present the reasons and motivations for addressing the subject. Then, the main objective of this thesis is introduced through the mission statement and the research questions. The hypotheses, which consist of the choices made, the constraints and the scope of the research, are presented before detailing the methodology.

The second chapter aims at defining the main terms of our research by exploring the pre-existing work also known as the state of the art.

The third chapter addresses the research question by collecting pre-existing studies that refer to it. Using the Systematic Literature Review method, we aggregate and assess the publications that present aesthetic evaluation measures for mobile interfaces.

A family of experiments is carried out to answer the research question and sub-questions. A chapter is dedicated to each of the three experiments. The first experiment focuses on the aesthetic measures offered by QUESTIM for the most popular sites of the SimilarWeb websites ranking. The second experiment evaluates the QUESTIM measures for the popular sites of the Alexa websites ranking. It also presents users' ratings of beauty. Finally, the third experiment introduces AIM aesthetic metrics, ergonomic criteria, and complexity and aesthetics user ratings for the most popular sites of SimilarWeb. Each chapter contains the deployed methodology of the experiment. The results of a descriptive statistical analysis are also presented and discussed for each of the three experiments. Finally, an inferential analysis is carried out to test the various hypotheses of the second and third experiments.

In the last chapter dedicated to the conclusion, we present our contributions. In addition, we explain the benefits and the limits of our analysis, before detailing future research that could be carried out.

# Chapter 2

## Related work

In this chapter, we will define the principal concepts of our research question which addresses the aesthetics evaluation of Graphical User Interfaces (see Section 1.2.2). Our review of the literature is structured along two lines. The first regards the Graphical User Interface while the second concerns the concept of aesthetics.

### 2.1 Graphical User Interfaces

The Graphical User Interface (GUI) can be seen as “a visual operating display” (Jansen, 1998, p.22). Specifically, the GUI of a device is the visual platform that allows the user to interact with the system while favouring a visual logic. It consists in graphic items such as menus, navigation bars and buttons (Zen and Vanderdonckt, 2014; Zen, 2017).

Historically, the first application with a Graphical User Interface, called Xerox Star, was created in the late seventies. Despite its lack of success, the application has influenced the design of the Apple Lisa as some designers have worked on both projects. Through history, some inventions have impacted the development of GUIs design more than others. The Macintosh, for example, was the first to develop icons, menus and mouse clicking and pointing processes. The Macintosh interface also allowed, for the first time, the user to be guided through the possible choices in a given context (Jansen, 1998).

#### 2.1.1 Mobile User Interfaces

In this thesis, we will study a specific type of Graphical User Interface, the Mobile User Interface (MUI). A Mobile User Interface is defined, by Soui et al. (2020), as “the graphical part of the mobile device that allows the user to interact with the device’s apps, features, content and functions.” (Soui et al., 2020, p.7685). Therefore, MUIs refer to mobile devices such as mobile phones and handheld computers (Niranjanamurthy et al., 2013).

## 2.2 Aesthetics

### 2.2.1 Concept evolution and definitions

“Aesthetics” was introduced three centuries ago and comes from a Greek word which is related to senses perceptions (Lavie and Tractinsky, 2004; Online Etymology Dictionary, 2021). The word “aesthetics” was adopted in the general language at the end of the 18<sup>th</sup> century (Lavie and Tractinsky, 2004). Over time, the term has evolved and has been examined in various disciplines such as philosophy and engineering (Zen, 2017). Through this thesis, we focused on the notion of aesthetics as a reference to what has a pleasant appearance (Lavie and Tractinsky, 2004). Indeed, in the literature of human-computer interaction, aesthetics relates to beauty or pleasing appearance (Zen, 2017). According to Bessghaier et al. (2021), the aesthetics is “the level of the visual appeal of a design.” (Bessghaier et al., 2021, p.2).

The concept of beauty, which invokes aesthetics, is not recent. In the antiquity, aesthetics considerations were seen as essential in architecture. Later, beauty also referred to the human body and nature. However, the concept has remained absent in the human-computer interaction field for a long period of time. Performance, functionality and usability were more discussed than aesthetics. Over the year, discussions and studies regarding usability and aesthetics have become more balanced. Literature started to focus on the effects of aesthetics on users and recognised its predicting power over user impressions and preferences (Lavie and Tractinsky, 2004).

### 2.2.2 Importance of aesthetics

Aesthetics is an essential interface feature as it has numerous impacts on the user impressions, preferences and behaviours. First, the visual appeal impacts enjoyment (Lavie and Tractinsky, 2004) and is positively correlated with users’ satisfaction (Choi and Lee, 2012). Moreover, pleasing GUIs help to engage users whereas unappealing websites are more likely to have their users switch to a competing site. Indeed, the aesthetics of a website determines the engaging of users and has an effect on their loyalty (Bessghaier et al., 2021; Reinecke et al., 2013). Furthermore, an interface with superior aesthetic qualities benefits from a greater credibility, which is the feeling of trust the user might feel facing the informational content. On the contrary, an interface with low aesthetic qualities induces negative perception of credibility at first sight (Robins and Holmes, 2008). The GUI visual appeal can also influence the user acceptability, comprehensibility and productivity (D. C. L. Ngo et al., 2003).

In addition, the strong link between both perceived usability and perceived aesthetics before, as well as after performing tasks, has been shown in the literature. Specifically,

the extent to which a system is aesthetically pleasing impacts the aesthetics and usability perceptions after use. This means that appealing interfaces tend to be seen by users as usable systems. Besides, the perception of usability after usage is linked to the aesthetics and not to the actual usability (Tractinsky et al., 2000).

Visual appeal is an aspect that should not be neglected in interface design since human judgement about aesthetics is formed quickly and is time-persistent (Tractinsky et al., 2006). When the user is confronted with the interface, the first impression is established in the first 50 milliseconds (Lindgaard et al., 2006). Moreover, it has been shown that this judgement persists over time. Indeed, the attractiveness ratings of a web page after 500 ms and after 10s are strongly linked (Tractinsky et al., 2006). Therefore, it is essential to focus on GUIs visual appeal as, in a fraction of a second, users form an opinion about the website that persists over time.

### 2.2.3 Measuring aesthetics

Aesthetics is a complex concept (Miniukovich and De Angeli, 2014a) which needs to be understood comprehensively (Zen and Vanderdonckt, 2014). It is therefore interesting to find measures to evaluate the aesthetics and explore its underlying aspects.

Two different approaches aim to evaluate interface aesthetics. The first one, defined as qualitative and subjective, collects the assessment of experts or users (Bessghaier et al., 2021). Subjective approach may appear as not reliable. For instance, the notion of what is aesthetically pleasing or not may seem completely individual. Nevertheless, by collecting a number of judgements, it is possible to achieve a certain degree of objectivity or to identify some common and accepted elements of evaluation among the participants (Zen and Vanderdonckt, 2016). The second method is characterized as quantitative and objective. It computes measures based on the interface features. Focusing on quantitative evaluation and, in particular, on automatically calculated measures of aesthetics appears to require less time and may lead to less errors (Bessghaier et al., 2021).

That being said, subjective approaches might be carried out to confirm objective ones. Indeed, automatically computed metrics need to align with human judgement. In order to do so, two techniques are possible. The first one consists in training an artificial intelligence in order to deduce measure values based on aesthetics information given by users. This method takes advantage of the human understanding of the measure. The second technique consists in verifying the consistency between the results given by metric formula and the human judgement (Zen and Vanderdonckt, 2016). After quantifying the aesthetics objectively by means of measurements, their alignments with human judgements can be checked thanks to user ratings for instance.

As mentioned previously, measures allow a more objective and rigorous approach to interface aesthetics evaluation. According to Zen (2017), “A visual measure is the application of a mathematical formula based on a visual principle resulting in a quantitatively descriptive measure.” (Zen, 2017, p.51).

Ngo et al. (2000) were the first to link a mathematical formula to the principles and notions elaborate previously on interface visual design (Zen and Vanderdonckt, 2016). Their contribution was one of the most important in the research on aesthetic measures. Through their work, they have proposed a set of measures that are empirically validated, commonly accepted and recognised in the scientific community (D. C. L. Ngo et al., 2000; D. C. L. Ngo and Byrne, 2001; D. C. L. Ngo et al., 2003). While researchers have become increasingly interested in the aesthetics of GUIs (Lavie and Tractinsky, 2004), the literature on the visual appeal of MUIs is still relatively new. As a matter of fact, no consensus has been reached on how to evaluate the visual design of MUIs (Soui et al., 2017).

## **2.2.4 Visual complexity**

Complexity is linked to the aesthetics concept. Indeed, complexity can predict the perceived aesthetics of GUIs (Miniukovich and De Angeli, 2014b). Moreover, it has been shown in the literature that a negative correlation exists between the two notions. While some studies started by assuming that complexity and aesthetics were strongly correlated, they did not explore the underlying aspects of complexity (Miniukovich and De Angeli, 2015). Especially, few studies address the complexity of the measures of mobile devices, without human participation (Riegler and Holzmann, 2018)

In the human-computer interaction field, three kinds of complexity can be identified: visual, information design and task complexity (Miniukovich and De Angeli, 2015). Through this thesis, we will explore the concept of visual complexity as comprehensively as possible.

Visual complexity is an essential aspect of GUIs as it significantly impacts the user. Firstly, complexity has a strong influence on initial impressions (Reinecke et al., 2013). Specifically, the visually complex interfaces lead to a more negative first impression than the less complex ones. As users are sensitive to this GUI feature, they rather browse a not too complex website (Tuch et al., 2012). Secondly, very complex screens are confusing and decrease user productivity (Riegler and Holzmann, 2018). Furthermore, interface complexity increases face tension, arousal and reaction time which is linked to the performance of a visual searching task. A complex interface will also impact the memory capacity of the user navigating it (Tuch et al., 2009).

# Chapter 3

## Systematic Literature Review

To answer the research question, “Which are the measures to assess the perceived aesthetics of a website in general and of electronic commerce websites in particular, on mobile devices?”, we tried to identify and aggregate the published studies on the subject. To this end, we adopted the Systematic Literature Review (SLR) procedure. In this chapter, we will describe the method and present the different results obtained with this approach.

### 3.1 Method

To get an overview of the studies carried out on the subject, we deployed a Systematic Literature Review. A SLR is a rigorous and systematic approach to locate, assess, and aggregate studies on a specific and defined topic, question or phenomenon. This scientific method aims to aggregate knowledge and produce objective summaries of the results related to a defined topic. The generated outcomes attempt to be unbiased, reproducible, and reliable (Kitchenham et al., 2010; Pearl Brereton et al., 2007; Villarreal-Narvaez et al., 2020). We deployed the SLR through 4 phases: identification, screening, eligibility, and inclusion (Altman et al., 2009; Villarreal-Narvaez et al., 2020).

The first phase is called the identification and consists in identifying the potential relevant studies of a digital library thanks to a specific query (Casteleyn et al., 2014; Kitchenham et al., 2010). We started by realising our research on Association for Computing Machinery Digital Library<sup>1</sup> (ACM DL) as it is one of the key electronic databases in the field (Table 3.1). We began with a fairly restrictive query. Since we did not obtain many results after going through the stages of the Systematic Literature Review, we changed the terms and relaxed the constraints imposed by them. More detailed information on the subject can be found below (see Section 3.2). Furthermore, we tested another

---

<sup>1</sup>See <https://dl.acm.org/>

database by performing our research on Google Scholar <sup>2</sup> which lists articles from a wider range of fields.

In this identification phase, we extracted the 100 most relevant publications for each query. As we deployed five queries, we saved, in total, 500 papers from the 25<sup>th</sup> of October to the 28<sup>th</sup> of November 2021. The rules applied to each library are specified in Table 3.1. For all digital libraries, results were sorted in order of relevance. Therefore, our findings rely on the relevance ranking algorithm of the research engines. Before moving to the next phase, we removed the double entries.

The screening step consists in discarding some studies according to the form and content criteria (Casteleyn et al., 2014). The form criterion suggests to focus only on articles for which the full text is available, written in English and peer-reviewed. The peer-reviewed references include publications such as studies published in peer-reviewed journals, workshops, and conferences. Conversely, the un-reviewed references are, for instance, PhD. theses, master theses, books, tutorials, patents and slideshows (Casteleyn et al., 2014; Villarreal-Narvaez et al., 2020). As regards the content criteria, we retained articles that explicitly mention a measure of aesthetics, visual complexity or visual design for mobile devices. Papers related to other topics were excluded.

The third step is called eligibility and excludes articles that do not match predefined criteria related to the study and report characteristics (Altman et al., 2009). More precisely, we only accepted articles containing automatic, formal and calculable aesthetics or complexity measures related to the visual design of a mobile interface.

Finally, the inclusion phase verifies quantitative and qualitative information about the articles (Villarreal-Narvaez et al., 2020). Nevertheless, this step will not be explained in detail in this thesis.

---

<sup>2</sup>See <https://scholar.google.com/>

| Q | Databases      | Searching strings                                                                                                                                  | Advanced search options                                                                                           | Number of results<br>(in thousands) |
|---|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| 1 | ACM            | (automat*) AND (eval* OR assess*)<br>AND (aesthetic* OR visual OR complexity)<br>AND (interface*) AND<br>((smartphone*) OR (tablet*) OR (mobile*)) | Search items from:<br>ACM Guide to Computing Literature<br>Publication date: All dates<br>Search within: Anywhere | 57                                  |
| 2 | ACM            | (eval* OR assess*)<br>AND (aesthetic* OR visual OR complexity)<br>AND (interface*) AND<br>((smartphone*) OR (tablet*) OR (mobile*))                | Search items from:<br>ACM Guide to Computing Literature<br>Publication date: All dates<br>Search within: Anywhere | 93                                  |
| 3 | ACM            | (eval* OR assess*)<br>AND (aesthetic* OR visual OR complexity)<br>AND<br>((smartphone*) OR (tablet*) OR (mobile*))                                 | Search items from:<br>ACM Guide to Computing Literature<br>Publication date: All dates<br>Search within: Anywhere | 172                                 |
| 4 | ACM            | (automat*) AND (metric* OR measure*)<br>AND (aesthetic* OR visual)<br>AND (interface*) AND<br>((smartphone*) OR (tablet*) OR (mobile*))            | Search items from:<br>ACM Guide to Computing Literature<br>Publication date: All dates<br>Search within: Anywhere | 23                                  |
| 5 | Google Scholar | (interface AND mobile)<br>(aesthetic OR aesthetics OR<br>visual OR complexity)                                                                     | Find articles: with all the words<br>with at least one of the words<br>Search within: Anywhere in the article     | 2,570                               |

Table 3.1: Searching strings and options of the libraries

## 3.2 Results and discussion

During our study, we went through the different phases of the SLR several times. Each time, we changed the query and/or the library used (Table 3.1). The reasoning behind these modifications will be explained hereafter, as well as the obtained results.

### 3.2.1 Query 1

We started by creating a query according to the theme of our research question on ACM DL, one of the most well-known digital libraries in the field:

$$Q = (\text{automat*}) \text{ AND } (\text{eval* OR assess*}) \text{ AND } (\text{aesthetic* OR visual OR complexity}) \text{ AND } (\text{interface*})$$

Nevertheless, this explanatory query allowed to identify roughly 173,000 results. Given the fact that we tried to find few but very relevant publications, we chose to refine our query by specifying the type of devices:

$$Q = (\text{automat*}) \text{ AND } (\text{eval* OR assess*}) \text{ AND } (\text{aesthetic* OR visual OR complexity}) \text{ AND } (\text{interface*}) \text{ AND } ((\text{smartphone*}) \text{ OR } (\text{tablet*}) \text{ OR } (\text{mobile*}))$$

The latter allows to identify approximately 57,000 results from which we extracted the 100 most relevant ones.

#### 3.2.1.1 Measuring the visual impressions and the complexity of mobile apps

One of the main interesting papers obtained thanks to the first query, in position 4 of the ranking, is the Miniukovich and De Angeli’s research (2014b). The study is a continuation of their previous paper which focused on the automatic evaluation of complexity on desktop (Miniukovich and De Angeli, 2014a). Through the present study, the researchers tried to extend the usage of six automatic complexity metrics to mobiles (more precisely mobile apps) by making minor adjustments. They also verified the predictive power of these measures with regard to the complexity and aesthetic assessments left by users. According to the authors, complexity is composed of three determinants (Miniukovich and De Angeli, 2014b).

The first complexity determinant is the amount of information which is made up of three components: the visual clutter, the dominant colours, which is related to the number of different colours, and the colours depth, which is linked to the colour gradual transition. Different low-level measures were computed to evaluate the three components of the amount of information. These submeasures were deployed in a factor analysis.

Firstly, two low-level measures are loaded on one factor that we can describe as the dominant colour: the number of static clusters and the number of dynamic clusters. The number of static clusters is obtained after clustering the pixel colour values. It is noteworthy that the colour values that appear less than five times are ignored. This threshold was adapted specifically for mobile devices. As far as the number of dynamic clusters is concerned, two colours belong to the same cluster when the distance between their positions in the colour cube is equal to or lower than three. Secondly, three measures are loaded on a factor which can be interpreted as the colours depth: the distinct pixel RGB values, the PNG file size, and the proportion of different RGB values on the number of dynamic clusters. Thirdly, the visual clutter is composed of the feature congestion, the file size (JPEG), the subband entropy, and the edge density ratio.

The second complexity determinant is the organisation of information and is assessed through symmetry. This latter concept is measured by the number of symmetrical pixels (viewed from a vertical axis) compared to the number of edge pixels. This ratio is, then, normalised by the edge density. It is worth highlighting that the symmetry threshold was adapted to suit mobile devices.

The third and last determinant of complexity is the discriminability of information which is evaluated by the figure-ground contrast and the edge congestion. Regarding the figure-ground contrast, the authors computed “the proportions of edge pixels on several consecutive luminance levels. Then we weighted the proportions ..., summed them up and normalized by the overall number of edge pixels.” (Miniukovich and De Angeli, 2014b, p.37). The algorithm was adjusted for mobiles by slightly decreasing the consecutive luminance levels to obtain ranges between 0.05 and 0.65. Indeed, mobiles seem to include less background contrast. The result is interpreted as the figure-ground contrast. As far as the edge congestion is concerned, this measure is defined as the proportion of congested pixels compared to the total number of edge pixels. A pixel is congested when the edge pixel is within 20 pixels of a pixel belonging to another edge.

This article is interesting as it presents measures of complexity adapted to mobile devices. Some of them are well described, such as symmetry and edge congestion contrast, whereas others could be explained more clearly such as figure-ground contrast. All measures are explicit while not all are formal, neither calculable. Indeed, no detailed information or formula is available to calculate these measures in the same way as the two authors did. This represents a limit to the reproducibility of the experiment.

However, the article links other references to the measures. The article previously written by the same two authors is mainly quoted. Therefore, we investigated if this article is more accurate in the definition of the measures by clearly explaining their meaning as well as their formulas. The measures were computed automatically thanks to existing or

new algorithms. Indeed, the authors created algorithms, implemented in MatLab, for the computation of symmetry, edge congestion and the figure-ground contrast. The related formulas are presented in the article and will be detailed hereafter. The article also briefly explains the submeasure of visual clutter, colour depth and dominant colours which were computed thanks to existing algorithms. However, these submeasures are not represented with their related formulas in the article (Miniukovich and De Angeli, 2014a).

**Symmetry** refers to the contours reflection around a vertical axis. The algorithm created by the two authors detects edges and filters them, before computing the symmetry. Indeed, contours, which are parallel to the horizontal axis, are discarded as they are, most of the time, symmetrical along the vertical axis. Moreover, when a group of pixels is detected as belonging to an edge, the algorithm sets aside the pixels which are within a radius of 4 pixels from what we will call a key point (a specific contour pixel that is not filtered). In their paper, Miniukovich and De Angeli (2014a) calculate the symmetry thanks to the ratio of the number of matches among the key points to the total number of key points. This ratio is normalised by the probability of matches among the key points and is defined as the symmetry measure (equation 3.1) (Miniukovich and De Angeli, 2014a):

$$Sym_{norm} = \frac{k_{sym}}{k_{all}} \times (1 - \frac{(k_{all} - 1) \times S_a}{S_s})^{-1} \quad (3.1)$$

where  $k_{all}$  are the key points which represent the unfiltered edge pixels whereas  $k_{sym}$  are the key points which possess a matching pixel across the vertical axis.  $S_a$  is the size of the area where the matching pixels are searched. In this paper, this variable is constant and equals to 4.  $S_s$  represents the size of the screenshot.

The **edge congestion** is related to the phenomenon that below a certain distance users struggled to differentiate objects. According to Miniukovich and De Angeli (2014a), the edge congestion is defined as the proportion of congested edge pixels to the total number of edge pixels (equation 3.2). Besides, to be considered as congested, a contour pixel must be within 20 pixels of another edge pixel. Moreover, the contour pixel is detected thanks to an algorithm which observes the colour differences in the neighbourhood of the latter (Miniukovich and De Angeli, 2014a).

$$cong = \frac{p_c}{p_a} \quad (3.2)$$

where  $p_c$  is the number of pixels considered as congested whereas  $p_a$  is the number of contour pixels.

The **figure-ground contrast** is linked to the colour and luminance changes of neighbouring zones. These pixel differences in colour and luminance allow the edge detec-

tion of different edge strength levels thanks to sensitivity thresholds. A high threshold is only able to identify the strong contours whereas the low one identifies the weak edges as well. The figure-ground contrast is a function of the number of edges pixels in the different levels. The sum of the pixels subtractions, which are computed between the consecutive levels, is normalised and weighted (equation 3.3). Indeed, edges from level close to one obtained a high weight as they lead to visual differentiation troubles (Miniukovich and De Angeli, 2014a).

$$E_{norm} = \frac{\sum_{l=1}^6 (E_l - E_{l+1})}{E_l - E_6} \times (1 - \frac{(l-1)}{6}) \quad (3.3)$$

where  $E_l$  is the number of pixels belonging to a level  $l$ .

### 3.2.1.2 Enumerating several complexity factors

In his paper, which appears in position 10 of the most relevant articles related to the first query, Gaffar (2009) presents the different complexity aspects of mobile applications. The complexity factors are divided in three categories: design time challenges, user experience, and, finally, content and capabilities of the system. Nevertheless, the article does not present calculable formulas and most of the factors do not address visual complexity of the interface.

### 3.2.1.3 Assessing the visual design of Android apps

Among the results of the first query, we found, in position 36, a paper which introduces CodeMaster UI Design (Solecki et al., 2020). This tool evaluates the visual design of mobile applications which were developed using a specific programming environment for Android apps, called App Inventor. On CodeMaster UI Design, the visual design evaluation is processed by extracting the application code and assessing the performance level of predefined criteria. These criteria are represented by a question and cover various aspects of the app design such as colours, typography, and layout. For instance, one criterion related to the colour category is: “How many colors are used (other than black, white, and gray)?” (Solecki et al., 2020, p. 53). The tool displays an overall design grade as well as the score of each category and criterion. Besides, the category score is the sum of the criteria scores included in this category. The assessment grade is computed as follows (equation 3.4)(Solecki et al., 2020):

$$Grade = \frac{\text{score achieved}}{\text{maximum score of the assessed criteria}} \times 10 \quad (3.4)$$

However, the article does not explain in detail how the performance level of the criteria is computed and what the formulas behind it are. This limits the ability to

replicate this experiment by evaluating the visual design in a similar way to what was done in the paper.

#### **3.2.1.4 Measuring the simplicity of smartphones**

The first search on ACM DL provides, in position 95, an article about the interface simplicity of smartphones. Choi and Lee (2012) model simplicity as composed of three aspects. Two are more cognitive and functional (information design and task complexity), whereas one is related to the visual design (visual aesthetics). Nevertheless, they assessed the three simplicity aspects, which were divided in construct and items, using Likert scales. Therefore, the study results rely on user’s points of view (Choi and Lee, 2012).

The article presents so-called self-reporting measures which required the collection of user evaluations. Therefore, this procedure is subject to variations and may be far from reality. Indeed, what is reported by a user may not correlate with the objective reality (Abeele et al., 2013).

### **3.2.2 Query 2**

Most of the publications retrieved from the first search satisfy the query syntactically but not semantically (see Section 3.2.1). Indeed, they deal with topics other than the subject of our research question and are, therefore, not relevant to our study. To overcome this, we implemented several strategies (Table 3.1). The first one consists in relaxing the constraints imposed by the query. We carried out new searches by, first, removing the “*automat\**” truncation from the query (where the asterisk represents any unknown character). This second search retrieved more articles than the first one (roughly 93,000). It should be noted that the publication of Miniukovich and De Angeli (2014b) about the complexity of mobile interfaces appears once again among the first results (in position 9). Even if this second query allows to locate a larger number of articles, they are not necessarily more relevant regarding our research objective.

### **3.2.3 Query 3**

In a third stage, we have pursued the strategy, which consists in relaxing the constraints imposed by the query terms, by removing the “*interface\**” truncation. Similarly, this query does not enable us to identify more interesting articles. In addition, it is noteworthy that the article of Miniukovich and De Angeli (2014b) appears in position 95. Therefore, this new query seems less effective in identifying articles relevant to our research.

### 3.2.4 Query 4

The second strategy consists in removing and replacing some terms of the first query. Through the first research, we have observed that the “complex\*” truncation often leads to articles that are not related to visual complexity. Indeed, they are rather related to other complexity aspects of systems, interfaces, etc. For this reason, we removed this truncation from our query. Moreover, we replaced the notion of “eval\*” and “assess\*” by that of “measure\*” and “metric\*”.

This fourth query identifies 23,000 publications, among which the article written by Miniukovich and De Angeli (2014b) appears in the first ranking position. In position 15, the study conducted by Bessghaier et al. (2021) seems highly relevant and will be detailed hereafter. As regards the irrelevant articles, they could have been discussed, by analogy with what was realized in the previous sections (see Sections 3.2.1 and 3.2.2).

#### 3.2.4.1 Detecting the aesthetic defects of mobile interfaces with measures

In their research, Bessghaier et al. (2021) introduce a tool which detects aesthetic defects on MUIs using automatic metrics. In order to detect design mistakes, they create defect detection rules with an optimisation algorithm by combining structural measures. In total, they present 15 measures with their related formulas: balance, economy, cohesion, layout uniformity, simplicity, density, regularity, sequence, homogeneity, grouping, unity, integrality, number of elements, complexity and, finally, clarity (Bessghaier et al., 2021).

According to Bessghaier et al. (2021), the **density** ( $Dst$ ) is related to the elements distribution across the different parts of the screen (equation 3.5):

$$Dst = 1 - 2|0.5 - \frac{\sum_i^n a_i}{a_{frame}}| \in [0, 1] \quad (3.5)$$

where  $a_{frame}$  and  $a_i$  are, respectively, the areas of the frame and an object  $i$ , while  $n$  is the number of objects on the layout.

The **sequence** ( $SQM$ ) is linked to the elements organisation. The measure is defined based on the area  $a_{ij}$  of an object  $i$  located in the quadrant  $j$  (which is Upper-Left (UL), Upper-Right (UR), Lower-Left (LL) or Lower-Right (LR)) (equation 3.6) (Bessghaier et al., 2021):

$$SQM = 1 - \frac{\sum_{j=UL,UR,LL,LR} |q_j - v_j|}{8} \in [0, 1] \quad (3.6)$$

where  $v_j$  and  $w_j$  are defined as follows:

$$v_j = \begin{cases} 4, & \text{if } w_j \text{ is the largest in } w \\ 3, & \text{if } w_j \text{ is the } 2^{nd} \text{ largest in } w \\ 2, & \text{if } w_j \text{ is the } 3^{rd} \text{ largest in } w \\ 1, & \text{if } w_j \text{ is the smallest in } w \end{cases}$$

$$w_j = q_j \sum_i^{n_j} a_{ij} \quad (3.7)$$

with:  $q_j = \{ q_{UL}, q_{UR}, q_{LL}, q_{LR} \} = \{ 4, 3, 2, 1 \}$  and  $w = \{ w_{UL}, w_{UR}, w_{LL}, w_{LR} \}$

The **regularity** ( $Rgl$ ) refers to elements spacing on an horizontal and a vertical axis. Bessghaier et al. (2021) computed the regularity measure based on  $RM_{alignment}$  and  $RM_{spacing}$  which are, respectively, related to the minimisation and the consistency spacing of the alignment points (equation 3.8) (see Appendix A.1):

$$Rgl = \frac{|RM_{alignment}| + |RM_{spacing}|}{2} \in [0, 1] \quad (3.8)$$

The **homogeneity** ( $HM$ ) relates to the elements' distribution among the four quadrants and is a function of the number of possible objects arrangements across these four screen parts  $W$  (equation 3.9) (see Appendix A.2) (Bessghaier et al., 2021):

$$HM = \frac{W}{W_{max}} \in [0, 1] \quad (3.9)$$

According to Bessghaier et al. (2021), the **economy** ( $Ecm$ ) is inversely proportional to the number of different element sizes ( $n_{size}$ ) and aims to favour a clear and simple display (equation 3.10):

$$Ecm = \frac{1}{n_{size}} \in [0, 1] \quad (3.10)$$

The **simplicity** ( $Smpl$ ) involves the minimization of the total number of objects ( $n$ ) and the number of vertical ( $n_{vap}$ ) as well as horizontal alignment points ( $n_{hap}$ ) (equation 3.11) (Bessghaier et al., 2021):

$$Smpl = \frac{3}{n_{vap} + n_{hap} + n} \in [0, 1] \quad (3.11)$$

The **balance** ( $BM$ ) refers to the elements' distribution across the interface. The concept is divided into two dimensions: the vertical  $BM_{vertical}$  and the horizontal  $BM_{horizontal}$

balance (equation 3.12) (see Appendix A.3) (Bessghaier et al., 2021):

$$BM = 1 - \frac{|BM_{vertical}| + |BM_{horizontal}|}{2} \in [0, 1] \quad (3.12)$$

The **cohesion** ( $CM$ ) measures the extent to which the elements are interrelated. It is a function of the two height-width ratios of the layout and screen ( $CM_{fl}$ ), on the one hand, and of the two height-width ratios of the objects and layout ( $CM_{lo}$ ), on the other hand (equation 3.13) (see Appendix A.4) (Bessghaier et al., 2021):

$$CM = \frac{|CM_{fl}| + |CM_{lo}|}{2} \in [0, 1] \quad (3.13)$$

The **layout uniformity** ( $LU$ ) is related to the layout consistency and the components structure (equation 3.14) (Bessghaier et al., 2021):

$$LU = 100 \times (1 - \frac{(N_h + N_w + N_t + N_l + N_b + N_r) - M}{6N - M}) \in [0, 1] \quad (3.14)$$

where  $N_h$  and  $N_w$ , are respectively, the number of different heights and widths while  $N$  is the number of objects.  $N_t$ ,  $N_l$ ,  $N_b$  and  $N_r$  are, respectively, the number of top, left, bottom, and right alignments.  $M$  is used for adjustment purposes and can be computed as follows (equation 3.15) (Bessghaier et al., 2021):

$$M = 2 + 2[2\sqrt{N}] \quad (3.15)$$

The **unity** is related to the impression of screen homogeneity. The unity measure ( $Unt$ ) is computed thanks to  $UM_{form}$ , which quantifies how similar objects are in size and shape, and  $UM_{space}$ , which is related to the space between widgets and that of the margins (equation 3.16) (see Appendix A.5) (Bessghaier et al., 2021):

$$Unt = \frac{|UM_{form}| + |UM_{space}|}{2} \in [0, 1] \quad (3.16)$$

The **integrality** intends to minimise the number of objects of different sizes and the spaces between the objects and the frame. The integrality measure ( $Intg$ ) is defined as follows (equation 3.17) (Bessghaier et al., 2021):

$$Intg = 1 - (0.5 \times [\frac{n_{size} - 1}{n} + \frac{s_{sc} + \sum i^n a_i}{2a_{MUI}}]) \in [0, 1] \quad (3.17)$$

where  $n$  represents the number of objects and  $n_{size}$  is the number of distinct objects sizes.  $a_{MUI}$ ,  $a_{sc}$  and  $a_i$  are respectively the areas of the mobile interface, the screen and the

interactive object  $i$ . It is noteworthy that  $s_{sc}$ , present in the equation 3.17, is not defined in the article. Therefore,  $s_{sc}$  might be, in fact,  $a_{sc}$  due to a typo.

The **grouping** encourages the grouping of objects in the center of the interface. According to Bessghaier et al. (2021), grouping ( $Grp$ ) can be defined as a function of  $AL$  and  $UM_{space}$  presented in the equations 3.20 and 3.19 (equation 3.18):

$$Grp = (0.85f) \times UM_{space} + (0.15f) \times AL \in [0, 1] \quad (3.18)$$

where  $AL$  is a function of the number of object type  $i$  ( $Ni$ ) as well as the number of, respectively, horizontal ( $nHap$ ) and vertical ( $nVap$ ) alignment points (equation 3.19) (Bessghaier et al., 2021):

$$AL = \min \left( 1, \frac{(4 \times Ni - (\frac{nHap \times nVap}{Ni}))}{4 \times Ni - 4} \right) \quad (3.19)$$

The  $UM_{space}$  is computed based on the area of the layout ( $a_{layout}$ ), the area of the frame ( $a_{frame}$ ) and the sum of the objects areas (SumOfArea) (equation 3.20) (Bessghaier et al., 2021):

$$UM_{space} = 1 - (a_{layout} - \frac{\text{SumOfAreas}}{a_{frame}} - \text{SumOfAreas}) \quad (3.20)$$

It should be noted that the term  $f$ , in the equation 3.18, is not defined in the article. Also, no additional information is given to help us understand the different possible types of object  $i$  (equations 3.19).

The **clarity** ( $CL$ ) is a measure of the elements comprehensibility (equation 3.21) (Bessghaier et al., 2021):

$$CL = \frac{\sum_{i=1}^{i=n} \text{label} + \text{title} + \text{hint}}{\sum_{i=1}^{i=n} \text{button} + \text{item} + \text{EditText}} \in [0, 1] \quad (3.21)$$

where  $label$ ,  $title$ ,  $hint$ ,  $button$ ,  $item$  and  $EditText$  are, respectively, the numbers of android text, title, hint label, buttons, menu items and editTexts elements.

The **number of elements** ( $NM$ ) is related to the number of components on an interface (equation 3.22) (Bessghaier et al., 2021):

$$NM = \sum i^{nb} Widgets \in [0, 1] \quad (3.22)$$

It is worth highlighting that the variables of the equation 3.22 are not defined in the article.

The **complexity** is, according to Bessghaier et al. (2021), the sum of all other measures presented by the authors (see Section 3.2.4.1). Therefore, the complexity (*Compl*) includes all the other structural aspects of the interface (equation 3.23):

$$Compl = \frac{Dst + SQM + Rgl + HM + Ecm + Smpl + BM}{\dots} \quad (3.23)$$

$$\frac{+CM + LU + Unt + Intg + Grp + CL + NB}{14} \in [0, 1]$$

where *Dst*, *SQM*, *Rgl*, *HM*, *Ecm*, *Smpl*, *BM*, *CM*, *LU*, *Unt*, *Intg*, *Grp*, *CL* and *NB* are, respectively, the density, the sequence, the regularity, the homogeneity, the economy, the simplicity, the balance, the cohesion, the layout uniformity, the unity, the integrality, the grouping, the clarity and the number of elements.

### 3.2.5 Query 5

In the last strategy, we changed the library. We chose the multipublisher source Google Scholar, which gathers publications of numerous fields. As this search engine allows for less complex queries than those used on ACM DL, we downgraded and adapted our request (Table 3.1). The first publication retrieved, in position 1, is highly relevant with regard to our research question and detailed hereafter (Riegler and Holzmann, 2018). By contrast, the other searching results will not be introduced. Indeed, they are not directly related to the subject of this thesis, rely on user evaluations, or do not detail a formula for measure computation.

#### 3.2.5.1 Measuring the complexity of mobile applications

Riegler and Holzmann (2018) established measures and their related formulas to assess the complexity of mobile applications. These measures are computed automatically, after having used a region detection algorithm. They also have been subject to verification through an experiment.

The **element smallness** relates to the sizes of the screenshot elements. According to Riegler and Holzmann (2018), the element smallness ( $E_s$ ) is a function of the elements averages width and height over all the groups of a screen  $s$  noted  $\overline{width_s}$  and  $\overline{height_s}$ , respectively (equation 3.24) (see Appendix A.6):

$$E_s = \frac{\left(1 - \frac{\overline{width_s}}{\overline{width_s}}\right) + \left(1 - \frac{\overline{height_s}}{\overline{height_s}}\right)}{2} \quad (3.24)$$

It is noteworthy that  $\overline{width_s}$  and  $\overline{height_s}$  are normalised in the range included between 0 and 1. Indeed, a larger score of element smallness stands for smaller elements on

the screen which are more difficult to identify.

Riegler and Holzmann (2018) present a measure of **misalignment** as aligned elements are perceived as more aesthetically pleasant than non-aligned ones. The misalignment measure ( $M_s$ ) is composed of three submeasures: the average number of alignments of a screen  $s$ , from a horizontal ( $\overline{numh_s}$ ), a vertical ( $\overline{numv_s}$ ) and a central ( $\overline{numc_s}$ ) perspective. More specifically, the misalignment is defined as the opposite of these three alignment submeasures (equation 3.25) (see Appendix A.7) (Riegler and Holzmann, 2018):

$$M_s = 1 - (\overline{numh_s} + \overline{numv_s} + \overline{numc_s}) \quad (3.25)$$

**Density** ( $D_s$ ) is related to the space left empty and the extent to which the elements occupy the screen (equation 3.26) (Riegler and Holzmann, 2018):

$$D_s = \frac{1}{A_s} \sum_{u=1}^{U_s} A_{u,s} \quad (3.26)$$

where  $A_{u,s}$  is the area of an element  $u$  (included in  $U_s$ , the set of all elements of a screen  $s$ ) belonging to a screen  $s$  whereas  $A_s$  is the area of a screen.

According to Riegler and Holzmann (2018), **imbalance** is composed of the mean horizontal balance and the mean vertical balance. The weighted mean of the vertical balance related to a screen  $s$  ( $\overline{balv_s}$ ) is a function of the top and bottom margins of an element whereas the weighted mean of the horizontal balance ( $\overline{balh_s}$ ) is computed, analogously, with the left and right margins. Besides, the element margins are the distances between the closest elements or the borders in a specific direction such as the top or the left. A measure of imbalance ( $B_s$ ), contrary to the measure of balance, can be defined as follows (equation 3.27) (see Appendix A.8) (Riegler and Holzmann, 2018):

$$B_s = 1 - \frac{\overline{balh_s} + \overline{balv_s}}{2} \quad (3.27)$$

A **colour complexity** measure is introduced by Riegler and Holzmann (2018) as colours are an important aspect of the interfaces. Indeed, using colours positively impacts the performance, the visual search time, the memorisation and the perception of users. Furthermore, colour misuses can be distracting for the users and reduce their usability perception of the system. The colour complexity measure is related to the number of dominant colours ( $numc_s$ ) of a screen  $s$  which appear more frequently than others. For each dominant colour pair combination, a colour complexity value ( $ccm_{c1,c2}$ ) is calculated as a function of the H (hue), S (saturation) and V (value) values of the two dominant colours ( $c_1$  and  $c_2$ ). The colour complexity ( $C_s$ ) can be defined using the later measures

(equation 3.28) (see Appendix A.9) (Riegler and Holzmann, 2018).

$$C_s = \frac{\frac{numc_s}{numcmax} + (1 - \overline{ccm_s}) + c_{h_s, h_{s+1}}}{3} \quad (3.28)$$

where  $numcmax$  is the number of dominant colours and  $c_{h_s, h_{s+1}}$  is the colour combination match of a screen's histogram ( $h_s$ ) and that of its successor ( $h_{s+1}$ ).  $\overline{ccm_s}$  is the mean complexity value of the colour combination.

The **typographic complexity** expresses the fact that a higher contrast between the text and background colour increases the legibility. Furthermore, a small font raises the fixation duration. Therefore, the typographic complexity measure ( $T_s$ ) is defined as a function of the text and background colour contrasts as well as the text size (equation 3.29) (Riegler and Holzmann, 2018):

$$T_s = \frac{\frac{numt_s}{numtmax} + \left(1 - \frac{\overline{t_s}}{tmax}\right) + (1 - \overline{ccm_s})}{3} \quad (3.29)$$

where  $numt_s$  is the number of unique text sizes (or text heights) while  $numtmax$  is its maximum. Similarly,  $\overline{t_s}$  is the average text size for a screenshot  $s$  and  $tmax$  is the maximum text size.  $\overline{ccm_s}$  is the mean of the text foreground/background colour combination match (see colour complexity measure in Section 3.2.5.1).

Consistency is an important interface feature. Indeed, it eases the recognition of the interface central objectives. The side information is also more easily filtered. In addition, an inconsistent interface regarding the colours used results in a reduction of user satisfaction. According to Riegler and Holzmann (2018), **inconsistency** ( $I$ ) includes various visual facets of an interface. More specifically, the measure is defined using the  $N$  metrics presented in this study through different screens of an application (equation 3.30):

$$I = \frac{1}{N} \sum_{n=1}^N \frac{1}{xmax_n} \sqrt{\frac{1}{J-1} \sum_{j=1}^J (x_{j_n} - \overline{x_n})^2} \quad (3.30)$$

where  $x_{j_n}$  is the value of the metric  $n$  for the screen  $j$ .  $xmax_n$  and  $\overline{x_n}$  are, respectively, the maximum value and the average of a metric  $n$  across the different captured screens.  $n$  is included in  $N$  which is the set of metrics presented by Riegler and Holzmann (2018): element smallness, the misalignment, the density, the imbalance, the colour complexity, the typographic complexity and the number of elements. Besides, the number of elements which is the sum of the screen's elements is not defined by a written mathematical formula in the article.

### 3.3 Conclusion of the Systematic Literature Review

In our Systematic Literature Review, we collected articles that explicitly and formally present calculable measures of visual complexity or aesthetics for mobile interfaces. We also briefly presented the related formulas and measures. Through this approach, we identified two relevant articles that perfectly meet the stated criteria (Bessghaier et al., 2021; Riegler and Holzmann, 2018). A third paper represents an extension of complexity measures from desktops to mobiles. Although the research extension does not present any formula, it clearly mentions the earlier publication of the two authors and indicates the adjustments made to the previous formulas (Miniukovich and De Angeli, 2014b). However, the other articles retrieved are irrelevant to our research question.

Firstly, among our different queries results, numerous articles satisfy the search syntactically but are semantically irrelevant. Indeed, although they fulfil the criteria imposed by the query, they do not address our research question. For example, a significant number of papers deal with topics other than the evaluation of visual interface design, such as the complexity of video coding or system portfolios.

Secondly, other papers present explicit MUIs measures of aesthetics or complexity. Nevertheless, the measures appear to be self-reported. These non-automatic measures depend on what is reported by users. They are subject to many biases since they rely on the users' perceptions. Given that perceptions are subjective and not always in line with reality, these measures are not ideal for obtaining transposable and generalisable results (Abeele et al., 2013).

Finally, other articles present explicit and automatic measures. However, they do not formally define them, nor do not present any formula. Without having access to the formula and the entire description behind the measures, it is difficult to reproduce the experiment.

In conclusion, despite the precision of the queries, the results are not very meaningful. There may be two reasons behind this observation. Firstly, the terms and queries used would not allow an efficient identification of articles on the subject. Indeed, our findings rely on the terms of our query and the indexation of published studies. Secondly, there could be little amount of work on the subject. This hypothesis seems plausible since we noticed that the existing studies on the visual design measures of mobile interfaces are recent. Moreover, the previously identified works mention the fact that there are few other works that relate to the exact same subject (Riegler and Holzmann, 2018).

# Chapter 4

## Experiment 1 - SimilarWeb / QUESTIM

In this chapter and the next two, we will explore through different experiments some aesthetics measures of websites in order to answer our research question.

**Which are the measures to assess the perceived aesthetics of a website in general and of electronic commerce websites in particular, on mobile devices?**

Therefore, we will deploy our experiments on e-commerce sites viewed on mobile devices and use two kinds of approaches to tackle this question: qualitative evaluation (beauty ratings, complexity ratings, and ergonomic criteria) and quantitative evaluation based on two sets of measures (the QUESTIM measures and the AIM measures).

In this first experiment, we will go through and analyse the aesthetics measures offered by Quality Estimator using Measures<sup>1</sup> (QUESTIM). QUESTIM is an application available online which evaluates the aesthetics of interfaces thanks to visual measures (Zen, 2017). These 16 measures will be semi-automatically computed for 33 e-commerce websites. These sites were retrieved, thanks to UILAB, from SimilarWeb<sup>2</sup> which suggests websites rankings based on the traffic registered on every site. Besides, UILAB is a software used to carry out experiments on GUIs visual design. This tool was developed to ease the design of experiments and support their reproducibility (Burny and Vanderdonckt, 2021).

### 4.1 Experimental method

The first step of the research method started with the screenshot collection in October 2020. We extracted the URL and a homepage screenshot of the first 50 most popular websites in the “E-commerce and Shopping” category of SimilarWeb thanks to UILAB. The homepages were captured in a portrait mode with a 414x732 resolution and saved in

---

<sup>1</sup>See <http://questimapp.appspot.com/>

<sup>2</sup>See <https://www.similarweb.com/fr/top-websites/>

a PNG file format. After downloading manually each screenshot in a file locally stored, the screenshot gallery was filtered. The websites which were not accessible were eliminated, such as [www.costco.com](http://www.costco.com). Furthermore, in December 2020, the screenshots of some websites were replaced by manually captured screenshots if they presented a default. For example, for some screenshots, the portrait mode has not been correctly loaded when the screen was captured or the website detected too rapid navigation which blocked the access. It concerns seven websites which were captured thanks to Snagit<sup>3</sup> (a software which allows to make advanced screenshots) and Window Resizer<sup>4</sup> (a Google Chrome extension to define different browser resolutions). Finally, after collecting, editing and excluding some screenshots, the gallery consists of 33 different screenshots.

Secondly, we proceeded manually to the zone segmentation of the screenshots thanks to QUESTIM. Each zone, characterized by a rectangle, was manually defined by applying a consistent reasoning to the entire screenshot gallery. For example, non-interacting banners and text menus are defined as separate zones. Banners with interactions (e.g., “click me”) also have a zone delimiting this button (Figure 4.1). This zone segmentation outcomes in a JSON file displaying all areas defined by  $x$  and  $y$  coordinates, a width, and a height.

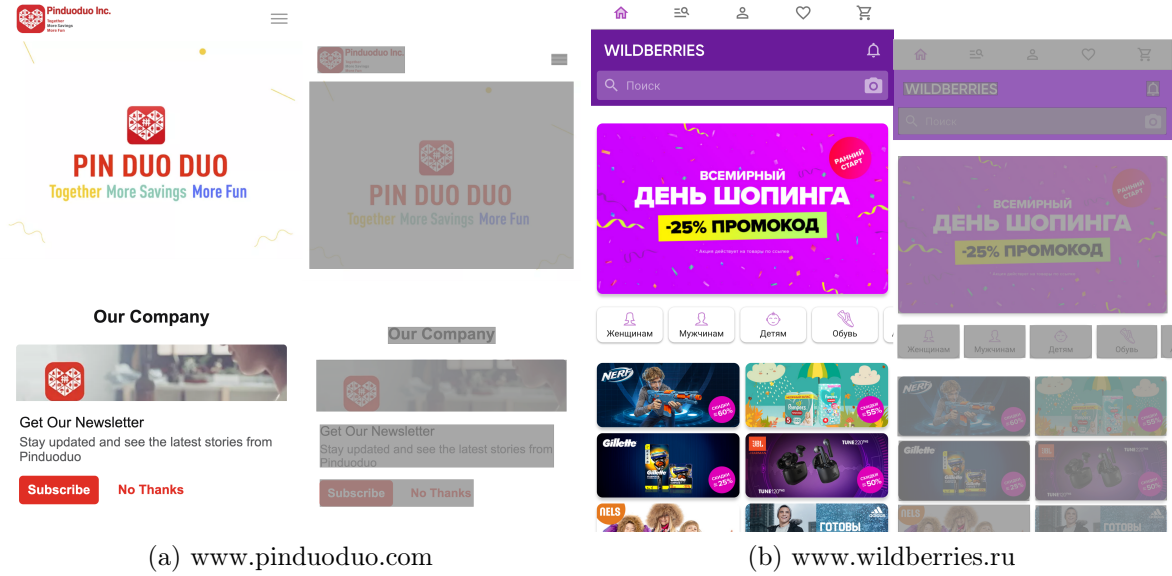


Figure 4.1: Zone segmentation on QUESTIM

On QUESTIM, the 16 aesthetics measures of each screenshot were computed automatically after having manually identified the different zones of the screenshot. It is noteworthy that opening the screenshots on QUESTIM, defining the zones manually, letting the system calculate the 16 measures automatically and downloading the output

<sup>3</sup>See <https://snagit.fr.softonic.com/>

<sup>4</sup>See <https://coolx10.com/window-resizer/>

files, separately, for each of the 33 images, took 2 hours and 36 minutes in total. The 16 QUESTIM measures used in our analysis are briefly described below (Zen, 2017).

- The external alignment is the proportion of zone sides that are aligned vertically or horizontally.
- The center alignment determines the proportion of the zones that have their center aligned vertically or horizontally.
- The balance is the distribution of optical weight, according to D. C. L. Ngo and Byrne (2001), and is inversely proportional to the vertical and horizontal balance. These two measures take into account the area, the shape, the colours of the zones, the distance between the zones and a central axis from, respectively, a vertical and a horizontal point of view.
- The symmetry is the identical correspondence of zones around a center or an axis.
- The density describes to what extent the zones cover the image.
- The simplicity is the directness and singleness character of an arrangement and depends on different factors such as the alignment of the zones.
- The equilibrium is linked to the disposition of the zones on the screenshot. The more the center of the layout corresponds to the center of the screenshot, the closer we get to the equilibrium (D. C. L. Ngo and Byrne, 2001).
- The concentricity describes to what extent the areas are gathered in the middle or placed in the corners of the screenshot.
- The colourfulness describes the intensity of colours.
- The colour density refers to the proportion of colours covering the screen.
- The compression complexity measures the size of the file after the compression of the image by an algorithm.
- The saliency balance refers to the balance computed thanks to a saliency algorithm. The latter evaluates the visibility of each zone by following the human perception as precisely as possible.
- The border detection identifies the borders of objects by detecting visual discontinuities. Therefore, the border balance is computed by counting the pixel luminance in each of the four quadrants of the screenshot. The border density is linked to the number of white pixels on the screenshot.

Before answering the research questions, descriptive statistics were completed to explore each variable separately as well as the links between features. The different

analyses were developed thanks to RStudio<sup>5</sup> and GraphPad Prism<sup>6</sup>.

In a nutshell, the experiment was performed with two different kinds of variables. The screenshot is a nominal variable which is linked to the website from which it is originally taken. Each screenshot also possesses the zone segmentation summary presented in a JSON file. The QUESTIM aesthetics measures are numerical and scale variables (between 0 and 1). A screenshot is linked to these 16 measures computed semi-automatically.

## 4.2 Results and discussion

QUESTIM produces 16 scale variables between 0 and 1. Figure 4.2 depicts the box plots of the aesthetics measures for the 33 websites and the data are summarised in the table 4.1. It should be noted that the data of this table 4.1 were observed on RStudio as well as on GraphPad Prism. The variables are pretty volatile as evidenced by the means, the standard deviations and ranges which are noticeably different. The means and standard deviations vary between 0.058 and 0.976 and between 0.008 and 0.265, respectively.

The equilibrium, border density and symmetry variables are the less spread variables and possess lower variances which are equal to or lower than 0.001. Conversely, colourfulness is the most dispersed variable. Indeed, it possesses the highest standard deviation, variance, and range.

In addition, the border density variable only possesses values lower than 0.163 whereas equilibrium and horizontal balance have values higher than 0.950 and 0.717, respectively. Therefore, we can conclude that the studied screenshots have a low border density and a high equilibrium and horizontal balance. It is noteworthy that the balance and the equilibrium are linked, as improving the balance enhances the equilibrium (Zen, 2017). Moreover, given that the balance is a function of the average of the horizontal and vertical balance, it can be deduced that the horizontal balance pulls up the balance values.

Finally, the Shapiro-Wilk tests reveal that every variable follows the normal distribution ( $p\text{-value} > 0.01$ ) except the border density, compression complexity, horizontal balance and the equilibrium ( $p\text{-value} < 0.01$ ). It is notable that the Shapiro-Wilk tests were carried out on RStudio and on GraphPad Prism and lead to the same results (see AppendixB.1).

The figure 4.3 is a scatter plot between each couple of variables and contains, on the upper part, Spearman correlations and, on the diagonal, histograms. The variables

---

<sup>5</sup>See <https://www.rstudio.com/>

<sup>6</sup>See <https://www.graphpad.com/scientific-software/prism/>

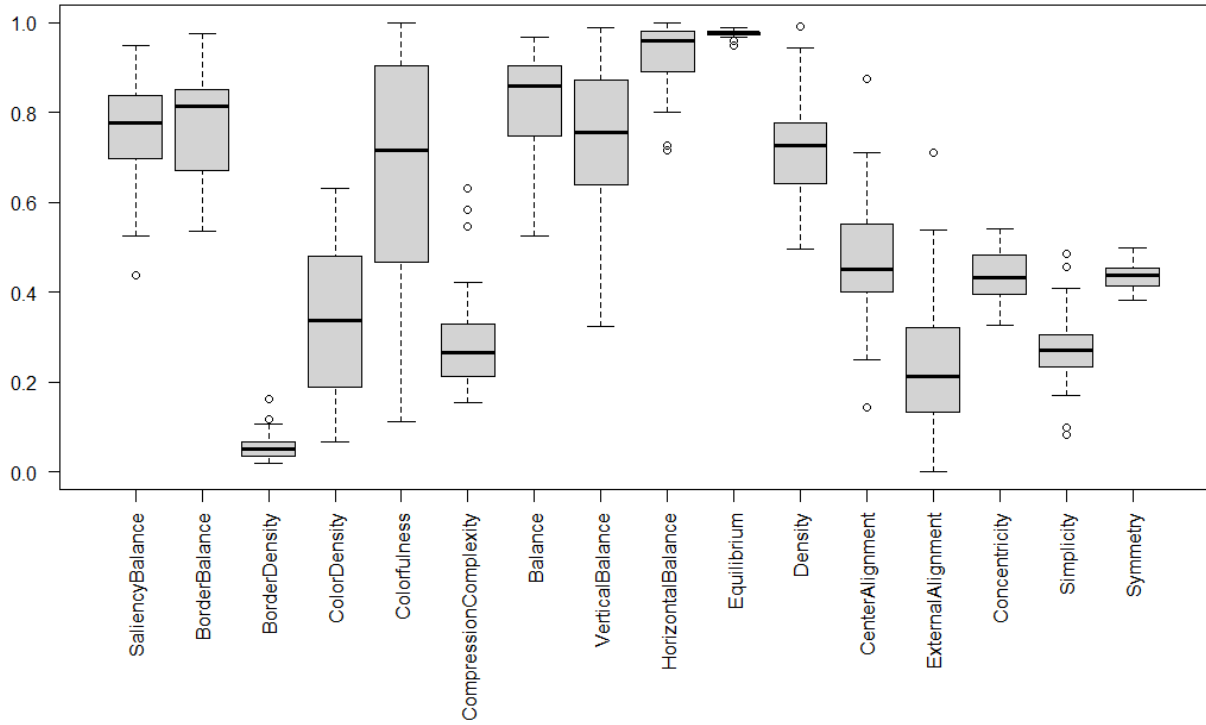


Figure 4.2: Box plot of the QUESTIM measures for SimilarWeb websites

that are most strongly correlated are vertical balance and balance (correlation=0.98), border density and compression complexity (correlation=0.68) and external alignment and compression complexity (correlation=0.64). As balance is a function of horizontal and vertical balance, the vertical balance and the balance are strongly correlated. The strongest negative correlations are associated with the density and simplicity (correlation=-0.59), concentricity and colour density (correlation=-0.4) and horizontal balance and simplicity (correlation=-0.36) variable couples. The more dense a screenshot is, the less simple it will be, as evidenced by the correlation.

QUESTIM proposes a set of measures to assess the aesthetics of Graphical User Interfaces through different aspects. In the following chapter 6, we will explore the link between these measures and the human judgement of aesthetics.

| Measure               | Minimum | Maximum | Range | Mean  | SD    | Variance |
|-----------------------|---------|---------|-------|-------|-------|----------|
| SaliencyBalance       | 0.436   | 0.949   | 0.513 | 0.758 | 0.124 | 0.015    |
| BorderBalance         | 0.536   | 0.975   | 0.439 | 0.774 | 0.121 | 0.015    |
| BorderDensity         | 0.018   | 0.163   | 0.145 | 0.058 | 0.033 | 0.001    |
| ColorDensity          | 0.068   | 0.632   | 0.564 | 0.349 | 0.166 | 0.028    |
| Colorfulness          | 0.111   | 1.000   | 0.889 | 0.677 | 0.265 | 0.070    |
| CompressionComplexity | 0.153   | 0.632   | 0.479 | 0.287 | 0.120 | 0.014    |
| Balance               | 0.526   | 0.968   | 0.442 | 0.829 | 0.106 | 0.011    |
| VerticalBalance       | 0.324   | 0.990   | 0.665 | 0.727 | 0.192 | 0.037    |
| HorizontalBalance     | 0.717   | 0.999   | 0.282 | 0.931 | 0.072 | 0.005    |
| Equilibrium           | 0.950   | 0.989   | 0.039 | 0.976 | 0.008 | 0.000    |
| Density               | 0.497   | 0.991   | 0.494 | 0.722 | 0.114 | 0.013    |
| CenterAlignment       | 0.143   | 0.875   | 0.732 | 0.475 | 0.144 | 0.021    |
| ExternalAlignment     | 0.000   | 0.711   | 0.711 | 0.241 | 0.150 | 0.023    |
| Concentricity         | 0.326   | 0.542   | 0.216 | 0.437 | 0.056 | 0.003    |
| Simplicity            | 0.081   | 0.487   | 0.405 | 0.275 | 0.087 | 0.008    |
| Symmetry              | 0.383   | 0.498   | 0.115 | 0.439 | 0.028 | 0.001    |

Table 4.1: Descriptive statistics of the QUESTIM measures



Figure 4.3: Pair plot of the QUESTIM measures for SimilarWeb websites

# Chapter 5

## Experiment 2 - Alexa ranking / QUESTIM

The second experiment was conducted as part of Antoine Kirsch’s master thesis (Kirsch, 2019). Through this experiment, we will analyse the measures collected by QUESTIM for shopping (e-commerce) websites from the Alexa ranking (no longer available by category since September 2020). To verify that these measures are indeed correlated with human judgement, we will also collect users’ beauty evaluations.

Furthermore, during this experiment, the popularity of the sites will be investigated. In the literature, the link between beauty and usability has been explored through various experiments which have, for example, sought to verify the following statement: “what is beautiful is usable” (Tractinsky et al., 2000, p.127). Instead of conducting an additional experiment on the link between these two variables, we will prefer to study aesthetics from the perspective of popularity. Specifically, we will test whether or not aesthetics is related to the popularity of websites. To exploit these data and to verify the previous statements, the research question can be broken down into the following subquestions linked with their hypotheses which are displayed below.

- Research Question 1.1: Do the QUESTIM measures satisfactorily predict the perceived aesthetics of electronic commerce websites on mobile devices?
  - Hypothesis 1.1: All QUESTIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.
- Research Question 1.2: Are popular electronic commerce websites perceived as aesthetic on mobile devices?
  - Hypothesis 1.2.1: The popular electronic commerce websites are perceived by users as aesthetic on mobile devices.

- Hypothesis 1.2.2: The popular electronic commerce websites are aesthetic (under QUESTIM measures) on mobile devices.

## 5.1 Experimental method

As the experimental methodology is similar to the one used in the previous experiment (see Section 4.1), we will not explain in detail the redundant parts of the experimental method. We started with the screenshot collection of the 100 most popular websites in the shopping category of the Alexa ranking<sup>1</sup> (which no longer offers classification by website category). It is noteworthy that Alexa ranks the websites based on their traffic which is an association of the page views and the number of visitors in a one-month time frame. Among the 100 websites which possess a 414x732 resolution, 21 screenshots were discarded. Indeed, their homepage was not available during the experiment or the QUESTIM measures were not correctly collected.

Secondly, the zone segmentation was completed manually on QUESTIM and 16 measures were automatically computed. Online aesthetics rating were collected on CHILab<sup>2</sup> which is an application dedicated to support experiments on interfaces assessment. Participants were asked to assess a selection of randomly assigned screenshots, one by one, on a 5-point scale (1 = the least aesthetic and 5 = the most aesthetic).

Descriptive statistics were performed to explore the variables and the potential links between them, mainly on RStudio. GraphPad Prism was used to double check the descriptive statistics data and the normality tests, also obtained on RStudio. To answer the research questions, inferential statistics were performed with RStudio. Figure 5.1 summarises the different variables as well as the links that may exist between them. The different statistical tools deployed to answer the research questions are also mentioned.

As far as the variable types are concerned, the screenshot is a nominal variable which is related to its website. The Alexa ranking is an ordinal and numerical variable. This feature refers to the popularity ranking of Alexa for the sites classified in the shopping category. The average beauty rating is a numerical and scale variable (between 1 and 5) which is the average rating left by a sample of participants for a screenshot. The semi-automatic QUESTIM aesthetics measures are numerical and scale between 0 and 1.

---

<sup>1</sup>See <https://www.alexa.com/topsites>

<sup>2</sup>See <http://lrim-chilab.herokuapp.com/>

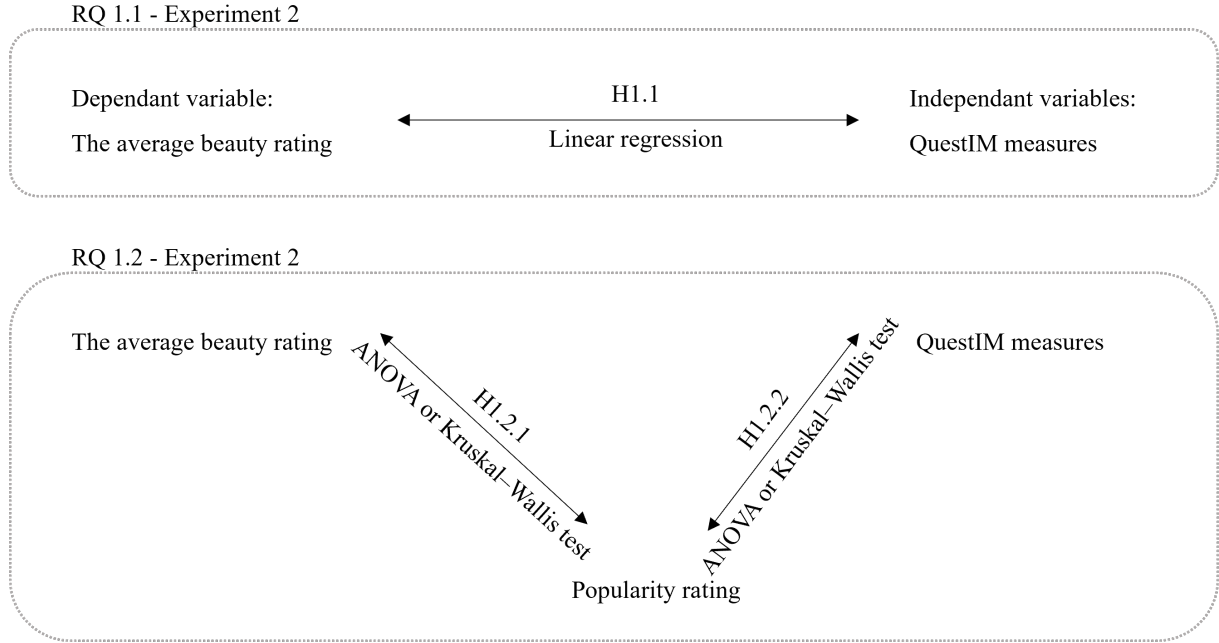


Figure 5.1: Visual summary of the inferential statistics (experiment 2)

## 5.2 Results and discussion

### 5.2.1 Descriptive statistics

The QUESTIM measures are represented by the box plots in figure 5.2 and summarised in the table 5.1. It should be noted that the observations made on these QUESTIM measures in this second experiment are similar to those made on the QUESTIM measures of the SimilarWeb shopping websites in the experiment presented earlier (see Section 4.2).

We observed that the means, the standard deviations and ranges are very different, meaning that the variables are volatile. Indeed, the means and standard deviations vary between 0.059 and 0.956 and between 0.024 and 0.269, respectively.

The symmetry, equilibrium and border density are the least dispersed variables, as already observed in the last experiment (see Section 4.2). They possess the lowest variances (lower than 0.002) and standard deviations (lower or equal to 0.4). The compression complexity and the external alignment are not dispersed either. On the contrary, the colourfulness and the vertical balance have the most spread distribution. Indeed, they own the highest variances (equal to 0.073 and 0.054, respectively) and standard deviations (equal to 0.269 and 0.232).

Furthermore, even if the horizontal balance does not possess a high minimum as in the last experiment (see Section 4.2), we can observe that the first quartile of the variable is high. 75% of the values of the horizontal balance are bigger than 0.8.

As regards the distribution of the variables, the Shapiro-Wilk normality tests show that each variable follows the normal distribution ( $p\text{-value} > 0.01$ ) except the border density, compression complexity, horizontal balance, equilibrium and external alignment ( $p\text{-value} < 0.01$ ) (see Appendix C.3).

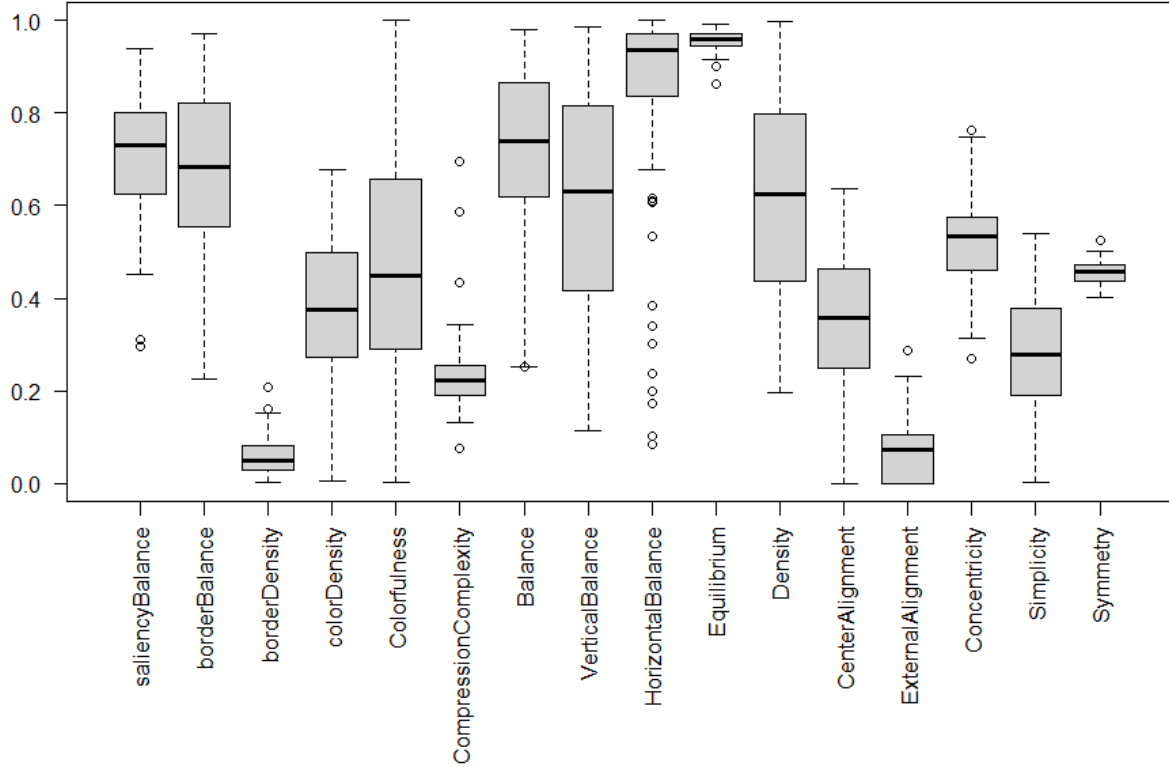


Figure 5.2: Box plot of the QUESTIM measures for Alexa websites

Figure 5.3 gathers the scatter plots of each pair of variables and contains on its right-hand side the Spearman correlations and, on the diagonal, histograms. We can notice that the simplicity and density have the strongest negative correlation which is equal to -0.93. Intuitively, it seems logical that a dense interface is not considered as simple as already observed and explained in the previous experiment (see Section 4.2). On the contrary, the balance and vertical balance are the most strongly and positively correlated variables (correlation=0.77). The balance and horizontal balance possess the second highest positive correlation (correlation=0.58). Indeed, the balance is a linear combination of the horizontal and vertical balance. The border density and the compression complexity have the third highest positive correlation (correlation=0.55).

| Measure               | Minimum | Maximum | Range | Mean  | SD    | Variance |
|-----------------------|---------|---------|-------|-------|-------|----------|
| SaliencyBalance       | 0.295   | 0.938   | 0.643 | 0.707 | 0.140 | 0.020    |
| BorderBalance         | 0.224   | 0.973   | 0.749 | 0.680 | 0.175 | 0.030    |
| BorderDensity         | 0.002   | 0.208   | 0.206 | 0.059 | 0.040 | 0.002    |
| ColorDensity          | 0.004   | 0.679   | 0.675 | 0.376 | 0.154 | 0.024    |
| Colorfulness          | 0.001   | 1.000   | 0.999 | 0.484 | 0.269 | 0.073    |
| CompressionComplexity | 0.076   | 0.695   | 0.619 | 0.234 | 0.088 | 0.008    |
| Balance               | 0.251   | 0.981   | 0.729 | 0.724 | 0.172 | 0.030    |
| VerticalBalance       | 0.115   | 0.987   | 0.872 | 0.616 | 0.232 | 0.054    |
| HorizontalBalance     | 0.085   | 1.000   | 0.915 | 0.832 | 0.242 | 0.059    |
| Equilibrium           | 0.862   | 0.993   | 0.131 | 0.956 | 0.024 | 0.001    |
| Density               | 0.196   | 0.997   | 0.801 | 0.614 | 0.215 | 0.046    |
| CenterAlignment       | 0.000   | 0.636   | 0.636 | 0.353 | 0.129 | 0.017    |
| ExternalAlignment     | 0.000   | 0.286   | 0.286 | 0.069 | 0.068 | 0.005    |
| Concentricity         | 0.268   | 0.763   | 0.494 | 0.527 | 0.104 | 0.011    |
| Simplicity            | 0.001   | 0.541   | 0.539 | 0.285 | 0.120 | 0.014    |
| Symmetry              | 0.403   | 0.524   | 0.121 | 0.457 | 0.024 | 0.001    |

Table 5.1: Descriptive statistics of the QUESTIM measures



Figure 5.3: Pair plot of the QUESTIM measures for Alexa websites

If we analyse the links between the QUESTIM measures and the average beauty rating (Table 5.2), we can observe that border density and simplicity are the most strongly correlated variables with the average beauty rating. In fact, the correlation between the border density and the beauty rating equals 0.25, whereas the correlation between the beauty variable and the simplicity is negative and equals -0.2.

| Measure               | Beauty |
|-----------------------|--------|
| SaliencyBalance       | 0.114  |
| BorderBalance         | -0.121 |
| BorderDensity         | 0.246  |
| ColorDensity          | 0.019  |
| Colorfulness          | -0.076 |
| CompressionComplexity | 0.025  |
| Balance               | 0.173  |
| VerticalBalance       | 0.127  |
| HorizontalBalance     | 0.167  |
| Equilibrium           | 0.086  |
| Density               | 0.080  |
| CenterAlignment       | -0.040 |
| ExternalAlignment     | 0.066  |
| Concentricity         | -0.164 |
| Simplicity            | -0.198 |
| Symmetry              | 0.123  |

Table 5.2: Correlations of the QUESTIM measures with the average beauty rating

## 5.2.2 Inferential statistics

Research Question 1.1: Do the QUESTIM measures satisfactorily predict the perceived aesthetics of electronic commerce websites on mobile devices?

Hypothesis 1.1: All QUESTIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.

Firstly, we address the research question 1.1 and its related hypothesis 1.1. We used linear regression to assess if the QUESTIM measures can satisfactorily predict the perceived aesthetics of e-commerce websites. A multiple linear regression was executed on RStudio to understand whether QUESTIM measures (explanatory independent variables) can satisfactorily explain and predict the beauty ratings given by the participants (dependent variable). However, collinearity and multicollinearity exist between the features and RStudio struggles to find some results of the linear regression. Indeed, links exist between the explanatory variables, which explains that the program cannot estimate the parameters and test the significance of the horizontal balance (see Appendix C.5). Moreover, multicollinearity affects the performance of the model.

In order to avoid collinearity and multicollinearity among the explanatory variables, we filtered some of them before realising the linear regression. The Variance Inflation Factor (VIF) could be used. The VIF is a factor associated with each variable, which indicates if the variable should be kept (VIF lower than 5) or not while realising the linear regression (Akinwande et al., 2015). Nevertheless, RStudio cannot compute the VIF as the QUESTIM measures are too collinear (see Appendix C.6). We therefore decided to observe manually the strongest correlations between the explanatory variables (Figure 5.3). The two strongest correlations are between the simplicity and density variables as well as between the balance and vertical balance variables. Firstly, we removed the density because it has a lower correlation with the average beauty rating compared to the simplicity (see Table 5.2). Secondly, we discarded the balance which is a function of the horizontal and vertical balance and has a low correlation with the score (see Table 5.2).

After having filtered the variables, multiple linear regression was performed using the 14 remaining QUESTIM measures. To select the features which possess the highest explanatory power, the AIC (Akaike Information Criterion) stepwise regression can be executed. This method consists in adding or subtracting model variables to minimise the AIC. The Akaike Information Criterion (AIC) is proportional to the quantity of information lost in the model and inversely proportional to the simplicity of the model determined by the number of features (Njike and Segers, 2017).

As some non-significant variables were still included in the model obtained after this stepwise method (see Appendix C.8), we discarded one by one the least significant features until the model was only composed of significant variables ( $p\text{-value} < 0.05$ ). This means that these remaining variables contribute to explain the average beauty rating and possess a coefficient different from zero. We obtained that the average beauty rating can be predicted by one feature, border density, thanks to the following model (equation 5.1, see Appendix C.9). The resulting model is not surprising as the border density is the QUESTIM variable most correlated with the average beauty rating (Table 5.2).

$$\text{AverageBeautyRating} = 2.67 + 4.39 \times \text{borderDensity} \quad (5.1)$$

The coefficient of determination, noted  $R^2$ , is the variance proportion of the average beauty rating that is explained by the model (Wackerly et al., 2014). This coefficient is equal to 8.61% (see Appendix C.9) for the previous model (equation 5.1). To conclude, we reject the hypothesis 1.1 because only one QUESTIM measure contributes to predict the perceived aesthetics of electronic commerce websites. Furthermore, the model does not satisfactorily predict the average beauty rating.

Research Question 1.2: Are popular electronic commerce websites perceived as aesthetic on mobile devices?

Secondly, we explored the behaviour of popular sites in relation to others. To do this, we compared the sites from the first half of the Alexa ranking with those from the second half. More specifically, we compared the average of a variable in the first half of the ranked sites with the average of this same variable in the second half of the ranked sites. To check whether two means come from the same population, the ANOVA test is used for normally-distributed variables and the Kruskal-Wallis test for the others (Wackerly et al., 2014). The null hypotheses state that the means of a variable are equal among the most popular and the least popular websites. We also need, for the ANOVA test, to verify that the two populations have the same variance. This condition has to be explicit on RStudio when carrying out the test. We assume that the highest popular sites are those with the most traffic according to the Alexa ranking.

Hypothesis 1.2.1: The popular electronic commerce websites are perceived by users as aesthetic on mobile devices.

We began by analysing the average beauty rating according to the popularity ranks. Visually, we can already notice that there seems to be no significant difference in evaluations between popular and non-popular websites (Figure 5.4). As the variable follows a normal distribution among the two populations (first half of the Alexa ranking sites and the second one), we performed an ANOVA test after verifying that the variances of the two populations were equal. This test allows us to state that there is no significant difference between the average ratings of popular and less popular websites ( $p\text{-value}=0.352$ , see Appendix C.10). Therefore, the hypothesis 1.2.1 can be rejected.

In an attempt to go further in the analysis, we decided to compare the average ratings for the 10 most popular and 10 least popular websites. The two populations composed of the most and least popular websites follow a normal distribution. Therefore, an ANOVA test was performed. The test shows that, once again, the means of the two populations were not significantly different ( $p\text{-value}=0.568$ ). As the results of this deeper analysis are not more promising than those of the first comparison, we decided not to investigate it further.

Hypothesis 1.2.2: The popular electronic commerce websites are aesthetic (under QUESTIM measures) on mobile devices.

In order to confirm or reject this hypothesis, we compared the behaviour of each QUESTIM measure separately among the most and the least popular websites. Depending on the distribution of the two populations (first half of the Alexa ranking and the second half) and the verification of the equal variances assumption, for normally-distributed

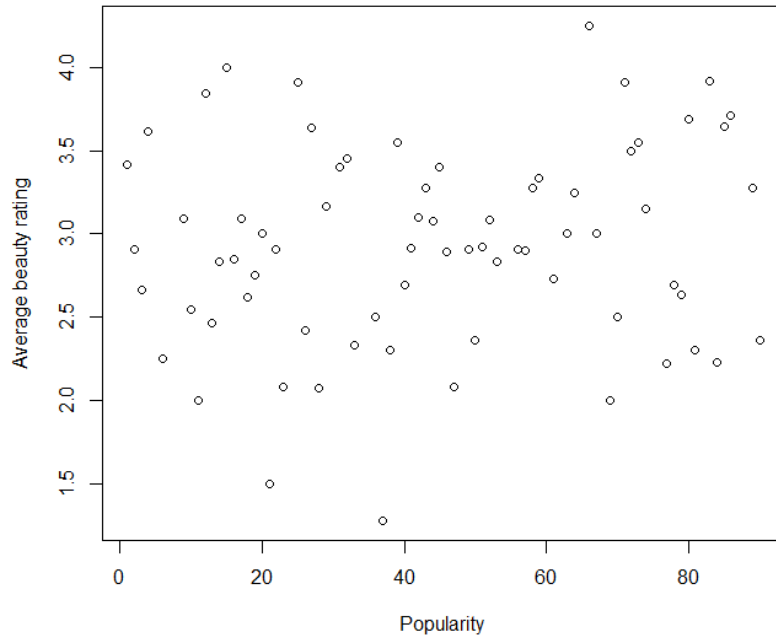


Figure 5.4: Scatter plot of the average beauty rating per popularity rank

populations, we performed means comparison tests. All p-values of the tests are between 0.062 and 0.981 (see Appendix C.10). This leads us to state for each measure that the average of a variable within the most popular sites is not significantly different from the average of the same variable within the least popular sites. As a result, we reject the hypothesis 1.2.2 because it cannot be affirmed that popular websites are more aesthetic.

## Chapter 6

# Experiment 3 - SimilarWeb / AIM / Ergonomic criteria

To address the research question “Which are the measures to assess the perceived aesthetics of a website in general and of electronic commerce websites in particular, on mobile devices?”, we will explore among other things two sources of evaluation: the Alto Interface Metrics (AIM)(an automatic set of measures), and the ergonomic criteria (a manual set of criteria). Thanks to UiLAB, AIM measures, ergonomic criteria, and beauty and complexity ratings will be collected for all websites categories listed in SimilarWeb. Indeed, these measures and criteria will be confronted with human judgement.

It is noteworthy that, apart from the study of aesthetics, we will explore interface complexity. Complexity is a rather complex and interesting concept. Indeed, it can be seen as a component of aesthetics and it also encompasses other underlying concepts. In the literature, it has been demonstrated that visual complexity can predict aesthetics ratings (Miniukovich and De Angeli, 2014b).

In addition, the difference between the website categories will be investigated since, as previously discussed, it has been argued in the literature that the visual design of sites varies depending on the category (Ivory et al., 2001). In this experiment, we will compare the behaviour of the measures according to the websites category. Specifically, we will look at e-commerce and shopping sites, on the one hand, compared to the other categories combined, on the other hand. To verify the previous statements and study the links between the variables, the research question can be divided into subquestions with associated hypotheses.

- Research Question 1.3: Do the AIM measures and the ergonomic criteria satisfactorily predict the perceived aesthetics and complexity of electronic commerce websites on mobile devices?

- Hypothesis 1.3.1: All AIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.
- Hypothesis 1.3.2: All AIM measures predict the perceived complexity of electronic commerce websites on mobile devices.
- Hypothesis 1.3.3: All ergonomic criteria predict the perceived aesthetics of electronic commerce websites on mobile devices.
- Hypothesis 1.3.4: All ergonomic criteria predict the perceived complexity of electronic commerce websites on mobile devices.
- Research Question 1.4: Compared to other sites, how are the complexity, the aesthetics and the ergonomics of the electronic commerce websites perceived and evaluated on mobile devices?
  - Hypothesis 1.4.1: Electronic commerce websites are perceived as aesthetic by users compared with the other websites on mobile devices.
  - Hypothesis 1.4.2: Electronic commerce websites are not perceived as complex by users compared with the other websites on mobile devices.
  - Hypothesis 1.4.3: The electronic commerce websites are more aesthetic (under AIM) than other websites on mobile devices.
  - Hypothesis 1.4.4: The electronic commerce websites are more ergonomic (under ergonomic criteria) than other websites on mobile devices.

## 6.1 Experimental method

In the third experiment, the research method started with the screenshot collection at the beginning of 2021. The URL and a homepage screenshot were extracted from the 50 most popular websites of each SimilarWeb category using UILAB. Thus, 50 homepages were captured in each of the 22 categories in a 414x712 resolution and saved in a PNG file format. The screenshot gallery was filtered. Indeed, unavailable websites were eliminated and website homepages hidden by an “accept cookies” window, for example, were manually captured. The final gallery consists of 1,059 websites distributed in 22 different website categories, which each contains between 43 and 50 screenshots. The zone segmentation, that constitutes the second step, was carried out thanks to an Aalto Interface Metrics (AIM) script implemented in UILAB. AIM aims to evaluate the design of Graphical User Interfaces and is composed of a service accessible online<sup>1</sup> and an open source code (Oulasvirta et al., 2018). This segmentation script automatically separates the screen

---

<sup>1</sup>See <https://interfacemetrics.aalto.fi/>

into different zones defined by an identifier,  $x$  and  $y$  positions, width, height and an image encoding data.

On UiLAB, the 53 AIM measures of each of the 1,059 screenshots were automatically computed based on the AIM script. More specifically, AIM presents 17 measures which together contain 53 submeasures in total (Oulasvirta et al., 2018). Here is a short presentation of those measures:

- The colour harmony is related to the HSV colour wheel and is composed of the distances to colour schemes (Cohen-Or et al., 2006).
- The colourfulness is defined by Hasler and Susstrunk (2003) using the different axes of the RGYB colour space.
- The HSV average contains the colour information represented in the Hue Saturation Value colour space (Hasler and Susstrunk, 2003; Miniukovich and De Angeli, 2014a).
- The LAB average is also related to colour information and is represented in LAB space (where L refers to lightness, A to the red and green axis, and B to the blue and yellow axis) (Hasler and Susstrunk, 2003).
- The standard deviation of the luminance is the variability of pixel luminance. (Miniukovich and De Angeli, 2014a).
- The number of colours contains the number of unique colours present (Miniukovich and De Angeli, 2014a; Miniukovich and De Angeli, 2015).
- The PNG (Portable Network Graphics) file size is expressed in bytes and is linked to the colours of the image (Miniukovich and De Angeli, 2014a).
- The wave score (Weighted Affective Valence Estimates) is the average of the colour preference scores of the image pixels based on the research of Palmer and Schloss (2010).
- The edge congestion is the proportion of congested pixels located on the edges (Oulasvirta et al., 2018; Rosenholtz et al., 2007).
- The edge density is the proportion of pixels located on the edges (Oulasvirta et al., 2018; Rosenholtz et al., 2007).
- The figure and ground contrast quantifies the contrast between the backgrounds, objects, and words based, in particular, on the colours (Oulasvirta et al., 2018; Miniukovich and De Angeli, 2014a; Reber et al., 1998).
- The number of alignment lines is a measure of grid quality and is related to the number of alignment points of the visual blocks that make up the image (Miniukovich and De Angeli, 2015).

- The JPEG (Joint Photographic Experts Group) file size describes the size of the file and is linked to the clutter of the image (Miniukovich and De Angeli, 2014a).
- The pixel symmetry refers to the symmetry across an axis (Miniukovich and De Angeli, 2014a).
- The quadtree measures are obtained thanks to the quadtree decomposition algorithm which splits an image into quadrants until those quadrants reach a certain level of entropy (Reinecke et al., 2013; Zheng et al., 2009).
- The white space is related to the spaces left without content on the screen (Oulasvirta et al., 2018; Miniukovich and De Angeli, 2015).

Eight ergonomic criteria were collected for each screenshot of the gallery. They were evaluated by the student of the human-computer interface course delivered by Jean Vanderdonckt (Vanderdonckt, 2021) after receiving explanations on the criteria. The ergonomic criteria are originally based on the work of Bastien and Scapin (Bastien and Scapin, 1993) and have been subject to validation (Bastien and Scapin, 1995; Scapin et al., 2001). Here is a short explanation of the ergonomic criteria used in our study and ranked in order of importance (Vanderdonckt, 1995; Vanderdonckt, 2021):

- The compatibility informs us whether the information and tasks are linked to the one of the real word.
- The consistency refers to the ease of identification and usage of information and actions.
- The workload refers to the quantity of information to deal with and the steps needed to perform a task on the website.
- The adaptation is linked to the possibility of adjustment of the interface to the user and his context.
- The dialogue control refers to the feeling of control that the user might have.
- The representativeness is related to the contents representative nature of some element of the interface such as labels and menu items.
- The guidance is linked to the information available that allow the user to be guided and to know his position in the task completion.
- The error management refers to the capacity of dealing and fixing errors done by the user.

Then, we collected online ratings for each website. Each participant was assigned a screenshot category. The experiment took place on UILAB where random screenshots of a category were displayed one after the other. On the bottom of the page, two rating

scales of seven stars were used to assess the beauty (1 = the least beautiful and 7 = the most beautiful) and the complexity (1 = the least complex and 7 = the most complex) of the site presented above.

Finally, the descriptive statistics of the different variables were completed before implementing inferential statistics to answer the research questions. The analyses were performed using mainly RStudio whereas some descriptive statistics data and the normality tests were performed on GraphPad Prism as well. The figure 6.1 depicts the different variables and the statistical tools deployed to address the research questions.

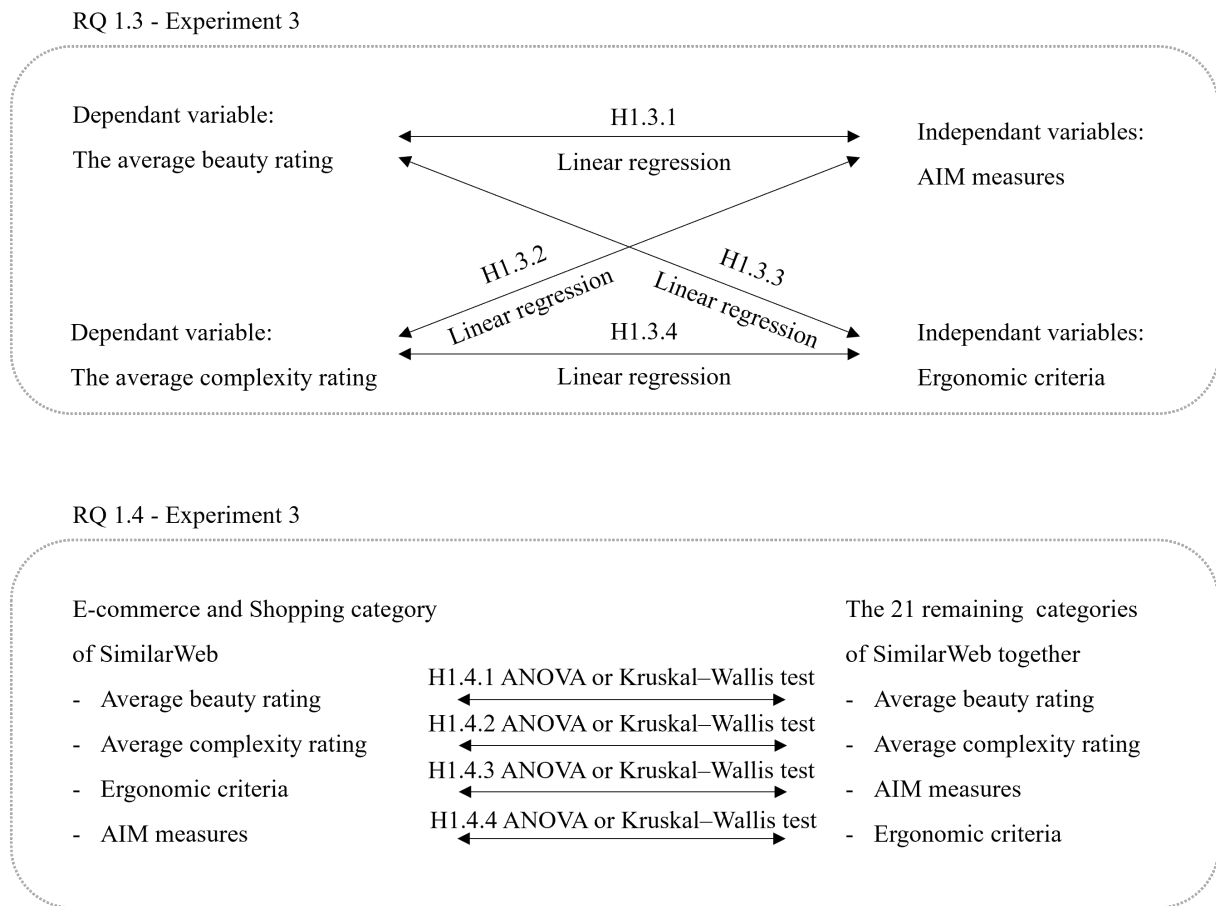


Figure 6.1: Visual summary of the inferential statistics (experiment 3)

Different kinds of variables were used during this experiment. As mentioned previously, the screenshot is still a nominal variable which is linked to a website and a SimilarWeb category (which represents a categorical variable). The AIM measures are numerical quantitative variables computed automatically thanks to UiLAB. The ergonomic criteria are numerical and scale variables between 1 and 7. The beauty and the complexity average ratings are also numerical and scale variables (between 1 and 7). They represent, respectively, the average of the beauty and complexity ratings left for one screenshot by a sample of participants.

## 6.2 Results and discussion

The statistical analysis was carried out on a data set containing 1,059 observations which represent the 1,059 homepage screenshots of the 22 categories of SimilarWeb. As explained previously, the database contains, for each website, its associated category, 53 AIM measures, eight ergonomic criteria and the average beauty and complexity rating. As one of the website is not associated with ergonomic criteria, we discarded this screenshot which was included in the “Business and Consumer Service” category. Our analyses were performed on 49 websites of the “E-commerce and Shopping” category. Nevertheless, the full database was used to answer the research question related to the comparison of the e-commerce and shopping category with the remaining categories of SimilarWeb (research question 1.4).

### 6.2.1 Descriptive statistics

The ergonomic criteria of the e-commerce and shopping category are included between 1 and 7 and their box plots are represented in figure 6.2. It can be quickly noticed that the error management variable is only composed of values equal to 4. Consequently, its standard deviation and variance are equal to 0 (Table 6.1). The error management is difficult to assess based on a screenshot because this dimension needs more dynamic interactions to evaluate its quality. We decided to set aside this variable in the analyses carried out on the e-commerce and shopping category. Indeed, the variable contains little information of interest for this specific category. As far as the other ergonomic criteria are concerned, they are not very volatile. Their means and standard deviations are included between 4.000 and 5.122 and 0.713 and 1.309, respectively. Moreover, thanks to the Shapiro-Wilk normality test (see Appendix D.3), it can be shown that the seven variables do not follow a normal distribution (p-values<0.01).

| Criterion          | Min. | Max. | Range | Mean  | Std. Deviation | Variance |
|--------------------|------|------|-------|-------|----------------|----------|
| Compatibility      | 2    | 7    | 5     | 4.469 | 1.043          | 1.088    |
| Consistency        | 2    | 6    | 4     | 4.714 | 0.890          | 0.792    |
| Workload           | 1    | 7    | 6     | 4.51  | 1.309          | 1.713    |
| Adaptation         | 3    | 7    | 4     | 4.694 | 0.713          | 0.509    |
| Dialog.Control     | 1    | 7    | 6     | 4.714 | 1.118          | 1.250    |
| Representativeness | 3    | 6    | 3     | 4.837 | 1.007          | 1.014    |
| Guidance           | 1    | 6    | 5     | 5.122 | 1.053          | 1.110    |
| Error Management   | 4    | 4    | 0     | 4     | 0              | 0        |

Table 6.1: Descriptive statistics of the ergonomic criteria

The figure 6.3 represents a scatter plot on its lower part, Spearman correlations on its upper part between each pair of eight ergonomic criteria, and the histograms of each

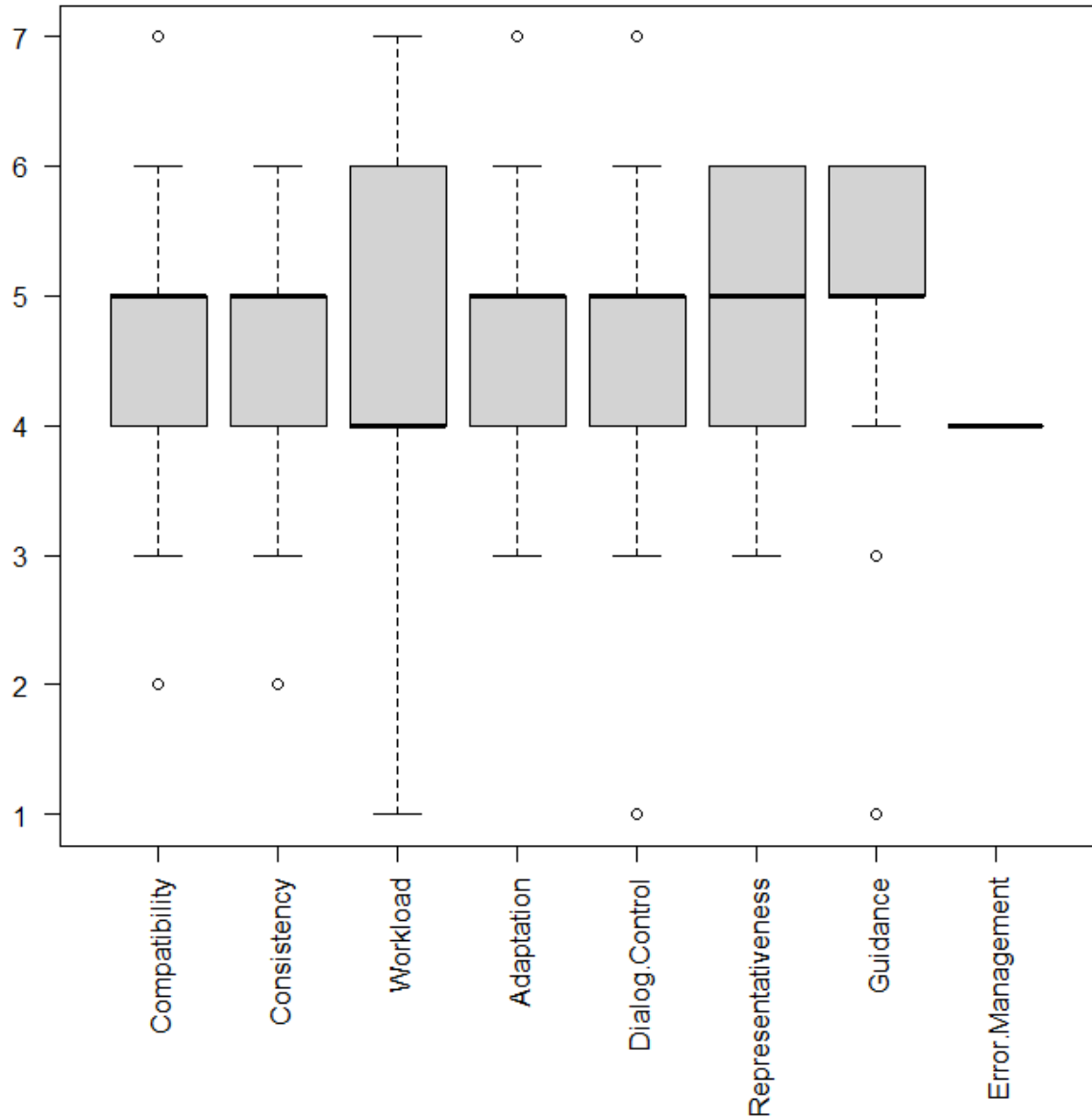


Figure 6.2: Box plot of the ergonomic criteria

variable on its diagonal. We noted that a lot of variables possess a positive correlation. The representativeness and the guidance are the most strongly correlated variables (correlation = 0.77). The consistency and the workload have also a strong correlation which is equal to 0.72.

As regards the links between the ergonomic criteria and the average ratings, we observed that some ergonomic criteria have a stronger link than others with the average beauty rating variable. Indeed, dialogue control and representativeness have, respectively, a correlation higher than 0.43 with the beauty rating variable. On the contrary, the ergonomic criteria do not seem to be highly correlated with the complexity average rating variable. The adaptation variable is the one with the highest correlation with complexity

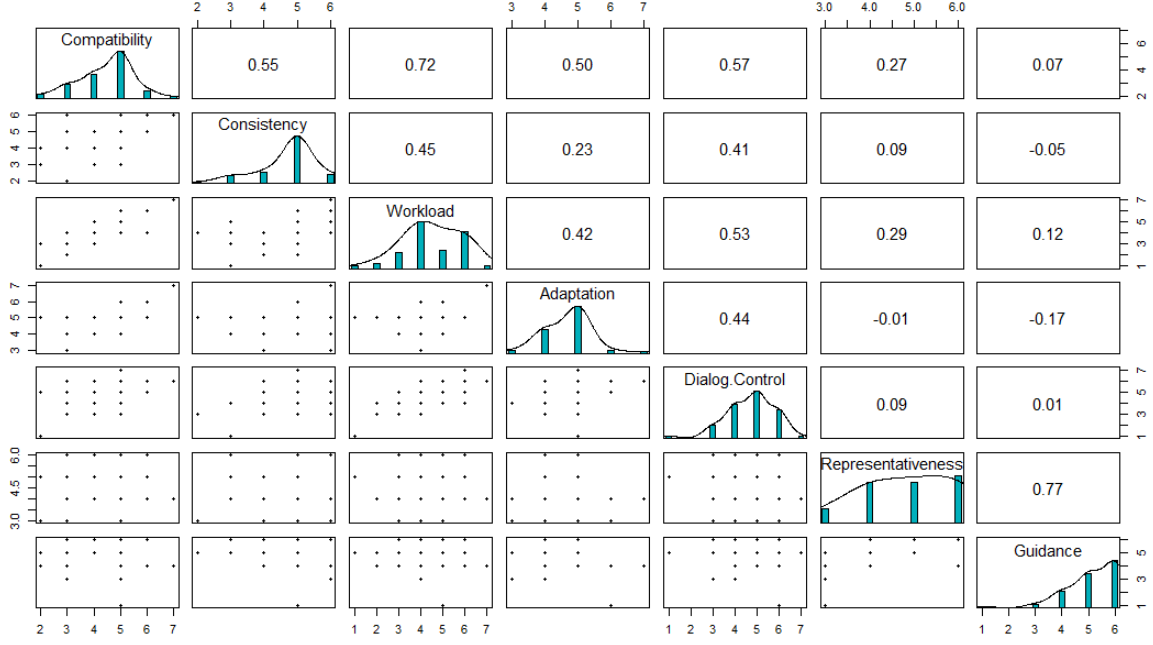


Figure 6.3: Pair plot of the ergonomic criteria

(correlation = -0.19) (Table 6.2).

| Criterion          | Beauty    | Complexity |
|--------------------|-----------|------------|
| Compatibility      | 0.2961939 | -0.0686303 |
| Consistency        | 0.2374147 | -0.0636639 |
| Workload           | 0.3107281 | -0.0650106 |
| Adaptation         | 0.1823291 | -0.1872986 |
| Dialog.Control     | 0.4398562 | 0.08882224 |
| Representativeness | 0.4790243 | 0.1522824  |
| Guidance           | 0.2916105 | 0.1271773  |

Table 6.2: Correlations of the ergonomic criteria with the average beauty and complexity rating

As far as the AIM measures of the e-commerce and the shopping category are concerned, we noticed that they have different units of measurement and distributions. For instance, the `color_harmony_i_distance` and the `HSVAvg_average_sat` variables have a range of 1,365,958 and 0.5771, respectively (see Appendix D.5). The Shapiro-Wilk tests show that 17 variables among 53 have a normal distribution ( $p\text{-values} > 0.001$ ) (see Appendix D.6). To evaluate the link between each pair of AIM variables, Spearman correlations were computed and revealed that some variables are strongly correlated. Indeed, the highest positive correlation between the e-commerce and shopping websites category is equal to 1 (`quadtrees_nb_leaves` and `quadtrees_equilibrium`) and the highest

negative one is -0.888 (HSVAvg\_average\_value and HSVAvg\_value\_stdev).

## 6.2.2 Inferential statistics

Research Question 1.3: Do the AIM measures and the ergonomic criteria satisfactorily predict the perceived aesthetics and complexity of electronic commerce websites on mobile devices?

We began by addressing the research question 1.3. To evaluate if the AIM measures and ergonomic criteria can predict the perceived aesthetics and complexity of the e-commerce websites, we performed linear regressions through different hypotheses.

Hypothesis 1.3.1: All AIM measures predict the perceived aesthetics of electronic commerce websites on mobile devices.

First, we predicted the average beauty rating (dependent variable) thanks to the AIM measures (independent variables), to verify the hypothesis 1.3.1. As the AIM measures are highly correlated (see Section 6.2.1), the linear regression cannot be computed correctly. RStudio is not able to estimate some parameters and to test the significance of the explanatory variables (see Appendix D.7). Indeed, collinearity and multicollinearity exist between the variable, which means that there are links between the features.

Given that multicollinearity affects the performance of the model, it can be interesting to filter the variables before performing a linear regression. One method could be to use the VIF as presented in section 5.2.2. Nevertheless, the AIM variables are too dependent on each other and the VIF cannot be computed on RStudio (see Appendix D.8).

In order to avoid multicollinearity between the variable, we created an algorithm. This algorithm selects, among the variables that are too closely related, those that are most correlated with the variable that we are trying to explain. Indeed, these variables possess a greater predictive power. More specifically, when two variables have a correlation superior to 0.8 or inferior to -0.8, the algorithm keeps the variable which is more correlated to the variable that we are trying to explain. For instance, color\_harmony\_L\_distance and color\_harmony\_i\_distance have a correlation equals to 0.86. The algorithm sets aside color\_harmony\_i\_distance because that variable is less correlated with the average beauty rating than color\_harmony\_L\_distance. It is noteworthy that the threshold of 0.8 was arbitrarily chosen because it allows to keep more or less 50% of the variables (see Appendix D.9). Knowing that the variables are, by definition, highly correlated, this threshold allows us to have a trade-off between filtering too many variables (20 remaining variables for a threshold of 0.7) and not enough variables (40 variables for a threshold of 0.9).

To predict or explain the average beauty rating using the 27 unfiltered AIM measures, we computed a linear regression. We used the stepwise regression to choose the features which possess the highest explanatory power (Njike and Segers, 2017). The output function of the stepwise regression is composed of 14 independent variables (see Appendix D.10). Then, we checked the significance of each variable one by one and removed the least significant one at a time until each of the remaining variables was significant (p-value<0.05). The final model is the following (equation 6.1, see Appendix D.11):

$$\begin{aligned}
AverageBeautyRating = & 6.497 - 0.00001 \times color\_harmony\_T\_distance \\
& + 0.002 \times color\_harmony\_L\_alpha \\
& - 0.002 \times color\_harmony\_T\_alpha \\
& - 3.131 \times HSVAvg\_average\_value \\
& + 0.016 \times LABAvg\_B\_mean \\
& + 0.00001 \times nb\_colors \\
& - 0.000009 \times jpeg\_file\_size \\
& + 1.706 \times quadtree\_symmetry
\end{aligned} \tag{6.1}$$

Of the eight variables, four have a positive coefficient, meaning that they have a positive impact on the average beauty rating. The three colour harmony variables (*color\_harmony\_T\_distance*, *color\_harmony\_L\_alpha*, *color\_harmony\_T\_alpha*), the HSV and the LAB variables can be grouped under the concept of colourfulness to which they relate, whereas the last 3 variables of the equation 6.1 (*nb\_colors*, *jpeg\_file\_size*, *quadtree\_symmetry*) are more linked to the complexity concept (Miniukovich and De Angeli, 2014a; Reinecke et al., 2013). Indeed, the more unique colours in the image, the more complex the image. Also, the more complex the image and the more information it contains, the larger the jpeg file size. Finally, the quadtree algorithm breaks down the image into low-entropy areas where the colour is more or less constant. The more unbalanced the image decomposition through the quadtree algorithm, the more asymmetric the decomposition tree will be.

According to the coefficient of determination,  $R^2$ , 68% of the variability of the average beauty rating is explained by the model, which is composed of eight explanatory variables (see Appendix D.11). Therefore, we can reject the hypothesis 1.3.1. Indeed, only eight AIM measures can satisfactorily predict the perceived aesthetics of electronic commerce websites on smartphones.

Hypothesis 1.3.2: All AIM measures predict the perceived complexity of electronic commerce websites on mobile devices.

Through this hypothesis, we explored the predictive power of the AIM measures (independent variables) regarding the complexity average rating (dependent variable). Unfortunately, we faced the same issues as presented above for the prediction of the average beauty rating thanks to the AIM measures (see hypothesis 1.3.1 in Section 6.2.1). Indeed, computing a linear regression and filtering the variables with the VIF is not possible because of the high collinearity between the variables.

To avoid collinearity between the features, we decided, as was also done previously, to deploy a filtering algorithm. This one examines the correlations between the AIM measures higher than 0.8 or lower than -0.8. This time, the algorithm keeps the most correlated variables with the complexity rating. This technique makes it possible to obtain 25 variables that are not too closely related to each other (see Appendix D.12). The threshold of 0.8 was again chosen arbitrarily as it allowed to keep 50% of the variables as opposed to the 0.7 and 0.9 threshold which eliminates, respectively, 33 and 14 variables among the 53 AIM measures.

After having filtered the measures, we used the AIC stepwise regression to select the features of the linear regression. The linear regression obtained possesses seven independent variables (see Appendix D.13). Among the features, some variables are still not significant. We removed the non-significant variables until reaching a model with only significant variables (p-value<0.05) (equation 6.2, see Appendix D.14).

$$\begin{aligned} \text{AverageComplexityRating} = & 3.390 + 0.001 \times \text{color\_harmony\_V\_alpha} \\ & - 0.0009 \times \text{color\_harmony\_T\_alpha} \\ & + 0.003 \times \text{HasslerSusstrunk\_RG\_stdev} \end{aligned} \quad (6.2)$$

The model indicates that the complexity can be predicted with the color\_harmony\_V\_alpha, the color\_harmony\_T\_alpha and the HasslerSusstrunk\_RG\_stdev. The color\_harmony\_T\_alpha is the only one which possesses a negative coefficient. This means that when the variable increases, the average complexity rating left by the participant decreases.

The hypothesis 1.3.2 can be rejected as only a small part of the AIM measures predict the perceived complexity. It is noteworthy that the model is not very efficient anyway as it only predicts 32% of the variability of the average complexity rating.

Hypothesis 1.3.3: All ergonomic criteria predict the perceived aesthetics of electronic commerce websites on mobile devices.

In order to address the research question 1.3 and the related hypothesis 1.3.3, we performed a linear regression using the average beauty rating (dependent variable) and the ergonomic criteria (independent variables) except for the error management variable which is constant. As previously explained (see Section 6.2.1), the ergonomic criteria are not too correlated and the VIF does not suggest eliminating certain variables (see Appendix D.15). Therefore, we performed a stepwise regression to select from the 7 explanatory variables those with the greatest predictive power (see Appendix D.16).

The resulting model contains two independent variables that are both significant ( $p\text{-value} < 0.05$ ) (equation 6.3). The two variables positively influence the perceived aesthetics. This means that the more control the user has over the interface, the more beautiful it will be perceived. Similarly, the more representative the interface elements (e.g. codes and labels) are of its content, the more attractive it will be to users. These results seem to be coherent with what we observed previously (Table 6.2). Indeed, dialogue control and representativeness possess the highest correlations with the beauty variable.

$$\begin{aligned} \text{AverageBeautyRating} = & 1.472 + 0.181 \times \text{Dialog.Control} \\ & + 0.245 \times \text{Representativeness} \end{aligned} \quad (6.3)$$

The hypothesis 1.3.3 can be rejected as only two ergonomic criteria predict the perceived aesthetics. Besides, the two explanatory variables do not possess a high predictive power as they explain only 34% of the variability of average beauty rating.

Hypothesis 1.3.4: All ergonomic criteria predict the perceived complexity of electronic commerce websites on mobile devices.

To test the last hypothesis of question 1.3, we also used a linear regression to predict the average complexity rating (dependent variable). This was done with all ergonomic criteria (independent variables) except for the error management variable as the variable is constant. Indeed, the criteria are not too correlated and the VIF does not advise us to put some variables aside (see Appendix D.17). The stepwise regression was deployed and resulted in a model with a single explanatory variable: the adaptation (see Appendix D.18). However, this variable is not significant and the final model contains only the intercept. This means that the ergonomic criteria do not predict the perceived complexity which leads to reject hypothesis 1.3.4. Given the observed correlations (Table 6.2), we are not surprised by this outcome. Indeed, the ergonomic criteria are weakly correlated with the complexity variable. The correlations in absolute values do not exceed 0.19.

Research Question 1.4: Compared to other sites, how are the complexity, the aesthetics and the ergonomics of the electronic commerce websites perceived and evaluated on mobile devices?

To tackle the research question 1.4, we tried to compare the variables of the e-commerce and shopping category of SimilarWeb with their equivalent for all the other categories of the database. We therefore compared the average of a variable in the category of interest (e-commerce and shopping) with the average of this same variable in the 21 other categories combined. As explained previously (See Section 5.2.2), the ANOVA test and the Kruskal-Wallis test can be used to verify if two averages are from the same population. The ANOVA test is applicable for the normally-distributed variables in the two populations. On the contrary, the Kruskal-Wallis test is suitable for the variables which do not possess a normal distribution. Through the hypotheses 1.4.1 to 1.4.4, we evaluated the behaviour of four types of variables separately: the average beauty and complexity rating, the AIM measures and the ergonomic criteria.

Hypothesis 1.4.1: Electronic commerce websites are perceived as aesthetic by users compared with the other websites on mobile devices.

To reject or confirm this hypothesis, we explored the average beauty rating which follows a normal distribution among the populations of the e-commerce and shopping websites and among the remaining categories. Thanks to the ANOVA test (with two equal variances), we can assert that there is not a statistically significant difference in beauty means between the e-commerce category and the remaining categories (see Appendix D.19). Indeed, the null hypothesis, which stands for the equality of the means, can not be rejected (p-value=0.361). Therefore, the e-commerce websites are not perceived as more aesthetic than others. This means that the hypothesis 1.4.1 cannot be confirmed.

Hypothesis 1.4.2: Electronic commerce websites are not perceived as complex by users compared with the other websites on mobile devices.

Then, we looked at the behaviour of the complexity variable among the different populations. This variable is only normally-distributed within the category of e-commerce and shopping websites. Therefore, we used the Kruskal-Wallis test. This test indicates that there is not a statistically significant difference in complexity means between the e-commerce and shopping category and the remaining categories (p-value=0.550) (see Appendix D.19). As a consequence, the hypothesis 1.4.2 can be rejected.

Hypothesis 1.4.3: The electronic commerce websites are more aesthetic (under AIM) than other websites on mobile devices.

Instead of studying 53 AIM measures in the category of interest (e-commerce and shopping) and in the 21 other categories combined, we decided to focus on a smaller

amount of variables. We selected the explanatory variables of the most promising model, which is the final linear regression of beauty on the AIM measures (equation 6.1): color\_harmony\_T\_distance, color\_harmony\_L\_alpha, color\_harmony\_T\_alpha, HSV-Avg\_average\_value, LABAvg\_B\_mean, nb\_colors, jpeg\_file\_size and quadtree\_symmetry. These explanatory variables are relatively successful in predicting the variance of beauty.

The Kruskal-Wallis tests show that HSVAvg\_average\_value is the only variable which possesses a statistically significant difference in means between the two populations studied (p-value=0.01). The value of the Hue Saturation Value (HSV) colour space is, on average, greater by 0.07 for the e-commerce and shopping websites (see Appendix D.20). This means that the colours can be seen as more intense for this specific category compared to the other categories combined. Due to these observations, we reject the hypothesis 1.4.3. Indeed, among the 8 variables tested, only one has a different distribution for e-commerce websites.

Hypothesis 1.4.4: The electronic commerce websites are more ergonomic (under ergonomic criteria) than other websites on mobile devices.

Finally, we focused on the behaviour of each ergonomic criterion separately among the two populations (e-commerce and shopping category and the remaining categories combined). Thanks to the Kruskal-Wallis tests, we observed that the null hypotheses, which stand for the equality of the means, are rejected for three variables (p-values<0.003). Indeed, there is a difference of adaptation, dialogue control and guidance means between the e-commerce and shopping category and the remaining categories (see Appendix D.19). More specifically, there is more or less a one-point difference between the guidance of the two populations. The guidance on e-commerce and shopping sites is on average evaluated at 5.12 while on other sites this same variable is equivalent to 4.16 (see Appendix D.20). It therefore seems that e-commerce sites guide the user more than other websites.

Concerning adaptation and dialogue control, we noticed that a difference of 0.67 and 0.62, respectively, exists between the means of the two populations. The adaptation and dialogue control are evaluated slightly more positively for e-commerce websites than for the other categories combined (see Appendix D.20). In other words, the user feels more in control of the interface. The websites also seem more adapted to the user and his context. However, we cannot stand that all ergonomic criteria possess a different distribution between the e-commerce category and the other categories. Therefore, we reject the hypothesis 1.4.4.

# Chapter 7

## Conclusion and perspectives

In this chapter, we will present our contributions with regards to our objective (see Section 1.2.1) and our research questions (see Section 1.2.2). Then, we will detail the benefits and the limits of our thesis. Finally, we will discuss the potential guidance for future research in the short and long-term.

### 7.1 Summary of the contributions

Through this work, we have addressed the literature gap that exists as regards the aesthetics measures of Mobile User Interfaces thanks to our research question.

**Which are the measures to assess the perceived aesthetics of a website in general and of electronic commerce websites in particular, on mobile devices?**

With the help of the Systematic Literature Review method, we identified and collected aesthetics and visual complexity measures with their related formulas to evaluate the visual design of mobile interfaces. Nevertheless, we obtained a low number of relevant results, despite the precision and the number of queries adopted. In our view, there are two possible explanations for this result. Firstly, relevant articles corresponding to our research question might not be indexed with the exact terms used in our queries. Secondly, it is possible that very few articles related to our subject and meeting our criteria are published. Indeed, we found that a surprising number of articles evaluate the visual design of mobile interfaces without using formulas that describe the measures employed. Moreover, this research topic is very recent. For example, the two most relevant articles of our Systematic Literature Review were published in 2018 and 2021.

In our family of experiments, we used different measures to assess the perceived aesthetics and complexity of e-commerce websites on mobile devices. In addition, to find automated and systematic measures to assess aesthetics, it is also essential to collect

measures that reflect human judgement.

In the first experiment, we analysed the aesthetics measures computed semi-automatically thanks to QUESTIM on the most popular web sites of SimilarWeb. In the second experiment, we tried to predict the aesthetics evaluation of users thanks to the measures collected on QUESTIM for the most popular websites of the Alexa ranking. We developed models predicting the evaluations left by the users. The resulting model, which is composed of one explanatory variable (border density), explains 8.61% of the variability of the average beauty rating (equation 5.1 related to hypothesis 1.1). We conclude that the measures proposed by QUESTIM are not efficient to predict the beauty perceived by users on mobile devices. In the second stage, the average of each QUESTIM measure (taken separately) and the perceived beauty, among the more popular websites, were compared with their equivalent among the less popular sites. We could not demonstrate that popular websites are significantly more aesthetically pleasing than less popular ones.

In the third experiment, we tried to predict the aesthetics and complexity ratings, separately, using AIM aesthetics measures and ergonomic criteria. The data were collected on the most popular websites of 22 different SimilarWeb categories.

First of all, we predicted the average beauty rating of e-commerce and shopping websites thanks to AIM measures. The model, composed of eight explanatory variables, possesses a coefficient of determination equals to 68% (equation 6.1 related to hypothesis 1.3.1). This means that the predicting model captures 68% of the beauty rating variability. The eight significant explanatory variables could be summarised under the colourfulness and complexity concepts: `color_harmony_T_distance`, `color_harmony_L_alpha`, `color_harmony_T_alpha`, `HSVavg_average_value`, `LABavg_B_mean`, `nb_colors`, `jpeg_file_size`, `quadtree_symmetry`.

Secondly, the two models predicting the average complexity rating thanks to the AIM measures and the average beauty rating based on the ergonomic criteria, are less efficient (equations 6.2 and 6.3 respectively linked with the hypotheses 1.3.2 and 1.3.3). Indeed, their respective coefficients of determination,  $R^2$ , were obtained and equal to 32% and 34%. Thus, these two predicting models capture less variability of the dependent variable. In contrast, our resulting model predicting the average complexity rating thanks to ergonomic criteria does not possess any explanatory variable (hypothesis 1.3.4). We conclude that ergonomic criteria cannot predict the complexity perceived by users.

Then, we tried to evaluate if the e-commerce sites are aesthetically different or have a complexity which is different from other sites. We found no statistically significant difference in beauty and complexity rating means between the e-commerce websites and the remaining categories combined. As regards the AIM measures, we only studied the behaviour of the variables that predict the beauty rating with a coefficient of de-

termination of 68% (equation 6.1 related to hypothesis 1.3.1). Among these eight AIM measures and the ergonomic criteria, the only variables that possess a statistically significant different means among the e-commerce websites and the remaining categories are the *HSV Avg\_average\_value*, the adaptation, the dialogue control and the guidance. E-commerce websites have a *HSV Avg\_average\_value*, a guidance, an adaptation and a dialogue control greater than 0.07, 0.97, 0.67 and 0.6, respectively. In other words, this means that the value of the Hue Saturation Value (HSV) colour space is on average higher for electronic commerce websites. The users are more guided through this category of websites than in the other categories combined. They also feel that they have more control over the interface. Furthermore, electronic commerce websites are more adapted to the user and his context.

## 7.2 Benefits and limits

In this thesis, we have improved our understanding of the literature dedicated to the evaluation of the aesthetics of mobile interfaces. By means of the Systematic Literature Review, we have identified studies carried out on the subject. In addition, we have improved our comprehension of the predictive models of user evaluations based on measurements. It shed some light on factors influencing the human judgement. However, we have also identified several limitations to our research.

Firstly, some limits are induced by the experiment methodology as well as the tools used in our family of experiments. On QUESTIM, the experimenters manually divide the screen into specific zones. In addition to being more time-consuming than an automatic method, the results might be affected by human subjectivity. Some variations may occur when different people or even the same person at different times perform this segmentation. Furthermore, each manual procedure is subject to variation, interpretation and error regardless of the quality of the established protocol. To ensure that measures always lead to the same result in the same conditions, it would be interesting to implement a fully automatic measure calculation as carried out by AIM and its segmentation zone script.

Secondly, manual intervention in the process of measuring the visual quality of an interface represents a limit to the reproducibility of the experiment. Reproducibility is an essential principle in science. It eases the cumulative process of knowledge where results are confronted to the previous ones. It refers to the performance of an experiment based on a previous one with a different experimental setup and a different team of experimenters. It should not be mistaken with the concept of replicability, that uses the same experimental setup, or repeatability, that uses the same experimental setup deployed by the same experimenters (Burny and Vanderdonckt, 2021). Conforming and striving to these notions is essential to generalise, infirm or expand previous results. For this

reason, human intervention must be limited during the deployment of an experiment and automatic measures must be favoured.

Furthermore, we discovered that deploying experiments takes time, resources to create a large dataset with user ratings, different measures, and several configurations (modifying the resolutions and/or the type of devices used). Consequently, the reproducibility of this type of experiment becomes even more complicated. Therefore, software such as UILAB should be preferred. Indeed, UILAB possesses features that ease the design of experiments and enhance reproducibility such as the ability to process multiple inputs at a time, to change automatically computed measures, and to customize the types of user data collected (Burny and Vanderdonckt, 2021).

As regards the ergonomic criteria, a dynamic interaction should be created with the interface to evaluate them correctly. In fact, our experiment is based solely on static screenshots while some criteria required dynamic interaction with the GUI to be evaluated properly. For instance, the error management is related to the capacity of the interface to deal and fix errors done by the user. The criteria are assessed more accurately when the evaluator has the possibility to browse through different pages of the websites and use the research field, for instance.

Some specific aspects of the perceived complexity and aesthetics could also limit our results. The user evaluations might vary according to demographic factors such as gender and age. On the one hand, studies have shown that colourfulness and complexity are affected by the user background, including, age and educational level. Gender might also impact the evaluation of aesthetics (Reinecke et al., 2013). On the other hand, some of the screenshots captured are in other languages or alphabets than the one used by users. Consequently, this aspect could influence the complexity perceived by users.

## 7.3 Future work

For future research, we advocate for an aesthetic assessment protocol that is reproducible and, therefore, based on objective measures. It is interesting to automate as much as possible the experiments process to be able to perform sophisticated studies.

### 7.3.1 Short term

Based on the data collected in the third experiment, we might replicate the analysis conducted on e-commerce websites for other SimilarWeb categories such as the “Games” category. Furthermore, it might be interesting to compare the obtained measures between two different website categories instead of making this comparison between a specific category and all other categories combined (as in hypotheses 1.4.1 to 1.4.4).

### 7.3.2 Midterm

On a longer term, it could be interesting to replicate or to reproduce the experiments carried out in this thesis. For example, experiments on the evaluation of the aesthetics of mobile interfaces could be conducted in other geographical areas. The same experimental set-up, deployed in one of our experiments, can be conducted in other countries. We could also make further investigations on the impact of various socio-demographic factors such as age and frequency of use of the device.

Moreover, we could replicate the third experiment by implementing other measures. Indeed, we could use different sets of aesthetic measures or potentially adjust the existing measures from computers to mobile devices.

While we focused on static evaluation, we could also capture the interface in a more dynamic way. It could be interesting to collect a series of screenshots or to record a video of a user's navigation through a website.

Furthermore, we could capture the screens of mobile applications instead of web pages given that mobile apps are specifically designed for this type of device. The visual design of other interactive systems can be assessed, such as commercial software.

Finally, the orientation of the devices could also be changed by capturing screens in landscape mode instead of portrait mode. More broadly, the visual design of other GUIs representations could be evaluated such as sketches and prototypes. This would allow to follow the evolution of the interface during the design and the development process, among other things.

# Bibliography

- Abeelee, M. V., Beullens, K., & Roe, K. (2013). Measuring mobile phone use: Gender, age and real usage level in relation to the accuracy and validity of self-reported mobile phone use. *Mobile Media & Communication*, 1(2), 213–236. <https://doi.org/10.1177/2050157913477095>
- Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, 5(07), 754–767. <https://doi.org/10.4236/ojs.2015.57075>
- Altman, D. G., Clarke, M., Devereaux, P., Gøtzsche, P. C., Ioannidis, J. P., Kleijnen, J., Liberati, A., Moher, D., Mulrow, C., & Tetzlaff, J. (2009). The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, 62(10), e1–e34. <https://doi.org/10.1016/j.jclinepi.2009.06.006>
- Bastien, C., & Scapin, D. (1993). *Ergonomic criteria for the evaluation of human-computer interfaces* (tech. rep. RT-0156). INRIA. <https://hal.inria.fr/inria-00070012>
- Bastien, C., & Scapin, D. (1995). Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction*, 7(2), 105–121. <https://doi.org/10.1080/10447319509526114>
- Bessghaier, N., Soui, M., Kolski, C., & Chouchane, M. (2021). On the detection of structural aesthetic defects of android mobile user interfaces with a metrics-based tool. *ACM Transactions on Interactive Intelligent Systems*, 11(1), 1–27. <https://doi.org/10.1145/3410468>
- Burny, N., & Vanderdonckt, J. (2021). Uilab, a workbench for conducting and reproducing experiments in gui visual design. *Proceedings of the ACM on Human-Computer Interaction*, 5, 1–31. <https://doi.org/10.1145/3457143>
- Casteleyn, S., Garrig'os, I., & Maz'on, J.-N. (2014). Ten years of rich internet applications: A systematic mapping study, and beyond. *ACM Transactions on the Web*, 8(3), 1–46. <https://doi.org/10.1145/2626369>
- Ceci, L. (2021). *Mobile internet usage worldwide - statistics & facts*. Retrieved December 7, 2021, from <https://www.statista.com/topics/779/mobile-internet/#dossierKeyfigures>

- Chittaro, L. (2011). Designing visual user interfaces for mobile applications. *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, 331–332. <https://doi.org/10.1145/1996461.1996550>
- Choi, J. H., & Lee, H.-J. (2012). Facets of simplicity for the smartphone interface: A structural model. *International Journal of Human-Computer Studies*, 70(2), 129–142. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2011.09.002>
- Clement, J. (2021). *Percentage of mobile device website traffic worldwide from 1st quarter 2015 to 1st quarter 2021*. Retrieved December 7, 2021, from <https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/>
- Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., & Xu, Y.-Q. (2006). Color harmonization. *ACM SIGGRAPH 2006 Papers*, 624–630. <https://doi.org/10.1145/1179352.1141933>
- Coppola, D. (2021). *E-commerce worldwide - statistics & facts*. Retrieved December 7, 2021, from <https://www.statista.com/topics/871/online-shopping/>
- Gaffar, A. (2009). Enumerating mobile enterprise complexity 21 complexity factors to enhance the design process. *Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, 270–282. <https://doi.org/10.1145/1723028.1723059>
- Gotteland, D., Haon, C., & Jolibert, A. (2012). Spécifier l’objet de la recherche. *Méthodologie de la recherche en sciences de gestion: Réussir son mémoire ou sa thèse* (pp. 63–104). Pearson Education France.
- Hasler, D., & Susstrunk, S. E. (2003). Measuring colorfulness in natural images. *Human Vision and Electronic Imaging VIII*, 5007, 87–95. <https://doi.org/10.1117/12.477378>
- ISO 25000. (2021). *Iso/iec 25010*. Retrieved December 17, 2021, from <https://iso25000.com/index.php/en/iso-25000-standards/iso-25010>
- Ivory, M. Y., Sinha, R. R., & Hearst, M. A. (2001). Empirically validated web page design metrics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 53–60. <https://doi.org/10.1145/365024.365035>
- Jacquemin, A. (2017). Penser sa methodologie. *Unpublished document, UCLouvain, Louvain-La-Neuve*.
- Jansen, B. J. (1998). The graphical user interface. *Special Interest Group on Computer-Human Interaction*, 30(2), 22–26. <https://doi.org/10.1145/279044.279051>
- Johnson, J. (2021a). *Global digital population as of january 2021*. Retrieved December 7, 2021, from <https://www.statista.com/statistics/617136/digital-population-worldwide/>
- Johnson, J. (2021b). *Internet usage worldwide - statistics & factss*. Retrieved December 7, 2021, from <https://www.statista.com/topics/1145/internet-usage-worldwide/#dossierKeyfigures>

- Kirsch, A. (2019). *Analysis of the aesthetics of e-commerce web sites on mobile devices* (Master's thesis). Université Catholique de Louvain, Louvain-La-Neuve.
- Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering – a tertiary study. *Information and Software Technology*, 52(8), 792–805. <https://doi.org/https://doi.org/10.1016/j.infsof.2010.03.006>
- Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, 60(3), 269–298. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2003.09.002>
- Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, 25(2), 115–126. <https://doi.org/10.1080/01449290500330448>
- Lund, S., Madgavkar, A., Manyika, J., Smit, S., Ellingrud, K., & Robinson, O. (2021). *The future of work after covid-19*. Retrieved December 7, 2021, from <https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19>
- Miniukovich, A., & De Angeli, A. (2014a). Quantification of interface visual complexity. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 153–160. <https://doi.org/10.1145/2598153.2598173>
- Miniukovich, A., & De Angeli, A. (2014b). Visual impressions of mobile app interfaces. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 31–40. <https://doi.org/10.1145/2639189.2641219>
- Miniukovich, A., & De Angeli, A. (2015). Computation of interface aesthetics. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1163–1172. <https://doi.org/10.1145/2702123.2702575>
- Ngo, D. C. L., & Byrne, J. G. (2001). Another look at a model for evaluating interface aesthetics. *International Journal of Applied Mathematics and Computer Science*, 11, 515–535.
- Ngo, D. C. L., Samsudin, A., & Abdullah, R. (2000). Aesthetic measures for assessing graphic screens. *Journal of information science and engineering*, 16(1), 97–116.
- Ngo, D. C. L., Teo, L. S., & Byrne, J. G. (2003). Modelling interface aesthetics. *Information Sciences*, 152, 25–46. [https://doi.org/https://doi.org/10.1016/S0020-0255\(02\)00404-8](https://doi.org/https://doi.org/10.1016/S0020-0255(02)00404-8)
- Ngo, D., Teo, L., & Byrne, J. (2000). Formalising guidelines for the design of screen layouts. *Displays*, 21(1), 3–15. [https://doi.org/https://doi.org/10.1016/S0141-9382\(00\)00026-3](https://doi.org/https://doi.org/10.1016/S0141-9382(00)00026-3)
- Niranjanamurthy, M., Kavyashree, N., Jagannath, S., & Chahar, D. (2013). Analysis of e-commerce and m-commerce: Advantages, limitations and security issues. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(6), 2360–2370.

- Njike, C., & Segers, J. (2017). Tp 4 : Régression linéaire multiple - linge1222 analyse statistique multivariée. *Unpublished document, UCLouvain, Louvain-La-Neuve*.
- Online Etymology Dictionary. (2021). *Aesthetics (n.)* Retrieved December 18, 2021, from [https://www.etymonline.com/word/aesthetics#etymonline\\_v\\_29562](https://www.etymonline.com/word/aesthetics#etymonline_v_29562)
- Oulasvirta, A., Miniukovich, A., Palmas, G., Weinkauff, T., Pascale, S., Koch, J., Langerak, T., Jokinen, J., Todi, K., & Kristhombuge, M. (2018). Aalto interface metrics (aim) a service and codebase for computational gui evaluation. *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, 16–19.
- Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19), 8877–8882. <https://doi.org/10.1073/pnas.0906172107>
- Pearl Brereton, O., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(4), 571–583. <https://doi.org/https://doi.org/10.1016/j.jss.2006.07.009>
- Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1), 45–48. <https://doi.org/10.1111/1467-9280.00008>
- Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013). Predicting users’ first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2049–2058. <https://doi.org/10.1145/2470654.2481281>
- Riegler, A., & Holzmann, C. (2018). Measuring Visual User Interface Complexity of Mobile Applications With Metrics. *Interacting with Computers*, 30(3), 207–223. <https://doi.org/10.1093/iwc/iwy008>
- Robins, D., & Holmes, J. (2008). Aesthetics and credibility in web site design. *Information Processing & Management*, 44(1), 386–399. <https://doi.org/https://doi.org/10.1016/j.ipm.2007.02.003>
- Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7(2), 17–17. <https://doi.org/10.1167/7.2.17>
- Saunders, M., Lewis, P., & Thornhill, A. (2009). Critically reviewing the literature. *Research methods for business students* (pp. 70–115). Pearson education.
- Scapin, D., Vanderdonckt, J., Farenc, C., Bastide, R., Bastien, C., Leulier, C., Mariage, C., & Palanque, P. (2001). Transferring knowledge of user interfaces guidelines to the web. In J. Vanderdonckt & C. Farenc (Eds.), *Tools for working with guidelines* (pp. 293–304). Springer London.

- Solecki, I., Porto, J., Alves, N. d. C., Gresse von Wangenheim, C., Hauck, J., & Borgatto, A. F. (2020). Automated assessment of the visual design of android apps developed with app inventor. *SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 51–57. <https://doi.org/10.1145/3328778.3366868>
- Soui, M., Chouchane, M., Gasmi, I., & Mkaouer, M. W. (2017). Plain: Plugin for predicting the usability of mobile user interface. *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 127–136. <https://doi.org/10.5220/0006171201270136>
- Soui, M., Chouchane, M., Mkaouer, M. W., Kessentini, M., & Ghedira, K. (2020). Assessing the quality of mobile graphical user interfaces using multi-objective optimization. *Soft Computing*, 24(10), 7685–7714. <https://doi.org/https://doi.org/10.1007/s00500-019-04391-8>
- Tractinsky, N., Katz, A., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13(2), 127–145. [https://doi.org/10.1016/S0953-5438\(00\)00031-X](https://doi.org/10.1016/S0953-5438(00)00031-X)
- Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies*, 64(11), 1071–1083. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2006.06.009>
- Tuch, A. N., Bargas-Avila, J. A., Opwis, K., & Wilhelm, F. H. (2009). Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies*, 67(9), 703–715. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2009.04.002>
- Tuch, A. N., Presslauer, E. E., Stöcklin, M., Opwis, K., & Bargas-Avila, J. A. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, 70(11), 794–811. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2012.06.003>
- Vanderdonckt, J. (1995). Accessing guidelines information with sierra. In K. Nordby, P. Helmersen, D. J. Gilmore, & S. A. Arnesen (Eds.), *Human—computer interaction: Interact '95* (pp. 311–316). Springer US. [https://doi.org/10.1007/978-1-5041-2896-4\\_52](https://doi.org/10.1007/978-1-5041-2896-4_52)
- Vanderdonckt, J. (2021). Linfo1311 interfaces homme-machine - slides. *Unpublished document, UCLouvain, Louvain-La-Neuve*.
- Villarreal-Narvaez, S., Vanderdonckt, J., Vatavu, R.-D., & Wobbrock, J. O. (2020). Dis '20: Proceedings of the 2020 acm designing interactive systems conference, 855–872. <https://doi.org/10.1145/3357236.3395511>
- Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

- Zen, M. (2017). *A methodology for assessing aesthetics of a graphical user interface of an information system: Visual measures-based automated evaluation* (Doctoral dissertation). Université Catholique de Louvain, Louvain-La-Neuve.
- Zen, M., & Vanderdonckt, J. (2014). Towards an evaluation of graphical user interfaces aesthetics based on metrics. *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 1–12. <https://doi.org/10.1109/RCIS.2014.6861050>
- Zen, M., & Vanderdonckt, J. (2016). Assessing user interface aesthetics based on the inter-subjectivity of judgment. *Proceedings of the 30th International BCS Human Computer Interaction Conference (HCI)*, 1–12. <https://doi.org/10.14236/ewic/HCI2016.25>
- Zheng, X. S., Chakraborty, I., Lin, J. J.-W., & Rauschenberger, R. (2009). Correlating low-level image statistics with users - rapid aesthetic and affective judgments of web pages. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–10. <https://doi.org/10.1145/1518701.1518703>

# Appendices

# Appendix A

## Aesthetics measures: additional formulas

### A.1 Regularity

According to Bessghaier et al. (2021), regularity is a function of  $RM_{alignment}$  and  $RM_{spacing}$  which are, respectively, linked to the minimisation and the spaces consistency of the alignment points. Firstly,  $RM_{alignment}$  is computed thanks to the number of vertical ( $n_{vap}$ ) and horizontal ( $n_{nap}$ ) alignment points (Bessghaier et al., 2021):

$$RM_{alignment} = \begin{cases} 1, & \text{if } n=1 \\ 1 - \frac{n_{vap}+n_{nap}}{2n}, & \text{Otherwise} \end{cases}$$

where  $n$  is the number of objects present on the layout.

Secondly,  $RM_{spacing}$  depends on the number of different distances between starting points of the columns and the rows (Bessghaier et al., 2021):

$$RM_{spacing} = \begin{cases} 1, & \text{if } n=1 \\ 1 - \frac{n_{spacing}-1}{2(n-1)}, & \text{Otherwise} \end{cases}$$

The regularity measure ( $Rgl$ ) is the mean of the absolute values of the two measures presented above (equation A.1) (Bessghaier et al., 2021):

$$Rgl = \frac{|RM_{alignment}| + |RM_{spacing}|}{2} \in [0, 1] \quad (\text{A.1})$$

## A.2 Homogeneity

Homogeneity can be computed thanks to the number of possible arrangements of  $n$  objects in the four quadrants ( $W$ ). This value is a function of the total number of objects ( $n$ ) and the number of objects in the upper-left ( $n_{UL}$ ), in the upper-right ( $n_{UR}$ ), in the lower-left ( $n_{LL}$ ) as well as in the lower-right ( $n_{LR}$ ) quadrant (equation A.2) (Bessghaier et al., 2021):

$$W = \frac{n!}{n_{UL}!n_{UR}!n_{LL}!n_{LR}} \quad (\text{A.2})$$

In addition, the maximum value of the number of possible arrangements  $W$  is calculated as follows (equation A.3) (Bessghaier et al., 2021):

$$W_{max} = \frac{n!}{(\frac{n}{4}!)^4} \quad (\text{A.3})$$

Then, homogeneity ( $HM$ ) is defined as the ratio of the two previously mentioned measures (equation A.4) (Bessghaier et al., 2021):

$$HM = \frac{W}{W_{max}} \in [0, 1] \quad (\text{A.4})$$

## A.3 Balance

According to Bessghaier et al. (2021), balance is decomposed into vertical and horizontal dimensions which are, both, computed based on  $w_j$  (equation A.5)(Bessghaier et al., 2021):

$$w_j = 1 - \sum_i^{n_i} a_{ij} \times d_{ij} \quad (\text{A.5})$$

where  $a_{ij}$  is the area of an object  $i$  on side  $j$  (knowing that  $j \in [L, R, T, B]$ ) and  $d_{ij}$  is the distance from an object to the frame.

The balance ( $BM$ ) is the association of its vertical ( $BM_{vertical}$ ) and horizontal ( $BM_{horizontal}$ ) components. The three measures are defined by the following formulas (equations A.6, A.7, A.8) (Bessghaier et al., 2021):

$$BM_{vertical} = \frac{w_L - w_R}{\max(|w_L|, |w_R|)} \quad (\text{A.6})$$

$$BM_{horizontal} = \frac{w_T - w_B}{\max(|w_T|, |w_B|)} \quad (\text{A.7})$$

$$BM = 1 - \frac{|BM_{vertical}| + |BM_{horizontal}|}{2} \in [0, 1] \quad (\text{A.8})$$

## A.4 Cohesion

Firstly, cohesion measure is a function of the height-width ratios regarding the layout and the screen ( $CM_{fl}$ ), according to Bessghaier et al. (2021). It is calculated using the layout height-width and the screen height-width ratios (Bessghaier et al., 2021):

$$CM_{fl} = \begin{cases} \frac{h_{layout}/b_{layout}}{h_{frame}/b_{frame}}, & \text{if } \frac{h_{layout}/b_{layout}}{h_{frame}/b_{frame}} \leq 1 \\ \frac{h_{frame}/b_{frame}}{h_{layout}/b_{layout}}, & \text{Otherwise} \end{cases}$$

where  $h_{layout}$  and  $b_{layout}$  are, respectively, the layout height and width. Similarly,  $h_{frame}$  and  $b_{frame}$  are, respectively, the frame height and width.

Secondly, the cohesion part related to the objects and the layout ( $CM_{lo}$ ) is computed as follows (equation A.9) (Bessghaier et al., 2021):

$$CM_{lo} = \frac{\sum_i^n f_i}{n} \quad (\text{A.9})$$

where  $n$  is the number of objects.  $f_i$  is calculated using the height-width ratios of the objects and the layout knowing that the width and the height of an object  $i$  are respectively noted  $b_i$  and  $h_i$  (Bessghaier et al., 2021):

$$f_i = \begin{cases} \frac{h_i/b_i}{h_{layout}/b_{layout}}, & \text{if } \frac{h_i/b_i}{h_{layout}/b_{layout}} = 1 \\ \frac{h_{layout}/b_{layout}}{h_i/b_i}, & \text{Otherwise} \end{cases}$$

Bessghaier et al. (2021) define the cohesion ( $CM$ ) as the mean of  $CM_{fl}$  and  $CM_{lo}$  (equation A.10) (Bessghaier et al., 2021):

$$CM = \frac{|CM_{fl}| + |CM_{lo}|}{2} \in [0, 1] \quad (\text{A.10})$$

## A.5 Unity

Unity is a function of an object similarity (size and shape) measure and a measure related to spaces. Specially,  $UM_{form}$  is related to the object similarity regarding its size and shape and is computed as follows (equation A.11) (Bessghaier et al., 2021):

$$UM_{form} = 1 - \frac{n_{size} + n_{shape} - 2}{2} \in [0, 1] \quad (\text{A.11})$$

where  $n_{size}$  and  $n_{shape}$  are respectively the number of sizes and shapes.

The relative measures linked to spaces ( $UM_{space}$ ) are computed based on the following formula (equation A.12) (Bessghaier et al., 2021):

$$UM_{shape} = 1 - \frac{a_{layout} - \sum_i^n a_i}{a_{frame} - \sum_i^n a_i} \in [0, 1] \quad (\text{A.12})$$

where  $a_{layout}$ ,  $a_i$  and  $a_{frame}$  are respectively the area of the layout, object  $i$  and frame whereas  $n$  is the number of object.

The unity measure ( $Unt$ ) is computed thanks to equations A.11 and A.12 (equation A.13) (Bessghaier et al., 2021):

$$Unt = \frac{|UM_{form}| + |UM_{space}|}{2} \in [0, 1] \quad (\text{A.13})$$

## A.6 Element smallness

The element smallness is computed, by Riegler and Holzmann (2018), as the average width and height of the different screenshot elements. Besides, the screenshot is, first, divided into groups, by using the background colours as a criterion. Then, the width and the height means are computed thanks to  $width_{u,g,s}$  and  $height_{u,g,s}$  which are the width and the height of an element  $u$  belonging to a group  $g$  and included in a screen  $s$ . The arithmetical width and height means over all the elements of a group  $U_{g,s}$  are calculated and weighted by the area of the group  $A_{g,s}$ . The weighted arithmetic means over all the groups  $G_s$  is then calculated (equations A.14 and A.15) (Riegler and Holzmann, 2018):

$$\overline{width}_s = \frac{1}{A_s} \sum_{g=1}^{G_s} A_{g,s} \frac{1}{U_{g,s}} \sum_{u=1}^{U_{g,s}} width_{u,g,s} \quad (\text{A.14})$$

$$\overline{height}_s = \frac{1}{A_s} \sum_{g=1}^{G_s} A_{g,s} \frac{1}{U_{g,s}} \sum_{u=1}^{U_{g,s}} height_{u,g,s} \quad (\text{A.15})$$

According to Riegler and Holzmann (2018), the element smallness ( $E_s$ ) is defined as a combination of  $\overline{width}_s$  and  $\overline{height}_s$  (equation A.16) (Riegler and Holzmann, 2018):

$$E_s = \frac{\left(1 - \frac{\overline{width}_s}{width_s}\right) + \left(1 - \frac{\overline{height}_s}{height_s}\right)}{2} \quad (\text{A.16})$$

## A.7 Misalignment

Riegler and Holzmann (2018) present a measure of misalignment ( $M_s$ ) which is composed of three submeasures: the average number of alignments of a screen  $s$ , from a horizontal ( $\overline{numh_s}$ ), a vertical ( $\overline{numv_s}$ ) and a central ( $\overline{numc_s}$ ) perspective. As these measures are calculated analogously, we will only introduce one of them. The average number of vertical alignments ( $\overline{numv_s}$ ) is computed using the number of vertical alignments ( $numv_{g,s}$ ) and the number of possible alignments ( $num_{g,s}$ , whether horizontal, vertical or central) in a group  $g$  and a screen  $s$  (equation A.17) (Riegler and Holzmann, 2018):

$$\overline{numv_s} = \frac{1}{A_s} \sum_{g=1}^{G_s} A_{g,s} \frac{numv_{g,s}}{num_{g,s}} \quad (\text{A.17})$$

where  $A_s$  and  $A_{g,s}$  are, respectively, the area of a screen  $s$  and the area of a group  $g$  (included in  $G_s$ ) and belonging to a screen  $s$ .

The misalignment measure ( $M_s$ ) is defined as the opposite of the three alignment submeasures (equation A.18) (Riegler and Holzmann, 2018):

$$M_s = 1 - (\overline{numh_s} + \overline{numv_s} + \overline{numc_s}) \quad (\text{A.18})$$

## A.8 Imbalance

Imbalance is composed, according to Riegler and Holzmann (2018), of the mean horizontal balance ( $\overline{balh_s}$ ) and the mean vertical balance ( $\overline{balv_s}$ ). These two latter measures are obtained from  $balh_{g,s}$  and  $balv_{g,s}$ .  $\overline{balv_s}$  is a function of the top and the bottom margin of an element  $u$  (included in  $U_{g,s}$ ) of a group  $g$  and a screen  $s$ , respectively noted  $mart_{u,g,s}$  and  $marb_{u,g,s}$  (equation A.19). Besides, the margins of an element are proportional to the distances between its closest elements or the borders. The vertical margins, as mentioned, include the top and bottom margins whereas the horizontal margins are related to the left and right margins. An analogous reasoning can be applied to compute the horizontal balance. The maximum horizontal and vertical margins are derived from the margin measures (Riegler and Holzmann, 2018).

$$balv_{g,s} = \frac{1}{U_{g,s}} \sum_{u=1}^{U_{g,s}} \frac{mart_{u,g,s} + marb_{u,g,s}}{2 \times \text{marvmax}_{g,s}} \quad (\text{A.19})$$

In addition, the measure  $balv_{g,s}$  (equation A.19) is weighted by the group area  $A_{g,s}$  to obtain the  $\overline{balv}_s$  (equation A.20) (Riegler and Holzmann, 2018).

$$\overline{balv}_s = \frac{1}{A_s} \sum_{g=1}^{G_s} A_{g,s} \times balv_{g,s} \quad (\text{A.20})$$

Then, the imbalance ( $B_s$ ), defined as the opposite of the measure of balance, can be computed as follows (equation A.21) (Riegler and Holzmann, 2018):

$$B_s = 1 - \frac{\overline{balh}_s + \overline{balv}_s}{2} \quad (\text{A.21})$$

## A.9 Colour complexity

According to Riegler and Holzmann (2018), the colour complexity, is computed based on the number of dominant colours ( $numc_s$ ) of a screen  $s$ . The dominant colours ( $c_1, c_2$ , etc.) appear more recurrently on a pixel-basis than other colours and are identified by a k-mean clustering algorithm. The colour complexity also refers to the colour complexity value ( $ccm_{c1,c2}$ ) associated to a dominant colour pair combination. This latter measure is calculated as a function of the H (hue), S (saturation) and V (value) values of the two colours (equation A.22) (Riegler and Holzmann, 2018):

$$ccm_{c1,c2} = 0.70 \times \left( \frac{\delta H_{c1,c2}}{H_{max}} \right) + 0.15 \times \left( \frac{S_{max} - \delta S_{c1,c2}}{S_{max}} \right) + 0.15 \times \left( \frac{V_{max} - \delta V_{c1,c2}}{V_{max}} \right) \quad (\text{A.22})$$

where the  $H_{max}$ ,  $S_{max}$  and  $V_{max}$  are the maximum hue, saturation, and value.

The colour complexity measure ( $C_s$ ) is computed thanks to the mean complexity value of the colour combination ( $\overline{ccm}_s$ ), the maximum number of dominant colours ( $numc_{max}$ ) and the colour combination match of a screen's histogram ( $h_s$ ) and that of its successor ( $h_{s+1}$ ) ( $c_{h_s, h_{s+1}}$ ) (equation A.23) (Riegler and Holzmann, 2018).

$$C_s = \frac{\frac{numc_s}{numc_{max}} + (1 - \overline{ccm}_s) + c_{h_s, h_{s+1}}}{3} \quad (\text{A.23})$$

# Appendix B

## Experiment 1

### B.1 Shapiro-Wilk normality tests of the QUESTIM measures of SimilarWeb

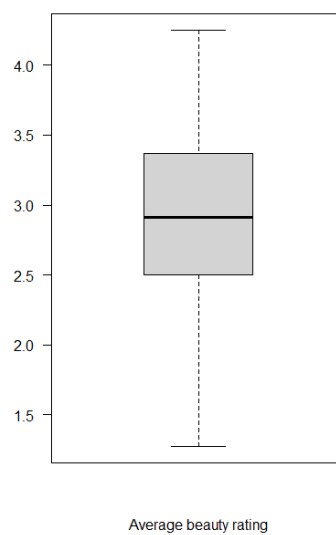
| Measure               | W      | P-value | Normal ( $\alpha=0.01$ )? | P-value summary |
|-----------------------|--------|---------|---------------------------|-----------------|
| SaliencyBalance       | 0.9623 | 0.300   | Yes                       | ns              |
| BorderBalance         | 0.9413 | 0.074   | Yes                       | ns              |
| BorderDensity         | 0.8866 | 0.002   | No                        | **              |
| ColorDensity          | 0.9559 | 0.198   | Yes                       | ns              |
| Colorfulness          | 0.9261 | 0.027   | Yes                       | *               |
| CompressionComplexity | 0.8459 | <0.001  | No                        | ***             |
| Balance               | 0.9243 | 0.024   | Yes                       | *               |
| VerticalBalance       | 0.9360 | 0.052   | Yes                       | ns              |
| HorizontalBalance     | 0.8079 | <0.001  | No                        | ***             |
| Equilibrium           | 0.9036 | 0.007   | No                        | **              |
| Density               | 0.9774 | 0.705   | Yes                       | ns              |
| CenterAlignment       | 0.9760 | 0.659   | Yes                       | ns              |
| ExternalAlignment     | 0.9405 | 0.070   | Yes                       | ns              |
| Concentricity         | 0.9826 | 0.862   | Yes                       | ns              |
| Simplicity            | 0.9596 | 0.252   | Yes                       | ns              |
| Symmetry              | 0.9833 | 0.878   | Yes                       | ns              |

Among the 16 QUESTIM variables, four do not follow a normal distribution: border density, compression complexity, horizontal balance and the equilibrium.

# Appendix C

## Experiment 2

### C.1 Box plot of the average beauty rating



The average beauty rating possesses a dispersed distribution.

## C.2 Descriptive statistics of the average beauty rating

|                | var_beauty_mean |
|----------------|-----------------|
| Minimum        | 1.273           |
| Maximum        | 4.25            |
| Range          | 2.977           |
| Mean           | 2.928           |
| Std. Deviation | 0.605           |
| Variance       | 0.366           |

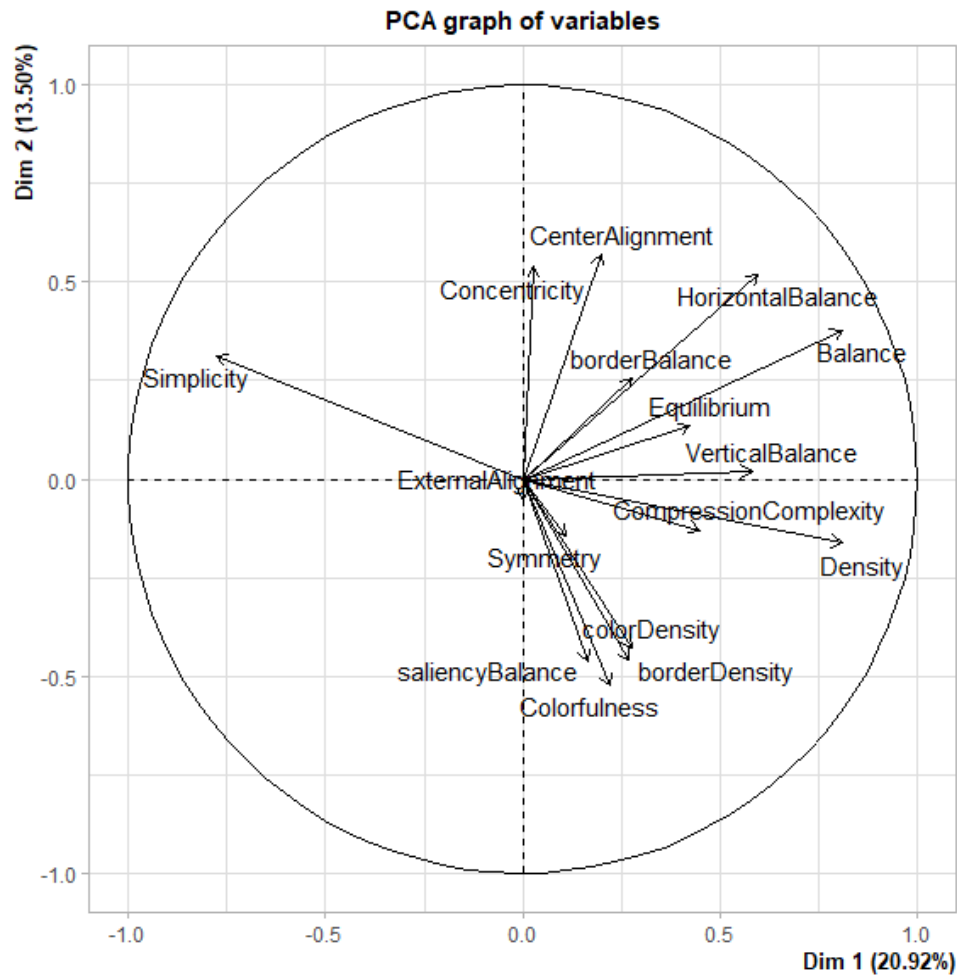
The mean of the average beauty rating is slightly greater than 2.5. This means that, on average, the participants did not rate websites as having a maximum or a minimum beauty on the evaluation scale.

## C.3 Shapiro-Wilk normality tests of the QUESTIM measures of Alexa

| Measure               | W      | P-value | Normal ( $\alpha=0.01$ )? | P-value summary |
|-----------------------|--------|---------|---------------------------|-----------------|
| SaliencyBalance       | 0.9599 | 0.0231  | Yes                       | *               |
| BorderBalance         | 0.9718 | 0.1109  | Yes                       | ns              |
| BorderDensity         | 0.8946 | <0.0001 | No                        | ****            |
| ColorDensity          | 0.9811 | 0.3619  | Yes                       | ns              |
| Colorfulness          | 0.9581 | 0.0185  | Yes                       | *               |
| CompressionComplexity | 0.7147 | <0.0001 | No                        | ****            |
| Balance               | 0.9477 | 0.0051  | No                        | **              |
| VerticalBalance       | 0.955  | 0.0125  | Yes                       | *               |
| HorizontalBalance     | 0.6733 | <0.0001 | No                        | ****            |
| Equilibrium           | 0.8494 | <0.0001 | No                        | ****            |
| Density               | 0.968  | 0.0666  | Yes                       | ns              |
| CenterAlignment       | 0.9742 | 0.1513  | Yes                       | ns              |
| ExternalAlignment     | 0.8816 | <0.0001 | No                        | ****            |
| Concentricity         | 0.9809 | 0.3522  | Yes                       | ns              |
| Simplicity            | 0.9669 | 0.0578  | Yes                       | ns              |
| Symmetry              | 0.9916 | 0.9201  | Yes                       | ns              |

Among the 16 QUESTIM variables, six do not follow a normal distribution: BorderDensity, CompressionComplexity, Balance, HorizontalBalance, Equilibrium and ExternalAlignment.

## C.4 Principal Component Analysis with the QUESTIM measures



By visualising the first and the second principal components which get 34.42% of the total variability of the model, we can confirm what was observed on the correlation matrix. For instance, the simplicity and the density are negatively correlated variables.

## C.5 Linear regression of the average beauty rating as a function of all QUESTIM measures

```
Call:
lm(formula = df_beauty_QUESTIM$Score ~ ., data = df_beauty_QUESTIM)

Residuals:
    Min       1Q   Median       3Q      Max
-1.47677 -0.35083  0.02265  0.34106  0.91141

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.33088     5.07375   2.036  0.04656 *
saliencyBalance  0.17982     0.54619   0.329  0.74323
borderBalance  -0.40697     0.40265  -1.011  0.31656
borderDensity   4.93201     2.45558   2.008  0.04951 *
colorDensity    0.70201     0.55164   1.273  0.20852
Colorfulness   -0.43557     0.28983  -1.503  0.13860
CompressionComplexity -1.72358     1.03642  -1.663  0.10200
Balance         1.34921     0.68615   1.966  0.05431 .
VerticalBalance -0.28805     0.46374  -0.621  0.53707
HorizontalBalance      NA          NA      NA      NA
Equilibrium      -3.60182     4.30871  -0.836  0.40681
Density          -3.44022     1.02298  -3.363  0.00141 **
CenterAlignment  0.55243     0.67444   0.819  0.41626
ExternalAlignment 0.94602     1.09374   0.865  0.39083
Concentricity    0.07534     0.80576   0.093  0.92585
Simplicity      -6.60912     1.99842  -3.307  0.00166 **
Symmetry        -1.87165     3.15935  -0.592  0.55600
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5396 on 55 degrees of freedom
Multiple R-squared:  0.3747,    Adjusted R-squared:  0.2042
F-statistic: 2.197 on 15 and 55 DF,  p-value: 0.01765
```

The significance test and the parameter estimation of the horizontal balance cannot be computed by RStudio.

## C.6 VIF - linear regression of the average beauty rating as a function of all the QUESTIM measures

```
> lr_score_Quest = lm(df_beauty_QUESTIM$score~., data= df_beauty_QUESTIM)
> library("car")
> vif(lr_score_Quest)
Error in vif.default(lr_score_Quest) :
  there are aliased coefficients in the model
```

The Variance Inflation Factor (VIF), related to the linear regression of the average beauty rating as a function of the 16 QUESTIM measures, cannot be computed on RStudio. Indeed, the QUESTIM measures are too linearly dependent.

## C.7 VIF - linear regression of the average beauty rating as a function of 14 unfiltered QUESTIM measures

```
> lr_score_Quest = lm(df_beauty_QUESTIM$score~., data= df_beauty_QUESTIM)
> vif(lr_score_Quest)
```

|                 |                       |                   |                   |
|-----------------|-----------------------|-------------------|-------------------|
| saliencyBalance | borderBalance         | borderDensity     | colorDensity      |
| 1.388826        | 1.186534              | 2.342919          | 1.594099          |
| colorfulness    | compressionComplexity | verticalBalance   | horizontalBalance |
| 1.465484        | 1.993956              | 1.593215          | 1.624427          |
| equilibrium     | centerAlignment       | externalAlignment | concentricity     |
| 2.201216        | 1.524309              | 1.310975          | 1.549748          |
| simplicity      | symmetry              |                   |                   |
| 1.986114        | 1.353423              |                   |                   |

According to the Variance Inflation Factor, we do not need to discard variables (no VIF upper than 5).

## C.8 Stepwise regression using 14 QUESTIM measures to predict the average beauty rating

```
Call:
lm(formula = df_beauty_QUESTIM$score ~ df_beauty_QUESTIM$borderDensity +
    df_beauty_QUESTIM$Equilibrium + df_beauty_QUESTIM$CompressionComplexity)

Residuals:
    Min       1Q   Median       3Q      Max
-1.58284 -0.43621  0.00717  0.40626  1.13618

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.9662     2.8848  -0.682   0.49787
df_beauty_QUESTIM$borderDensity    6.9107     2.1539   3.208   0.00205 **
df_beauty_QUESTIM$Equilibrium    5.0433     3.0021   1.680   0.09763 .
df_beauty_QUESTIM$CompressionComplexity -1.4215     0.9622  -1.477   0.14426
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5719 on 67 degrees of freedom
Multiple R-squared:  0.1443,    Adjusted R-squared:  0.1059
F-statistic: 3.765 on 3 and 67 DF,  p-value: 0.01467
```

The obtained model thanks to the stepwise method is composed of three explanatory variables. We can observe, on the last column of the figure, that some variables are still not significant ( $p\text{-value} > 0.05$ ).

## C.9 Final linear regression of the average beauty rating using the QUESTIM measures

```
call:
lm(formula = df_beauty_QUESTIM$Score ~ df_beauty_QUESTIM$borderDensity)

Residuals:
    Min       1Q   Median       3Q      Max
-1.75372 -0.43090 -0.01203  0.47056  1.06514

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.6690     0.1229   21.71  <2e-16 ***
df_beauty_QUESTIM$borderDensity  4.3932     1.7231    2.55   0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5824 on 69 degrees of freedom
Multiple R-squared:  0.0861,    Adjusted R-squared:  0.07285
F-statistic:  6.5 on 1 and 69 DF,  p-value: 0.01301
```

After the stepwise method, we removed the nonsignificant variables one by one ( $p\text{-value} > 0.05$ ). The resulting model is composed of one explanatory variable: BorderDensity.

## C.10 Analysis of the variables among the first half compared to the second half of the Alexa ranking sites

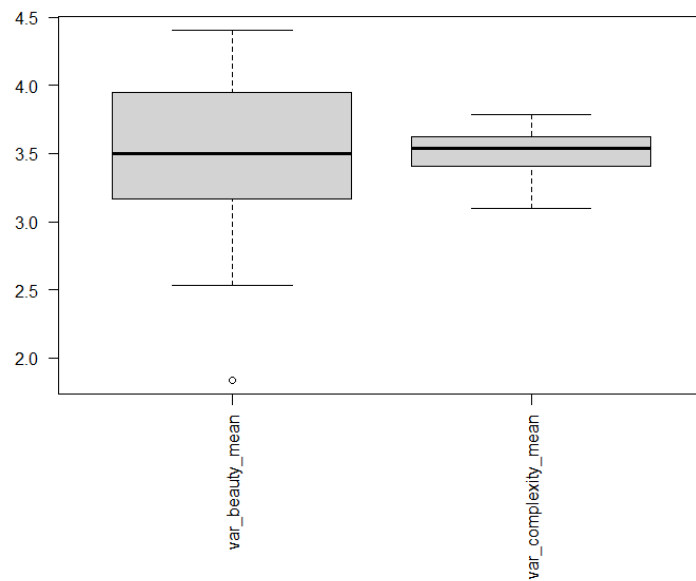
| Variable              | Popular normality | Less popular normality | Normality | Var. = | Test              | P-v. Test | Mean = |
|-----------------------|-------------------|------------------------|-----------|--------|-------------------|-----------|--------|
| Average beauty rating | 0.754             | 0.787                  | Yes       | Yes    | ANOVA (equal var) | 0.352     | Yes    |
| SaliencyBalance       | 0.366             | 0.048                  | Yes       | Yes    | ANOVA (equal var) | 0.062     | Yes    |
| BorderBalance         | 0.608             | 0.128                  | Yes       | Yes    | ANOVA (equal var) | 0.351     | Yes    |
| BorderDensity         | 0.008             | 0.001                  | No        | /      | Kruskal-Wallis    | 0.375     | Yes    |
| ColorDensity          | 0.592             | 0.370                  | Yes       | Yes    | ANOVA (equal var) | 0.981     | Yes    |
| Colorfulness          | 0.011             | 0.112                  | Yes       | Yes    | ANOVA (equal var) | 0.556     | Yes    |
| CompressionComplexity | 0.000             | 0.000                  | No        | /      | Kruskal-Wallis    | 0.327     | Yes    |
| Balance               | 0.125             | 0.044                  | Yes       | Yes    | ANOVA (equal var) | 0.898     | Yes    |
| VerticalBalance       | 0.318             | 0.031                  | Yes       | Yes    | ANOVA (equal var) | 0.481     | Yes    |
| HorizontalBalance     | 0.000             | 0.000                  | No        | /      | Kruskal-Wallis    | 0.927     | Yes    |
| Equilibrium           | 0.880             | 0.000                  | No        | /      | Kruskal-Wallis    | 0.482     | Yes    |
| Density               | 0.481             | 0.157                  | Yes       | Yes    | ANOVA (equal var) | 0.250     | Yes    |
| CenterAlignment       | 0.153             | 0.620                  | Yes       | Yes    | ANOVA (equal var) | 0.734     | Yes    |
| ExternalAlignment     | 0.000             | 0.002                  | No        | /      | Kruskal-Wallis    | 0.189     | Yes    |
| Concentricity         | 0.797             | 0.113                  | Yes       | Yes    | ANOVA (equal var) | 0.897     | Yes    |
| Simplicity            | 0.244             | 0.236                  | Yes       | Yes    | ANOVA (equal var) | 0.447     | Yes    |
| Symmetry              | 0.244             | 0.236                  | Yes       | Yes    | ANOVA (equal var) | 0.605     | Yes    |

This table summarises the analysis of the different variables among the popular websites and the less popular websites. The second column, named Popular normality, contains the p-values of the Shapiro-Wilk normality tests carried out on the most popular websites (first half of the Alexa ranking sites). The third column is similar to the second one. The only difference is that we tested the sites in the second part of the Alexa ranking. The column 4 indicates if both populations follow a normal distribution or not. If the two populations follow the normal distribution, an ANOVA test will be carried out, otherwise, the Kruskal-Wallis test will be performed. If the two populations have a normal distribution, the variance equality between the populations will be verified in column 5. Indeed, this information must be included while performing the ANOVA test. Besides, the test used to compare the two populations means is mentioned in the column 6. The related p-value and outcome are summarised in columns 7 and 8. The tests were performed using an  $\alpha=0.01$ .

# Appendix D

## Experiment 3

### D.1 Box plot of the average beauty and complexity rating



The average beauty rating is more spread than the complexity average rating.

## D.2 Descriptive statistics of the average beauty and complexity rating

|                | var_beauty_mean | var_complexity_mean |
|----------------|-----------------|---------------------|
| Minimum        | 1.833           | 3.097               |
| Maximum        | 4.405           | 3.788               |
| Range          | 2.572           | 0.690               |
| Mean           | 3.508           | 3.513               |
| Std. Deviation | 0.568           | 0.154               |
| Variance       | 0.323           | 0.024               |

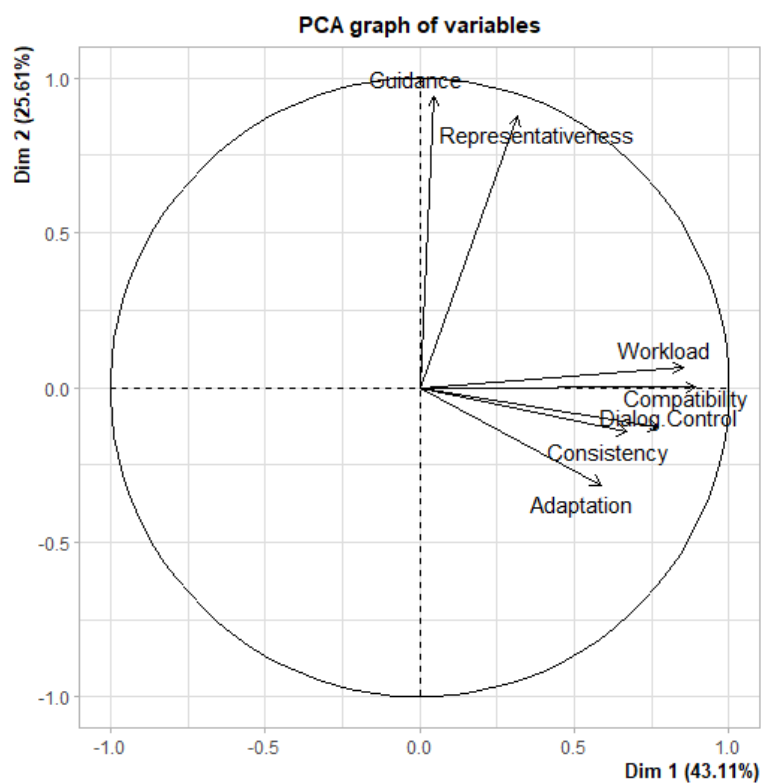
The average beauty rating has a larger range of values and a bigger standard deviation than the average complexity rating. This means that the beauty perception possesses a larger variability. Nevertheless, the two variables have a similar mean which is equal roughly to 3.5.

## D.3 Shapiro-Wilk normality tests of the ergonomic criteria

| Criterion          | W               | P-value | Normal ( $\alpha=0.01$ )? | P-value summary |
|--------------------|-----------------|---------|---------------------------|-----------------|
| Compatibility      | 0.8944          | <,001   | No                        | ***             |
| Consistency        | 0.7941          | <,001   | No                        | ***             |
| Workload           | 0.9091          | 0.001   | No                        | **              |
| Adaptation         | 0.7972          | <,001   | No                        | ***             |
| Dialog.Control     | 0.904           | <,001   | No                        | ***             |
| Representativeness | 0.8561          | <,001   | No                        | ***             |
| Guidance           | 0.7753          | <,001   | No                        | ***             |
| Error Management   | Input not valid |         |                           |                 |

The ergonomic criteria do not follow a normal distribution. It is noteworthy that the error management normality test is not carried out as the variable is constant in the e-commerce websites category.

## D.4 Principal Component Analysis with the ergonomic criteria



By visualising the first and the second principal components which get 68% of the total variability of the model, we can observe that the representativeness and the guidance seem strongly correlated.

## D.5 Descriptive statistics of the AIM measures

| Measure                         | Minimum   | Maximum     | Range       | Mean       | Std. Deviation | Variance        |
|---------------------------------|-----------|-------------|-------------|------------|----------------|-----------------|
| Color_harmony_i_distance        | 13900.000 | 1379858.000 | 1365958.000 | 314070.000 | 289586.000     | 83860223892.759 |
| Color_harmony_V_distance        | 3920.000  | 727974.000  | 724054.000  | 142255.000 | 146298.000     | 21402987603.018 |
| Color_harmony_L_distance        | 2398.000  | 318994.000  | 316596.000  | 55824.000  | 59482.000      | 3538082822.037  |
| Color_harmony_Linverse_distance | 1880.000  | 316619.000  | 314739.000  | 54450.000  | 58921.000      | 3471663920.082  |
| Color_harmony_I_distance        | 1550.000  | 416528.000  | 414978.000  | 77703.000  | 100562.000     | 10112783022.867 |
| Color_harmony_T_distance        | 542.600   | 179045.000  | 178503.000  | 16630.000  | 27555.000      | 759255408.745   |
| Color_harmony_X_distance        | 88.920    | 54362.000   | 54273.000   | 4551.000   | 8744.000       | 76461434.517    |
| Color_harmony_Y_distance        | 295.300   | 127559.000  | 127263.000  | 18141.000  | 29277.000      | 857138610.608   |
| Color_harmony_i_alpha           | 0.000     | 355.000     | 355.000     | 173.500    | 116.000        | 13448.379       |
| Color_harmony_V_alpha           | 3.000     | 356.000     | 353.000     | 191.000    | 106.600        | 11363.457       |
| Color_harmony_L_alpha           | 13.000    | 351.000     | 338.000     | 170.500    | 99.210         | 9841.922        |
| Color_harmony_Linverse_alpha    | 3.000     | 357.000     | 354.000     | 196.300    | 109.100        | 11910.991       |
| Color_harmony_I_alpha           | 2.000     | 175.000     | 173.000     | 46.270     | 47.610         | 2266.657        |
| Color_harmony_T_alpha           | 3.000     | 329.000     | 326.000     | 200.100    | 100.200        | 10047.901       |
| Color_harmony_X_alpha           | 13.000    | 312.000     | 299.000     | 157.000    | 85.970         | 7391.645        |
| Color_harmony_Y_alpha           | 8.000     | 346.000     | 338.000     | 129.100    | 104.800        | 10979.642       |
| HasslerSusstrunk_RG_mean        | 0.537     | 74.230      | 73.690      | 23.350     | 19.620         | 384.752         |
| HasslerSusstrunk_RG_stdev       | 2.502     | 95.180      | 92.670      | 36.990     | 22.080         | 487.307         |
| HasslerSusstrunk_YB_mean        | 4.972     | 77.290      | 72.320      | 22.660     | 15.080         | 227.303         |
| HasslerSusstrunk_YB_stdev       | 16.480    | 72.280      | 55.790      | 36.580     | 14.170         | 200.869         |
| HasslerSusstrunk_RGYB_mean      | 6.562     | 94.200      | 87.640      | 33.620     | 23.220         | 538.984         |
| HasslerSusstrunk_RGYB_stdev     | 21.200    | 108.200     | 86.980      | 53.740     | 22.430         | 503.032         |
| HasslerSusstrunk_Colorfulness   | 24.170    | 133.900     | 109.700     | 63.820     | 28.810         | 830.253         |
| HSVAvg_average_hue              | 2.032     | 357.700     | 355.700     | 213.100    | 159.900        | 25575.433       |
| HSVAvg_average_sat              | 0.031     | 0.608       | 0.577       | 0.204      | 0.127          | 0.016           |
| HSVAvg_sat_stdev                | 0.106     | 0.443       | 0.337       | 0.275      | 0.086          | 0.007           |

| Measure                | Minimum    | Maximum     | Range       | Mean        | Std. Deviation | Variance         |
|------------------------|------------|-------------|-------------|-------------|----------------|------------------|
| HSVAvg_average_value   | 0.673      | 0.974       | 0.301       | 0.846       | 0.082          | 0.007            |
| HSVAvg_value_stdev     | 0.056      | 0.351       | 0.295       | 0.229       | 0.069          | 0.005            |
| HSVAvg_nb_unique_hsv   | 5878.000   | 36199.000   | 30321.000   | 15155.000   | 6358.000       | 40422318.260     |
| HSVAvg_nb_unique_hue   | 8978.000   | 45956.000   | 36978.000   | 20310.000   | 9875.000       | 97515621.207     |
| HSVAvg_nb_unique_sat   | 11564.000  | 21761.000   | 10197.000   | 18433.000   | 2734.000       | 7472149.122      |
| HSVAvg_nb_unique_value | 220.000    | 256.000     | 36.000      | 254.700     | 5.869          | 34.449           |
| LABAvg_lightness_mean  | 57.030     | 96.470      | 39.440      | 78.920      | 9.614          | 92.426           |
| LABAvg_lightness_stdev | 9.140      | 35.420      | 26.280      | 26.070      | 6.366          | 40.530           |
| LABAvg_A_mean          | -8.418     | 23.780      | 32.200      | 5.192       | 7.735          | 59.836           |
| LABAvg_A_stdev         | 2.768      | 34.610      | 31.840      | 14.430      | 9.277          | 86.067           |
| LABAvg_B_mean          | -23.780    | 17.800      | 41.580      | 2.043       | 7.441          | 55.366           |
| LABAvg_B_stdev         | 7.220      | 35.690      | 28.470      | 18.580      | 7.880          | 62.097           |
| luminance_stdev        | 25.260     | 94.870      | 69.610      | 71.830      | 16.350         | 267.248          |
| Nb_colors              | 656.000    | 102453.000  | 101797.000  | 39150.000   | 23683.000      | 560899353.167    |
| Png_file_size          | 129183.000 | 2230332.000 | 2101149.000 | 1035428.000 | 485832.000     | 236032497760.599 |
| Wave_score             | 0.471      | 0.658       | 0.187       | 0.539       | 0.043          | 0.002            |
| Edge_congestion        | 0.035      | 0.196       | 0.161       | 0.093       | 0.039          | 0.001            |
| Edge_density           | 0.038      | 0.130       | 0.093       | 0.082       | 0.023          | 0.001            |
| FigureGround_contrast  | -12.100    | 18.890      | 30.980      | 1.053       | 3.363          | 11.308           |
| Nb_alignment_lines     | 33.000     | 259.000     | 226.000     | 127.500     | 55.590         | 3089.754         |
| Jpeg_file_size         | 124323.000 | 277683.000  | 153360.000  | 203882.000  | 34418.000      | 1184621087.969   |
| Pixel_symmetry         | 0.144      | 1.090       | 0.946       | 0.544       | 0.178          | 0.032            |
| Quadtree_balance       | 0.430      | 0.945       | 0.515       | 0.769       | 0.142          | 0.020            |
| Quadtree_symetry       | 0.426      | 0.912       | 0.486       | 0.677       | 0.139          | 0.019            |
| Quadtree_equilibrium   | 1.000      | 1.000       | 0.000       | 1.000       | 0.000          | 9.308E-10        |
| Quadtree_nb_leafs      | 2188.000   | 10936.000   | 8748.000    | 6023.000    | 1939.000       | 3760950.421      |
| White_space            | 0.103      | 0.894       | 0.791       | 0.483       | 0.186          | 0.035            |

## D.6 Shapiro-Wilk normality tests of the AIM measures

| Measure                         | W      | P-value | Normality ( $\alpha=0.01$ )? | P-value summary |
|---------------------------------|--------|---------|------------------------------|-----------------|
| Color_harmony_i_distance        | 0.8647 | <.001   | No                           | ***             |
| Color_harmony_V_distance        | 0.8214 | <.001   | No                           | ***             |
| Color_harmony_L_distance        | 0.7732 | <.001   | No                           | ***             |
| Color_harmony_Linverse_distance | 0.7662 | <.001   | No                           | ***             |
| Color_harmony_I_distance        | 0.6865 | <.001   | No                           | ***             |
| Color_harmony_T_distance        | 0.5266 | <.001   | No                           | ***             |
| Color_harmony_X_distance        | 0.4994 | <.001   | No                           | ***             |
| Color_harmony_Y_distance        | 0.6052 | <.001   | No                           | ***             |
| Color_harmony_i_alpha           | 0.8961 | <.001   | No                           | ***             |
| Color_harmony_V_alpha           | 0.9448 | 0.023   | Yes                          | *               |
| Color_harmony_L_alpha           | 0.9012 | <.001   | No                           | ***             |
| Color_harmony_Linverse_alpha    | 0.9075 | <.001   | No                           | ***             |
| Color_harmony_I_alpha           | 0.7038 | <.001   | No                           | ***             |
| Color_harmony_T_alpha           | 0.8432 | <.001   | No                           | ***             |
| Color_harmony_X_alpha           | 0.7962 | <.001   | No                           | ***             |
| Color_harmony_Y_alpha           | 0.837  | <.001   | No                           | ***             |
| HasslerSusstrunk_RG_mean        | 0.8708 | <.001   | No                           | ***             |
| HasslerSusstrunk_RG_stdev       | 0.9395 | 0.014   | Yes                          | *               |
| HasslerSusstrunk_YB_mean        | 0.8773 | <.001   | No                           | ***             |
| HasslerSusstrunk_YB_stdev       | 0.9296 | 0.006   | No                           | **              |
| HasslerSusstrunk_RGYB_mean      | 0.904  | <.001   | No                           | ***             |
| HasslerSusstrunk_RGYB_stdev     | 0.9375 | 0.012   | Yes                          | *               |
| HasslerSusstrunk_Colorfulness   | 0.9347 | 0.009   | No                           | **              |
| HSVAvg_average_hue              | 0.7026 | <.001   | No                           | ***             |
| HSVAvg_average_sat              | 0.9314 | 0.007   | No                           | **              |
| HSVAvg_sat_stdev                | 0.9664 | 0.173   | Yes                          | ns              |

| Measure                | W      | P-value | Normality ( $\alpha=0.01$ )? | P-value summary |
|------------------------|--------|---------|------------------------------|-----------------|
| HSVAvg_average_value   | 0.9331 | 0.008   | No                           | **              |
| HSVAvg_value_stdev     | 0.9676 | 0.194   | Yes                          | ns              |
| HSVAvg_nb_unique_hsv   | 0.9263 | 0.004   | No                           | **              |
| HSVAvg_nb_unique_hue   | 0.8944 | <.001   | No                           | ***             |
| HSVAvg_nb_unique_sat   | 0.926  | 0.004   | No                           | **              |
| HSVAvg_nb_unique_value | 0.2289 | <.001   | No                           | ***             |
| LABAvg_lightness_mean  | 0.9843 | 0.751   | Yes                          | ns              |
| LABAvg_lightness_stdev | 0.9493 | 0.035   | Yes                          | *               |
| LABAvg_A_mean          | 0.8661 | <.001   | No                           | ***             |
| LABAvg_A_stdev         | 0.9153 | 0.002   | No                           | **              |
| LABAvg_B_mean          | 0.9076 | <.001   | No                           | ***             |
| LABAvg_B_stdev         | 0.9101 | 0.001   | No                           | **              |
| luminance_stdev        | 0.9234 | 0.004   | No                           | **              |
| Nb_colors              | 0.9605 | 0.099   | Yes                          | ns              |
| Png_file_size          | 0.9778 | 0.476   | Yes                          | ns              |
| Wave_score             | 0.9687 | 0.214   | Yes                          | ns              |
| Edge_congestion        | 0.9069 | <.001   | No                           | ***             |
| Edge_density           | 0.9792 | 0.53    | Yes                          | ns              |
| FigureGround_contrast  | 0.4436 | <.001   | No                           | ***             |
| Nb_alignment_lines     | 0.9673 | 0.188   | Yes                          | ns              |
| Jpeg_file_size         | 0.9838 | 0.73    | Yes                          | ns              |
| Pixel_symmetry         | 0.9726 | 0.306   | Yes                          | ns              |
| Quadtree_balance       | 0.8839 | <.001   | No                           | ***             |
| Quadtree_symmetry      | 0.9492 | 0.034   | Yes                          | *               |
| Quadtree_equilibrium   | 0.8421 | <.001   | No                           | ***             |
| Quadtree_nb_leafs      | 0.9802 | 0.574   | Yes                          | ns              |
| White_space            | 0.9879 | 0.891   | Yes                          | ns              |

Among the 53 AIM measures, 17 possess a normal distribution. The same results were obtained both on RStudio and on GraphPad Prisma.

## D.7 Linear regression of the average beauty rating as a function of all AIM measures

```
Call:
lm(formula = df_ecomm$var_beauty_mean ~ ., data = df_beauty_AIM)

Residuals:
ALL 49 residuals are 0: no residual degrees of freedom!

Coefficients: (5 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.982e+01      NaN      NaN      NaN
color_harmony_i_distance -1.902e-05      NaN      NaN      NaN
color_harmony_v_distance  3.437e-05      NaN      NaN      NaN
color_harmony_L_distance  1.561e-04      NaN      NaN      NaN
color_harmony_Linverse_distance -1.398e-04      NaN      NaN      NaN
color_harmony_I_distance  1.875e-05      NaN      NaN      NaN
color_harmony_T_distance -6.279e-05      NaN      NaN      NaN
color_harmony_X_distance -1.078e-04      NaN      NaN      NaN
color_harmony_Y_distance -2.039e-05      NaN      NaN      NaN
color_harmony_i_alpha    -7.474e-03      NaN      NaN      NaN
color_harmony_v_alpha    -3.075e-03      NaN      NaN      NaN
color_harmony_L_alpha    -1.729e-03      NaN      NaN      NaN
color_harmony_Linverse_alpha -2.824e-03      NaN      NaN      NaN
color_harmony_I_alpha     1.302e-02      NaN      NaN      NaN
color_harmony_T_alpha     3.699e-03      NaN      NaN      NaN
color_harmony_X_alpha    -3.246e-03      NaN      NaN      NaN
color_harmony_Y_alpha     3.306e-03      NaN      NaN      NaN
Hasslersusstrunk_RG_mean  3.795e-01      NaN      NaN      NaN
Hasslersusstrunk_RG_stdev 9.351e-02      NaN      NaN      NaN
Hasslersusstrunk_YB_mean  2.277e-01      NaN      NaN      NaN
Hasslersusstrunk_YB_stdev -6.121e-02      NaN      NaN      NaN
Hasslersusstrunk_RGYB_mean -4.279e-01      NaN      NaN      NaN
Hasslersusstrunk_RGYB_stdev -1.448e-02      NaN      NaN      NaN
Hasslersusstrunk_Colorfulness NA          NA          NA          NA
HSVavg_average_hue       2.375e-03      NaN      NaN      NaN
```

RStudio fails to test the variables significance and to estimate all parameters. It should be noted that the screenshot only shows the results for a part of the variables as the rest does not provide any additional information.

## D.8 VIF - linear regression of the average beauty rating as a function of all AIM measures

```
> lr_beauty_AIM = lm(df_ecomm$var_beauty_mean~., data=df_beauty_AIM)
> library("car")
> vif(lr_beauty_AIM)
Error in vif.default(lr_beauty_AIM) :
  there are aliased coefficients in the model
```

The AIM variables are too linearly dependent. Therefore, the Variance Inflation Factor cannot be computed on RStudio.

## D.9 Remaining AIM measures after having filtered the highly correlated variables (average beauty rating prediction)

```
"color_harmony_L_distance"  
"color_harmony_I_distance"  
"color_harmony_T_distance"  
"color_harmony_Y_distance"  
"color_harmony_i_alpha"  
"color_harmony_v_alpha"  
"color_harmony_L_alpha"  
"color_harmony_I_alpha"  
"color_harmony_T_alpha"  
"color_harmony_X_alpha"  
"color_harmony_Y_alpha"  
"HasslerSusstrunk_YB_stdev"  
"HSVavg_average_hue"  
"HSVavg_average_sat"  
"HSVavg_average_value"  
"HSVavg_nb_unique_hue"  
"HSVavg_nb_unique_value"  
"LABavg_A_mean"  
"LABavg_A_stdev"  
"LABavg_B_mean"  
"nb_colors"  
"edge_congestion"  
"figureGround_contrast"  
"nb_alignment_lines"  
"jpeg_file_size"  
"pixel_symmetry"  
"quadtree_symmetry"
```

The AIM measures were filtered to avoid collinearity and multicollinearity between variables. We created an algorithm to select the measures thanks to Spearman correlations. If two variables had a correlation bigger than 0.8 or lower than -0.8, the variable with the highest correlation with the average beauty rating was kept. Among the 53 AIM measures, 27 will be used to predict the average beauty rating.

## D.10 Stepwise regression using the 27 AIM measures to predict the average beauty rating

```

call:
lm(formula = df_ecomm$var_beauty_mean ~ df_ecomm$color_harmony_L_distance +
  df_ecomm$color_harmony_T_distance + df_ecomm$color_harmony_V_alpha +
  df_ecomm$color_harmony_L_alpha + df_ecomm$color_harmony_T_alpha +
  df_ecomm$color_harmony_X_alpha + df_ecomm$HSVAvg_average_hue +
  df_ecomm$HSVAvg_average_value + df_ecomm$HSVAvg_nb_unique_value +
  df_ecomm$LABAvg_B_mean + df_ecomm$nb_colors + df_ecomm$jpeg_file_size +
  df_ecomm$pixel_symmetry + df_ecomm$quadtreesymetry)

Residuals:
    Min       1Q   Median       3Q      Max
-0.68284 -0.17222  0.00011  0.17291  0.54052

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.623e+00  2.472e+00   1.061 0.296104
df_ecomm$color_harmony_L_distance  2.509e-06  1.730e-06   1.450 0.156114
df_ecomm$color_harmony_T_distance -1.328e-05  3.365e-06  -3.945 0.000379 ***
df_ecomm$color_harmony_V_alpha    2.144e-03  7.700e-04   2.784 0.008709 **
df_ecomm$color_harmony_L_alpha    2.425e-03  7.145e-04   3.394 0.001765 **
df_ecomm$color_harmony_T_alpha   -4.016e-03  1.058e-03  -3.796 0.000579 ***
df_ecomm$color_harmony_X_alpha   -1.565e-03  7.198e-04  -2.173 0.036804 *
df_ecomm$HSVAvg_average_hue     -5.417e-04  4.298e-04  -1.260 0.216109
df_ecomm$HSVAvg_average_value   -2.939e+00  7.951e-01  -3.696 0.000766 ***
df_ecomm$HSVAvg_nb_unique_value  1.477e-02  9.121e-03   1.619 0.114591
df_ecomm$LABAvg_B_mean          1.604e-02  7.594e-03   2.112 0.042070 *
df_ecomm$nb_colors               8.969e-06  3.126e-06   2.869 0.007024 **
df_ecomm$jpeg_file_size        -8.681e-06  1.744e-06  -4.978 1.83e-05 ***
df_ecomm$pixel_symmetry         7.056e-01  3.243e-01   2.175 0.036650 *
df_ecomm$quadtreesymetry        1.526e+00  4.222e-01   3.614 0.000963 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3081 on 34 degrees of freedom
Multiple R-squared:  0.7916,    Adjusted R-squared:  0.7057
F-statistic: 9.222 on 14 and 34 DF, p-value: 7.196e-08

```

The output of the stepwise regression is a model with 14 explanatory variables. Significance tests, displayed on the last column, were performed for each variable. It can be observed that some variables are still not significant (p-value>0.05).

## D.11 Final linear regression of the average beauty rating using the AIM measures

```
call:
lm(formula = df_ecomm$var_beauty_mean ~ df_ecomm$color_harmony_T_distance +
  df_ecomm$color_harmony_L_alpha + df_ecomm$color_harmony_T_alpha +
  df_ecomm$HSVAvg_average_value + df_ecomm$LABAvg_B_mean +
  df_ecomm$nb_colors + df_ecomm$jpeg_file_size + df_ecomm$quadtree_symetry)

Residuals:
    Min       1Q   Median       3Q      Max
-0.78193 -0.17886  0.00636  0.20653  0.75019

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.497e+00  7.797e-01   8.332 2.82e-10 ***
df_ecomm$color_harmony_T_distance -9.878e-06  2.249e-06  -4.392 8.02e-05 ***
df_ecomm$color_harmony_L_alpha    2.275e-03  7.270e-04   3.129 0.003270 **
df_ecomm$color_harmony_T_alpha   -2.089e-03  7.820e-04  -2.671 0.010873 *
df_ecomm$HSVAvg_average_value   -3.131e+00  7.060e-01  -4.435 7.02e-05 ***
df_ecomm$LABAvg_B_mean          1.617e-02  7.706e-03   2.098 0.042246 *
df_ecomm$nb_colors              1.233e-05  3.046e-06   4.049 0.000230 ***
df_ecomm$jpeg_file_size        -8.906e-06  1.789e-06  -4.977 1.27e-05 ***
df_ecomm$quadtree_symetry       1.706e+00  4.699e-01   3.630 0.000796 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3502 on 40 degrees of freedom
Multiple R-squared:  0.6831,    Adjusted R-squared:  0.6197
F-statistic: 10.78 on 8 and 40 DF,  p-value: 6.305e-08
```

After having removed on by one each non-significant variable ( $p\text{-value} > 0.05$ ), we obtain a linear model with eight explanatory variables.

## D.12 Remaining AIM measures after having filtered the highly correlated variables (average complexity rating prediction)

```
"color_harmony_L_distance"  
"color_harmony_I_distance"  
"color_harmony_T_distance"  
"color_harmony_Y_distance"  
"color_harmony_i_alpha"  
"color_harmony_v_alpha"  
"color_harmony_L_alpha"  
"color_harmony_I_alpha"  
"color_harmony_T_alpha"  
"color_harmony_X_alpha"  
"color_harmony_Y_alpha"  
"HasslersSusstrunk_RG_stdev"  
"HSVAvG_average_hue"  
"HSVAvG_nb_unique_hsv"  
"HSVAvG_nb_unique_hue"  
"HSVAvG_nb_unique_value"  
"LABAvG_lightness_mean"  
"LABAvG_A_mean"  
"LABAvG_B_mean"  
"edge_congestion"  
"figureGround_contrast"  
"nb_alignment_lines"  
"jpeg_file_size"  
"pixel_symmetry"  
"quadtree_symmetry"
```

The algorithm keeps 25 explanatory variables. Its functioning is similar to the one developed for the prediction of the average beauty rating (see Appendix D.9). The only difference is that the variable kept has a higher correlation with the complexity average rating than with the average beauty rating.

## D.13 Stepwise regression using the 25 AIM measures to predict the average complexity rating

```

call:
lm(formula = df_complexity_AIM$var_complexity_mean ~ df_complexity_AIM$color_harmony_I_distance +
  df_complexity_AIM$color_harmony_V_alpha + df_complexity_AIM$color_harmony_L_alpha +
  df_complexity_AIM$color_harmony_T_alpha + df_complexity_AIM$HSVAvg_nb_unique_hsv +
  df_complexity_AIM$HasslerSusstrunk_RG_stdev + df_complexity_AIM$figureGround_contrast)

Residuals:
    Min       1Q   Median       3Q      Max
-0.33431 -0.06465 -0.01137  0.07130  0.17878

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.306e+00   5.621e-02  58.806 < 2e-16 ***
df_complexity_AIM$color_harmony_I_distance -5.502e-07   2.173e-07  -2.532 0.015259 *
df_complexity_AIM$color_harmony_V_alpha    9.574e-04   2.563e-04   3.735 0.000571 ***
df_complexity_AIM$color_harmony_L_alpha    4.533e-04   2.412e-04   1.880 0.067243 .
df_complexity_AIM$color_harmony_T_alpha   -1.143e-03   3.384e-04  -3.378 0.001612 **
df_complexity_AIM$HSVAvg_nb_unique_hsv    8.174e-06   3.209e-06   2.548 0.014691 *
df_complexity_AIM$HasslerSusstrunk_RG_stdev 2.819e-03   1.036e-03   2.721 0.009510 **
df_complexity_AIM$figureGround_contrast   -9.350e-03   5.356e-03  -1.746 0.088355 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1198 on 41 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.3968
F-statistic: 5.511 on 7 and 41 DF, p-value: 0.0001622

```

After having deployed the stepwise regression, we obtained a model composed of seven explanatory variables. It can be observed, on the last column, that some variables are still not significant ( $p\text{-value} > 0.05$ ).

## D.14 Final linear regression of the complexity average rating using the AIM measures

```
Call:
lm(formula = df_complexity_AIM$var_complexity_mean ~ df_complexity_AIM$color_harmony_V_alpha +
    df_complexity_AIM$color_harmony_T_alpha + df_complexity_AIM$HasslerSusstrunk_RG_stdev)

Residuals:
    Min       1Q   Median       3Q      Max
-0.36094 -0.08114 -0.01259  0.10739  0.25064

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3900581   0.0469811  72.158 < 2e-16 ***
df_complexity_AIM$color_harmony_V_alpha  0.0010284   0.0002762   3.724 0.000545 ***
df_complexity_AIM$color_harmony_T_alpha -0.0009009   0.0003171  -2.841 0.006730 **
df_complexity_AIM$HasslerSusstrunk_RG_stdev 0.0028784   0.0009690   2.971 0.004755 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1317 on 45 degrees of freedom
Multiple R-squared:  0.3159,    Adjusted R-squared:  0.2703
F-statistic: 6.928 on 3 and 45 DF,  p-value: 0.0006223
```

The non-significant features were removed one by one ( $p\text{-value} > 0.05$ ). The final linear regression, obtained after this filtering stage, is composed of three independent variables.

## D.15 VIF - linear regression of the average beauty rating as a function of the ergonomic criteria

```
> vif(lr_beauty_ergo) # do not remove any variable
df_ecomm$Compatibility      df_ecomm$Consistency      df_ecomm$workload      df_ecomm$Adaptation      df_ecomm$Dialog.Control
      3.017490              1.552564              2.504735              1.465669              1.781986
df_ecomm$Representativeness      df_ecomm$Guidance
      2.722518              2.636426
```

On RStudio, we observed that no Variance Inflation Factor is higher than 5. This means that all the ergonomic criteria can be used to perform a linear regression.

## D.16 Stepwise regression using the ergonomic criteria to predict the average beauty rating

```
call:
lm(formula = df_ecomm$var_beauty_mean ~ df_ecomm$Dialog.Control +
    df_ecomm$Representativeness)

Residuals:
    Min       1Q   Median       3Q      Max
-1.34068 -0.26492  0.03004  0.33305  0.85730

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.47177    0.42185   3.489 0.001080 **
df_ecomm$Dialog.Control  0.18056    0.06098   2.961 0.004840 **
df_ecomm$Representativeness 0.24500    0.06770   3.619 0.000733 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4706 on 46 degrees of freedom
Multiple R-squared:  0.3421,    Adjusted R-squared:  0.3135
F-statistic: 11.96 on 2 and 46 DF,  p-value: 6.58e-05
```

The stepwise regression leads to a model with two independent variables which are both significant.

## D.17 VIF - linear regression of the average complexity rating as a function of the ergonomic criteria

```
> vif(lr_complexity_ergo) # do not remove any variable
df_ecomm$compatibility    df_ecomm$consistency    df_ecomm$workload    df_ecomm$adaptation    df_ecomm$dialog.Control
      3.017490             1.552564             2.504735             1.465669             1.781986
df_ecomm$representativeness df_ecomm$guidance
      2.722518             2.636426
```

The Variance Inflation Factor suggests to keep all the ergonomic criteria to deploy a linear regression as its values are lower than 5.

## D.18 Stepwise regression using the ergonomic criteria to predict the average complexity rating

```
call:
lm(formula = df_ecomm$var_complexity_mean ~ df_ecomm$Adaptation

Residuals:
    Min       1Q   Median       3Q      Max
-0.40206 -0.11355  0.02775  0.10772  0.25397

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.71738    0.14663   25.352  <2e-16 ***
df_ecomm$Adaptation -0.04362    0.03089   -1.412    0.165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1526 on 47 degrees of freedom
Multiple R-squared:  0.0407,    Adjusted R-squared:  0.02029
F-statistic: 1.994 on 1 and 47 DF,  p-value: 0.1645
```

The output of the stepwise regression is a model with one explanatory variable. Nevertheless, on the last column, we noticed that this variable is not significant ( $p\text{-value} > 0.05$ ).

## D.19 Analysis of the variables among the e-commerce and shopping category compared to the rest of the categories

| Variable                  | Ecomm normality | Rest normality | Normality | Var. = | Test                | P-v. Test | Mean = |
|---------------------------|-----------------|----------------|-----------|--------|---------------------|-----------|--------|
| Beauty average rating     | 0.199           | 0.134          | Yes       | No     | ANOVA (unequal var) | 0.361     | Yes    |
| Complexity average rating | 0.452           | 0.006          | No        | /      | Kruskal-Wallis      | 0.550     | Yes    |
| Color_harmony_T_distance  | 0.242           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.386     | Yes    |
| Color_harmony_L_alpha     | 0.001           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.975     | Yes    |
| Color_harmony_T_alpha     | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.759     | Yes    |
| HSVAvG_average_value      | 0.008           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.010     | No     |
| LABAvG_B_mean             | 0.001           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.075     | Yes    |
| Nb_colors                 | 0.099           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.246     | Yes    |
| Jpeg_file_size            | 0.730           | 2.19E-08       | No        | /      | Kruskal-Wallis      | 0.432     | Yes    |
| Quadtree_symetry          | 0.034           | 8.52E-15       | No        | /      | Kruskal-Wallis      | 0.371     | Yes    |
| Compatibility             | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.833     | Yes    |
| Consistency               | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.363     | Yes    |
| Workload                  | 0.001           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.659     | Yes    |
| Adaptation                | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.001     | No     |
| Dialog.Control            | 0.001           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.003     | No     |
| Representativeness        | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.021     | Yes    |
| Guidance                  | 0.000           | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.000     | No     |
| Error.Management          | Constant        | < 2.2e-16      | No        | /      | Kruskal-Wallis      | 0.242     | Yes    |

The column 2, named Ecomm normality, gathers the p-values of the Shapiro-Wilk normality test performed on the screenshots of the e-commerce and shopping category of SimilarWeb. The column 3 is similar to the column 2. The only difference is that the normality test is performed on all categories except the e-commerce and shopping category (meaning that the normality is checked on the 21 categories together). The fourth column indicates if the two populations have a normal distribution or not. The variance equality is checked, on column 5, for the normal variables. Depending on the results indicated in the columns 4 and 5, a mean comparison test was chosen (column 6). The p-values and the outcomes of the mean tests are presented in columns 7 and 8. Besides, the tests were carried out with an  $\alpha=0.01$ .

## D.20 Means of some variables among the e-commerce and shopping category compared to the rest of the categories

| Category | HSVAvg_average_value | Adaptation | Dialog.Control | Guidance   |
|----------|----------------------|------------|----------------|------------|
| Ecomm    | 0.845833129          | 4.69387755 | 4.71428571     | 5.12244898 |
| Rest     | 0.770113796          | 4.02576809 | 4.09117939     | 4.1555996  |

The test which verifies the means equality among the two populations reveals that this equality is rejected for four variables. The means of those variable among the two populations are represented in the table



# **Evaluating the Visual Design of Graphical User Interfaces on Mobile Devices: A Family of Experiments**

## **Resume**

The growing expansion of e-commerce practices has challenged traditional companies in the retail industry. In order to increase or capture new market parts, companies have been encouraged to strengthen their online presence. Therefore, the visual platforms for user interaction, called Graphical User Interfaces (GUIs) benefit from a particular attention from companies. Indeed, the aesthetics and the visual complexity of these interfaces significantly impact the users and represent a mean to establish a competitive advantage. Our research focuses on the measures of visual design on mobile devices, as this kind of device has become increasingly popular. Moreover, few discussions exist in the literature regarding this subject and stay poorly developed. On the one hand, this thesis aims to cover the gap that exists in the literature. On the other hand, it aims to find measures in order to assess the aesthetics and visual complexity of mobile devices through a Systematic Literature Review (SLR) and a family of experiments. We found that the most promising model of prediction captures 68% of the beauty ratings variability through eight measures presented by Aalto Interface Metrics (AIM). Furthermore, no significant aesthetic difference was observed among the popular and less popular websites. We also observed that four variables possess a higher mean among the e-commerce websites category compared to the other categories.