Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional
Zacatenco Campus
Computer Science Department

Epistemology of Anticipated User Experience: Task, User and Heuristic Approaches

Thesis submitted by
Luis Martín Sánchez Adame

as fulfillment of the requirements for the degree of
Doctor in Computer Science

Advisors:
Dr. Sonia Guadalupe Mendoza Chapa
Dr. Beatriz Adriana González Beltrán

Mexico City November 2021

Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional
Unidad Zacatenco
Departamento de Computación

Epistemología de la Experiencia de Usuario Anticipada: Enfoques de Tarea, Usuario y Heurístico

Tesis que presenta

Luis Martín Sánchez Adame

Para obtener el grado de

Doctor en Ciencias en Computación

Directoras de Tesis:

Dra. Sonia Guadalupe Mendoza Chapa

Dra. Beatriz Adriana González Beltrán

Ciudad de México

Noviembre 2021

"... If, for example, you come at four o'clock in the afternoon, then at three o'clock, I shall begin to be happy. I shall feel happier and happier as the hour advances..."
To my fox

"... Monsters exist, but they are too few in number to be truly dangerous; more dangerous are the common men, the functionaries ready to believe and to act without asking questions..."
To the memory of Primo Levi

ABSTRACT

In a market that is saturated and full of possibilities, people choose an application not only for its functionality but also for how it makes them feel and how well they can express themselves. These feelings are the field of study of User eXperience (UX), which seeks to make applications useful, beautiful and cause happiness. How do we know what we know in UX? How do we ensure that systems have valuable insights? To answer these questions of an epistemic nature, it is necessary to carry out UX and usability evaluations. Currently, much study has been done on evaluation issues. However, a field that has been relegated is Anticipated User eXperience (AUX), which concerns users' expectations, beliefs, and hopes before using an application. In this work, we try to expand the frontiers of knowledge about AUX. This is done through three UX/usability assessment approaches: *1) Task Oriented:* studying users' hopes to perform basic tasks in various environments; *2) User Oriented:* collecting user beliefs to create a tool; And, *3) Heuristic Oriented:* compiling expectations in the state of the art to conduct evaluations with experts. The consequence of these three approaches will allow developers to create higher-value products. The results quantitatively confirmed that AUX seems to be mainly composed of pragmatic rather than hedonic aspects, i.e., elements directly related to efficacy and efficiency to solve tasks. The development of this idea could lead to improving existing evaluation methods and the creation of new ones.

RESUMEN

En un mercado saturado y lleno de posibilidades, las personas eligen una aplicación no sólo por su funcionalidad, sino también por cómo los hace sentir y qué tan bien pueden expresarse. Estos sentimientos son el campo de estudio de la Experiencia de Usuario (UX por sus siglas en inglés), que busca que las aplicaciones sean útiles, hermosas y provoquen felicidad. ¿Cómo sabemos lo que sabemos en UX?, ¿cómo nos aseguramos que los sistemas tengan percepciones valiosas? Para responder a estas preguntas de naturaleza epistémica es necesario realizar evaluaciones de UX y de usabilidad. Actualmente se ha estudiado mucho en temas de evaluación, sin embargo, un campo que ha sido relegado es el de la Experiencia de Usuario Anticipada (AUX por sus siglas en inglés), que concierne a las expectativas, creencias y esperanzas de los usuarios antes de utilizar una aplicación. En este trabajo intentamos ampliar las fronteras del conocimiento sobre AUX. Esto se hace a través de tres enfoques de evaluación de UX/usabilidad: *1) Orientado a Tareas:* estudiando las esperanzas de los usuarios para realizar tareas básicas en varios entornos; *2) Orientado al Usuario:* recolectando las creencias de los usuarios para crear una herramienta; y *3) Orientado a Heurísticas:* compilando expectativas en el estado del arte para realizar evaluaciones con expertos. La consecuencia de estos tres enfoques permitirá a los desarrolladores crear productos de mayor valor. Los resultados confirmaron cuantitativamente que la AUX parece estar compuesta principalmente de aspectos pragmáticos en lugar de hedónicos, i.e., elementos directamente relacionados con eficacia y eficiencia para la realización de tareas. El desarrollo de esta idea podría conducir a la mejora de los métodos de evaluación existentes y la creación de otros nuevos.

PUBLICATIONS

This list contains all the works we have published as a result of this thesis:

JOURNALS

- Sánchez-Adame, L. M., Mendoza, S., Urquiza, J., Rodríguez, J. & Meneses-Viveros, A. (2021). Towards a set of heuristics for evaluating chatbots. *IEEE Latin America Transactions*, *19*(12), 2037–2045. https://doi.org/10.1109/TLA.2021.9480145

- Sánchez-Adame, L. M., Urquiza-Yllescas, J. F. & Mendoza, S. (2020b). Measuring anticipated and episodic ux of tasks in social networks. *Applied Sciences*, *10*, 8199. https://doi.org/10.3390/app10228199

CONFERENCES

- Mendoza, S., Hernández-León, M., Sánchez-Adame, L. M., Rodríguez, J., Decouchant, D. & Meneses-Viveros, A. (2020). Supporting student-teacher interaction through a chatbot. In P. Zaphiris & A. Ioannou (Eds.), *Learning and collaboration technologies. human and technology ecosystems* (pp. 93–107). Springer International Publishing. https://doi.org/10.1007/978-3-030-50506-6_8

- Sánchez-Adame, L. M., Mendoza, S., González-Beltrán, B. A., Meneses-Viveros, A. & Rodríguez, J. (2020a). The man in the besieged castle: Heuristic evaluation of home security systems. In A. Moallem (Ed.), *Hci for cybersecurity, privacy and trust* (pp. 250–260). Springer International Publishing. https://doi.org/10.1007/978-3-030-50309-3_17

- Sánchez-Adame, L. M., Mendoza, S., Meneses Viveros, A. & Rodríguez, J. (2019a). Towards a set of design guidelines for multi-device experience. In M. Kurosu (Ed.), *Human-computer*

interaction. perspectives on design (pp. 210–223). Springer International Publishing. https://doi.org/10.1007/978-3-030-22646-6_15

- Sánchez-Adame, L. M., Mendoza, S., Viveros, A. M. & Rodríguez, J. (2019b). Consistency in multi-device environments: A case study. In K. Arai, R. Bhatia & S. Kapoor (Eds.), *Intelligent computing* (pp. 232–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-22871-2_17

- Sánchez-Adame, L. M., Mendoza, S., González-Beltrán, B. A., Rodríguez, J. & Viveros, A. M. (2018a). Aux and ux evaluation of user tools in social networks. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 104–111. https://doi.org/10.1109/WI.2018.0-101

- Sánchez-Adame, L. M., Mendoza, S., González-Beltrán, B. A., Rodríguez, J. & Viveros, A. M. (2018b). Ux evaluation over time: User tools in social networks. *2018 15th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 1–6. https://doi.org/10.1109/ICEEE.2018.8533950

- Sánchez-Adame, L. M., Mendoza, S., González-Beltrán, B. A., Meneses Viveros, A. & Rodríguez, J. (2018c). Towards an aux evaluation framework for user tools in virtual communities. In A. Rodrigues, B. Fonseca & N. Preguiça (Eds.), *Collaboration and technology* (pp. 25–33). Springer International Publishing. https://doi.org/10.1007/978-3-319-99504-5_3

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
AR	Augmented Reality
AUX	Anticipated User eXperience
CoP	Communities of Practice
CUX	Cumulative User eXperience
DIY	Do It Yourself
DSRM	Design Science Research Methodology
DUH	Developing Usability Heuristics
DUI	Distributed User Interface
EUX	Episodic User eXperience
FAQ	Frequently Asked Questions
GDS	Google Design Sprint
GUI	Graphical User Interface
HCI	Human-Computer Interaction
IoT	Internet of Things
MUX	Momentary User eXperience
NLP	Natural Language Processing
UI	User Interface
UX	User eXperience

1

INTRODUCTION

The Swiss-French architect Charles-Édouard Jeanneret (1887-1965), better known as Le Corbusier, postulated the need of man for beauty and conceptualised it, pointing out that a thing is beautiful when it responds to a need and architecture must be beautiful, since it is something capable to produce happy people (Gossman, 1998; Magaña, 2019).

Following that line of thought, we would have to define beauty, happiness and need. Regarding the first concept, we can turn to the discussion that Stephen Dedalus has in the novel *A Portrait of the Artist as a Young Man* (1916) by James Joyce (1882-1941). Dedalus takes up the ideas of Thomas Aquinas (1225-1274) and Plato (*circa* 427 BC - 347 BC), saying that three things are necessary for beauty: integrity, harmony and luminosity. Integrity refers to observing and perceiving a thing by the whole and not by parts. Once a thing is identified, the apprehension analysis follows: it is apprehended as a complex, multiple, divisible, separable, composed of its parts, and harmonious in the result, in the sum of them, i.e., harmony. Finally, luminosity is the force that allows us to generalise, converting images into universal æsthetics.

If we follow the logic of Le Corbusier, then we can argue that meeting needs lead to happiness. Abraham Maslow (1908-1970) was of a similar mind, for in his seminal work *A theory of human motivation* (Maslow, 1943), he presents his hierarchy of needs. Like a pyramid (see Figure 1.1), human needs are represented in ascending order of importance, from the base (physiological needs) to the top (self-actualisation).

Now, what does all this have to do with Computer Science? If we return to the premise of Le Corbusier, we can begin to discuss that many applications of Computing are beautiful because they respond to needs and sometimes make people happy. Moreover, although we have already very briefly defined these terms in literary, philosophical and psychological forms, it is clear that there is no simple way to approach them. In this way, how do we know in the strict context of Computer Science when an artifact, i.e., software,

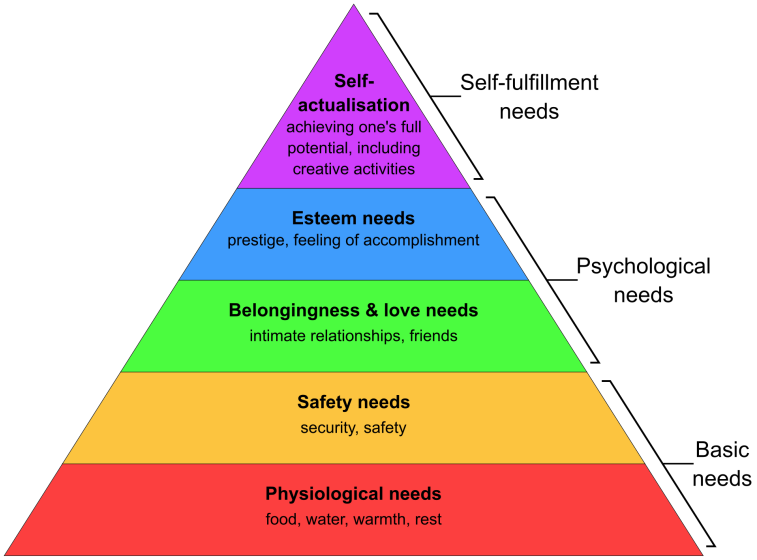


Figure 1.1: Maslow's hierarchy of needs (Maslow, 1943).

hardware or a combination of both, is useful, beautiful and makes people happy? To answer these questions is the work of Human-Computer Interaction (HCI).

1.1 MOTIVATION

Epistemology is the branch of Philosophy concerned with knowledge. Epistemologists study the nature, origin, scope of knowledge, epistemic justification, the rationality of belief, and various related issues. Epistemology is considered a major subfield of Philosophy and other major subfields such as Ethics, Logic, and Metaphysics. Epistemology aims to answer questions such as *What do we know?*, *What does it mean to say that we know something?*, *What makes justified beliefs justified?* and *How do we know that we know?* (Stroll & Martinich, 2021; Wenning, 2009).

We already stated that the question "how to know if an artifact is beautiful, useful and makes people happy?" corresponds to HCI as a Computer Science branch. Furthermore, as a discipline of HCI, and in the context of this work, that question is in the domain of User eXperience (UX). How do we know what we know in UX? using the techniques, tools and evaluation methods that exist (see Sections 2.3 and 2.5).

However, most works talk about UX by assessing an existing artifact, so evaluations are measured **during** and **after** the experience. In this way, Anticipated User eXperience (AUX), i.e., evaluations prior to a functional prototype or finished product, have been relegated (Yogasara et al., 2011).

As an important note, and before delving further into the subject, this work should not be confused or related to Epistemic modal logic or any formal representation of knowledge (Alvarado & Esquer, 1997; Alvarado & Sheremetov, 2001).

1.2 RESEARCH CONTEXT

According to the 2012 ACM Computing Classification System [1], our work is within the **Human-centred computing** classification, as a higher level category and as more specific categories, it can fall within: *User studies*, *Usability testing* and *Heuristic evaluations* (see Figure 1.2).

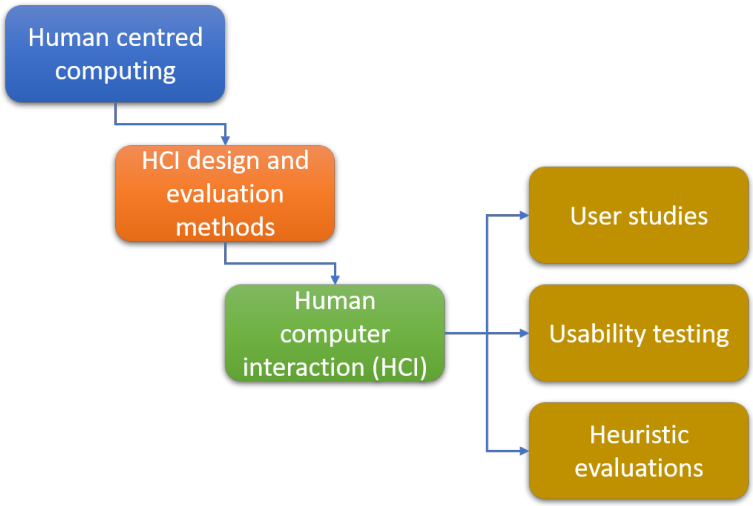


Figure 1.2: ACM 2012 classification.

1.3 PROBLEM STATEMENT

As we already mentioned, the problem we have is the lack of studies focused on AUX. In this way, we will explore various evaluation

1 https://dl.acm.org/ccs

methods to answer *how we know what we know about AUX?*. For this, our research will be oriented in three approaches: tasks, users and heuristics. In this way, we can obtain some knowledge similar to that which already exists in UX studies (see Figure 1.3).

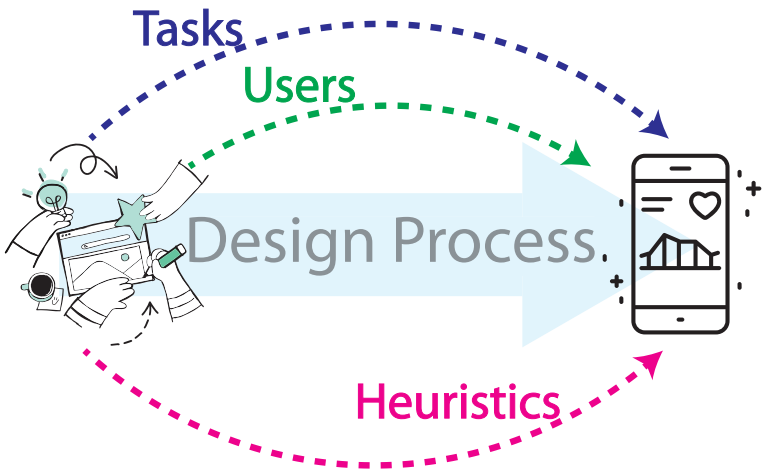


Figure 1.3: Increase what we know about AUX to help the design process.

1.4 OBJECTIVES

General

To develop AUX methods that help improve the design process of Graphical User Interfaces (GUIs).

Specifics

- To define an evaluation method that studies the expectations of users in basic tasks of an artifact.
- To create an artifact with requirements and beliefs of users.
- To produce heuristic evaluations with previous experiences in the state of the art.

1.5 CLASSIFICATION OF PROPOSALS

All our design and evaluation proposals are focused on obtaining more knowledge about AUX. However, to obtain more solid know-

ledge, it has to come from various sources. All our work is classified, according to its nature, into three groups:

- **Task oriented:** We seek to study the expectations of users to perform basic tasks in various environments[2].
- **User oriented:** We collect the needs and beliefs of users to create a tool.
- **Heuristic oriented:** We compiled state-of-the-art empirical knowledge so that it could help to carry out expert evaluations.

In this way, this helps us not only to organise and present our proposals, but also to the Related Work and Conclusions (see Figure 1.4).

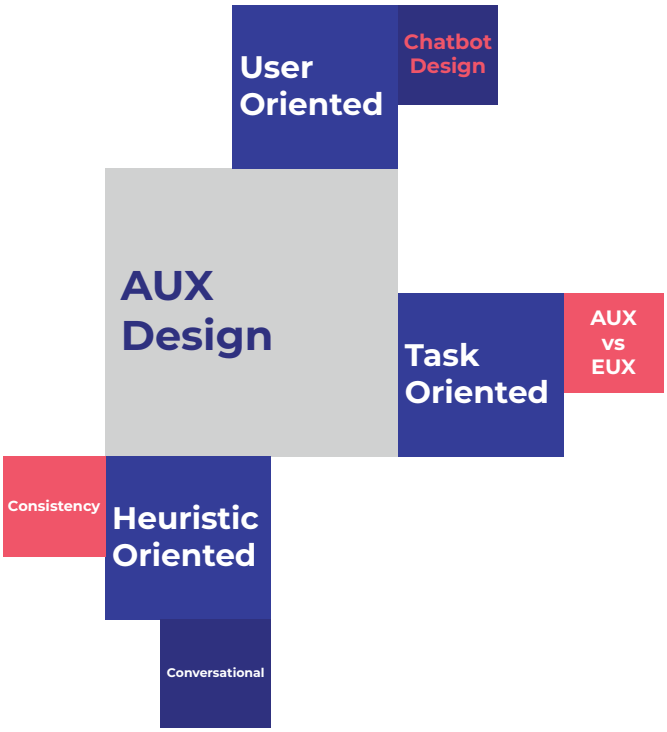


Figure 1.4: AUX Design Proposals.

2 Throughout this work *environment* and *context* are synonymous.

1.6 THESIS STRUCTURE

This document consists of ten chapters (see Figure 1.5). After having presented the motivation for the project and having raised the problem to be solved and the objectives to be pursued, the Theoretical Framework is exposed in Chapter 2, which is the theoretical foundation of our proposals. Next, Chapter 3 sets out the Related Work. Then the Research Methodology is explained in Chapter 4. Chapters 5, 6, 7 and 8 are our design proposals in AUX, classified according to their orientation. Afterwards, Chapter 9 contains the general Discussion of our research. Finally, in Chapter 10, we present our Conclusions and some ideas for Future Work.

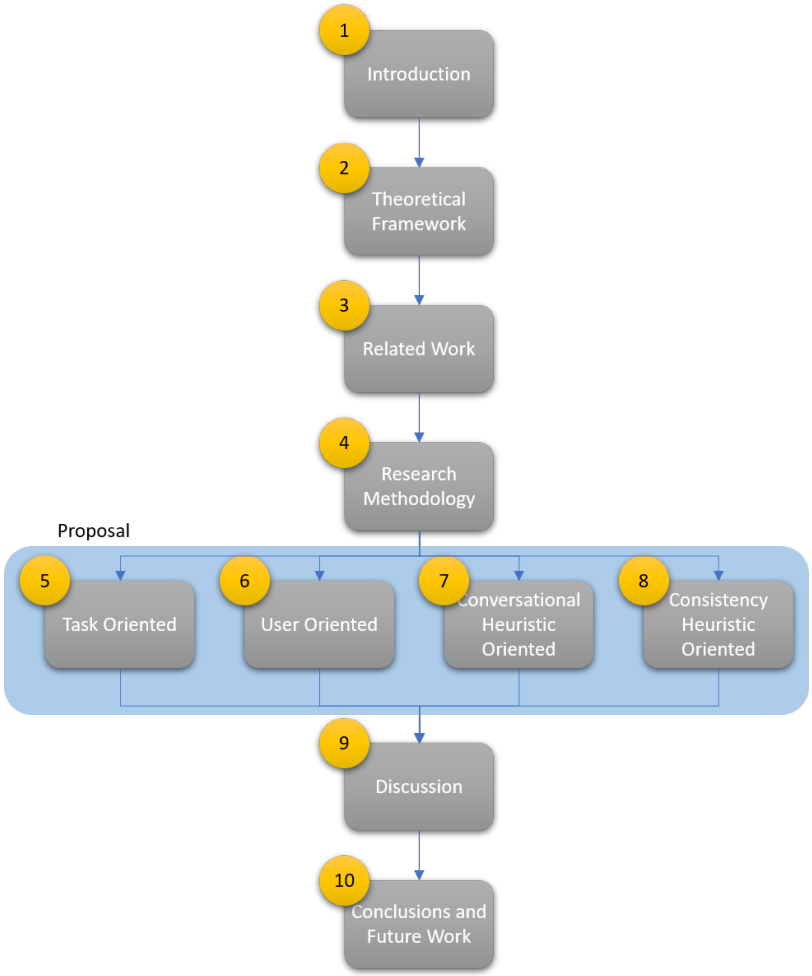


Figure 1.5: Document Organisation.

2

THEORETICAL FRAMEWORK

This Chapter contains the essential theoretical foundations necessary to understand the context and development of the thesis work. The content is related according to three epistemic knowledge questions: What is studied? How do we study it? Furthermore, Where do we study it? In this way, as shown in Figure 2.1, the Sections of the Chapter answer these questions.

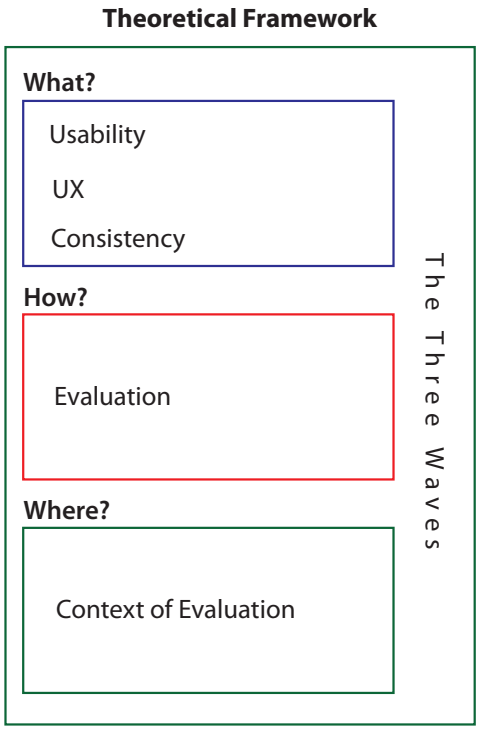


Figure 2.1: Theoretical Framework organisation.

2.1 THE THREE WAVES

The HCI area is not new. How it originated and where it comes from is debatable, but what we can agree on is that it draws on areas such as Human Factors, Ergonomics, Design in its multiple variants,

from graphic, through industrial, software and even architectural, Physiology, Philosophy and Cognitive Sciences (Grudin, 2018).

Tracking the evolution of HCI is fascinating and much more understandable thanks to the "Three Waves" scheme proposed by various authors (Bannon, 2011; Bødker, 2006, 2015; Duarte & Baranauskas, 2016; Rogers, 2012).

The first wave (*circa* 1948 - 1979) consists of the work of electrical and electronic engineers at the dawn of computing (see Figure 2.2). This stage focuses on Human Factors and Ergonomics, favouring concrete problems and performance metrics, e.g., studying whether a pilot can manoeuvre a new and complex system of controls without errors (Wiener, 1989). In this wave, pragmatic results were sought without emphasising theoretical aspects (Harrison et al., 2007).

Figure 2.2: Grace Hopper (1906-1992) was one of the most important characters in the first wave (Image courtesy of Jan Arkesteijn).

The second wave (*circa* 1980 - 1999) was a paradigm shift and is marked by the entry of Cognitive Sciences (see Figure 2.3). This represented a revolution, as the theory gained leadership, and the focus was on the human mind in terms of information processing, e.g., studying how the human mind processes information from a machine and how it communicates through a user interface (D. Norman, 2002). In this wave, theoretical foundations of cognition and activity were sought (Bødker, 2015).

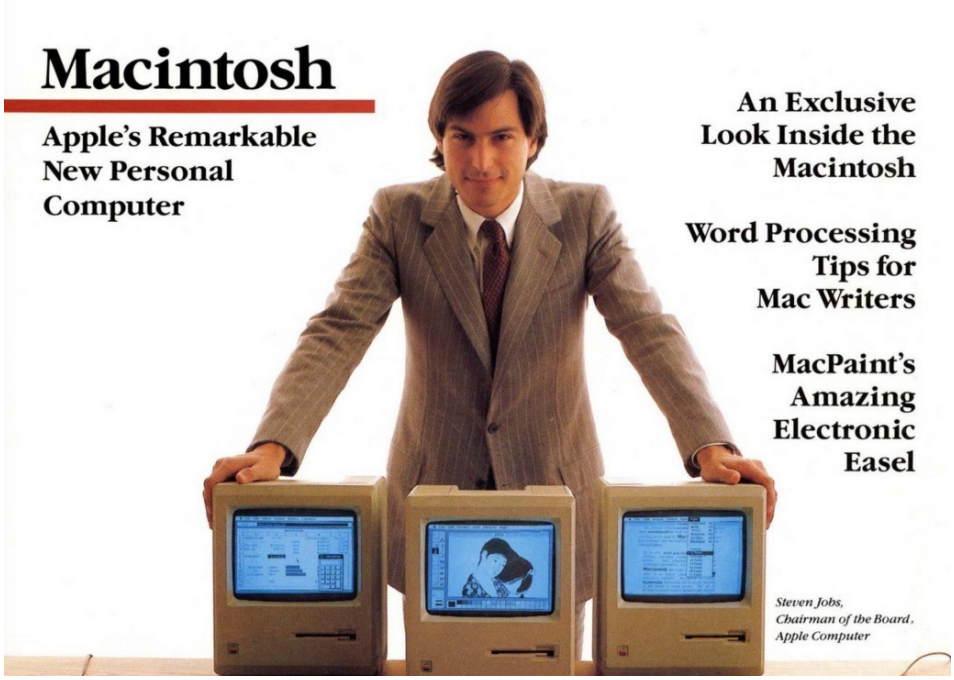


Figure 2.3: The birth of the personal computer marked the second wave (Poole et al., 1984).

The third wave (*circa* 2000 to present) brings with it previously displaced and forgotten elements, such as culture, values, and the role of the evaluator, e.g., studying how to reintroduce the humanities to HCI to stimulate social change (Bardzell & Bardzell, 2016). This wave seeks to move away from productivity and purposeful interactions to seek more pleasure and innovation through new technologies (Bødker, 2015).

It is in this third wave that we currently find ourselves (see Figure 2.4). There are experts like Bødker (2015) and Grudin (2018) who argue that we are probably on the edge of starting the fourth wave. What will this new stage consist of? There is no certainty, although it is most likely very intertwined with AI (Harper, 2019).

The importance of the wave change is in its epistemic paradigm, i.e., how knowledge was obtained in each period. This, of course, depends on the context of each era, the available methods, who were the users and evaluators of each era's systems, the objectives that were pursued, and the factors that limited them (Kaye, 2007). Table 2.1 summarises these facets for each wave.

Figure 2.4: Technology as a means of expression belongs to the third wave (Geffen, 2016).

Time frame	Who are the users?	Who are the evaluators?	What are the limiting factors?	Which is the paradigm?
1948 - 1979	Engineers and mathematicians	Engineers	Reliability	Human factors
1980 - 1999	White collars	Usability professionals	Time of the worker accomplishing their jobs	Usability
2000 - Present	People choosing to use technology	Us, and designers, and writers...	How to express oneself	UX

Table 2.1: HCI Waves.

How important is it to study HCI waves? As in any other discipline, history is the common thread that allows us to see the changes over time. In this case, the waves represent the paradigms that allowed studying the interaction that human beings have had (and have) with technology. In this way, it is clear that the first two stages evolved thanks to usability, while the third is about user experience. Although both terms are essential for the understanding of this thesis and, although close, they should never be confused with each other, that is why we dedicate a section to each one to study them in greater detail.

2.2 USABILITY

"Usability is about human behaviour. It recognises that humans are lazy, get emotional, are not interested in putting a lot of effort into, say, getting a credit card and generally prefer things that are easy to do vs those that are hard to do." -David McQuillen, ex-Swiss banker and founder of Sufferfest cycling workout resources.

Until the early 2000s (beginning of the third wave), usability was the decisive criterion to determine how well was the interaction with a product or a service. Usability is based on tasks, goals, and performance and is measured in terms of efficiency, e.g., error rate and mental workload. Effectiveness, e.g., task completeness. And satisfaction, e.g., qualitative and quantitative attitudes (Jordan, 1998).

The best-known definition of usability is proposed by the ISO (9241-11: 2018): "The extent to which specific users can use a product to achieve specific objectives with effectiveness, efficiency and satisfaction in a specific context of use." This definition encompasses three main axes: **1) specific users:** not just any user, but the specific user for whom the product was designed; **2) specific objectives:** specific users have to share an objective for the product, i.e., that their purpose is the objective of the product; **3) specific use context:** the product has to be designed to work in the environment where users will use it (Barnum, 2011a).

There are many other important quality attributes. A key one is utility, which refers to the design's functionality: Does it do what users need? Usability and utility are equally important and determine whether something is useful: It matters little that something is easy if it is not what the user wants. It is also no good if the system can hypothetically do what the user wants, but the designer cannot make it happen because the user interface is too difficult. In this way, we have that (Nielsen, 2012):

- **Utility** = whether it provides the features the user needs.
- **Usability** = how easy and pleasant these features are to use.
- **Useful** = usability + utility.

On the Web, for example, usability is a necessary condition for survival. If a website is challenging to use, people leave. If the homepage fails to state what a company offers and what users can do on the site, people leave. If users get lost on a website, they leave. If a website's information is hard to read or does not answer users' key questions, they leave (Nielsen, 2012).

Usability is an essential part of the UX, and they are complementary concepts but never interchangeable (Petrie & Bevan, 2009). Thus, it is necessary to explore into the depths of UX.

2.3 USER EXPERIENCE

"No product is an island. A product is more than the product. It is a cohesive, integrated set of experiences. Think through all of the stages of a product or service – from initial intentions through final reflections, from first usage to help, service, and maintenance. Make them all work together seamlessly." Don Norman, inventor of the term User Experience.

Defining UX is a complex matter since there is no uniform explanation covering all contexts (E. Law et al., 2008; E. L.-C. Law et al., 2009). However, the ISO (9241-11: 2018) can be used again, which gives an excellent general concept of UX: "The perceptions and responses of a person resulting from the use or anticipated use of a product, system or service." To better understand this definition and get a glimpse of UX's complexity, it is best if we study the design tool proposed by Peter Morville, the *honeycomb* (Morville, 2005).

2.3.1 *Morville's Honeycomb*

Peter Morville is president of Semantic Studios, an information architecture and findability consulting firm. He has worked in the HCI area since 1994 and is considered one of the founding fathers of information architecture (Wikipedia, 2021).

The inspiration for the honeycomb comes from when he was working on topics related to AI. He wanted to find a point of balance between context, content and users (see Figure 2.5).

UX focuses on having a deep understanding of users, what they need, what they value, their abilities, and their limitations. It also takes into account the business goals and objectives of the group managing the project. UX best practices promote improving the user's interaction with and perceptions of the product and any related services. The UX honeycomb is a tool that explains the various facets of UX design (see Figure 2.6). Since there are many aspects of this field far beyond usability, Peter felt that this new diagram would help to educate clients. The honeycomb helps to find a sweet spot between the various areas of a good UX (usability.gov, 2014; Wesolko, 2016).

In this way, Peter defines these seven facets of UX as follows (Morville, 2016; Wesolko, 2016):

- **Useful:** A business's product or service needs to be helpful and fill a need. If the product or service is not useful or fulfilling

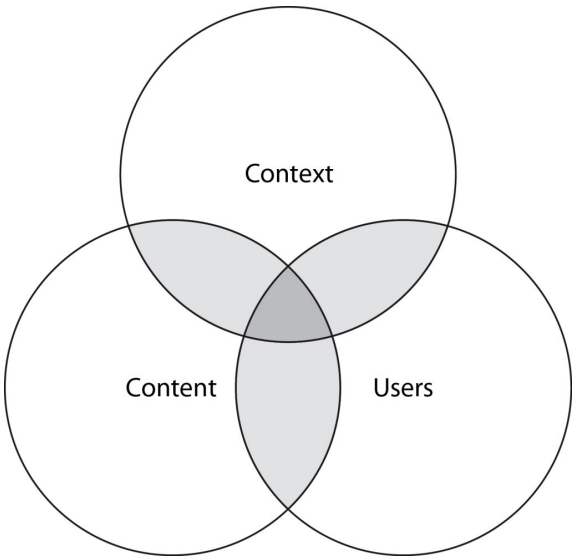


Figure 2.5: Finding the equilibrium among these elements is key to develop the UX design (Morville, 2016).

the user's wants or needs, then there is no real purpose for the product itself.

- **Usable:** The system in which the product or service is delivered needs to be simple and easy to use. Systems should be designed in a way that is familiar and easy to understand. The learning curve a user must go through should be as short and painless as possible.
- **Desirable:** The visual aesthetics of the product, service, or system need to be attractive and easy to translate. Design should be minimal and to the point.
- **Findable:** Information needs to be findable and easy to navigate. If the user has a problem, they should be able to find a solution quickly. The navigational structure should also be set up in a way that makes sense.
- **Accessible:** The product or services should be designed so that even users with disabilities can have the same UX as others.
- **Credible:** The company and its products or services need to be trustworthy.

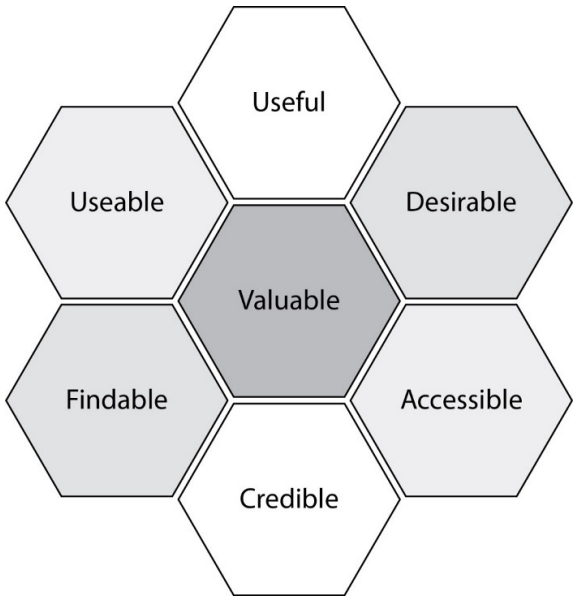


Figure 2.6: The UX honeycomb (Morville, 2016).

- **Valuable:** The product or services must deliver value to the sponsors. For non-profits, the UX must advance the mission. With for-profits, it must contribute to the bottom line and improve customer satisfaction.

This UX approach is essential for several reasons. The first is that it creates a landscape beyond usability. The second is that the modular design allows focusing on one element at a time to start on what is most critical or be carried out according to budget needs. Finally, each element is "a crystal with which you can see the UX", transforming the standard analyses, which helps the evolution of the UX (Morville, 2016).

Recently, the honeycomb received a small addition, where the components are reorganised according to three dimensions: thought, feeling, and use (Karagianni, 2018). This addendum can be seen in Figure 2.7

In this way, the seven elements of the honeycomb were grouped based on how the user interacts with a product (uses, thinks, feels). The elements were also rearranged within the honeycomb so that the relationship between them is visible. Finally, colour coding and labelling make the groupings clear. Therefore, the dimensions are depicted as follows (Karagianni, 2018):

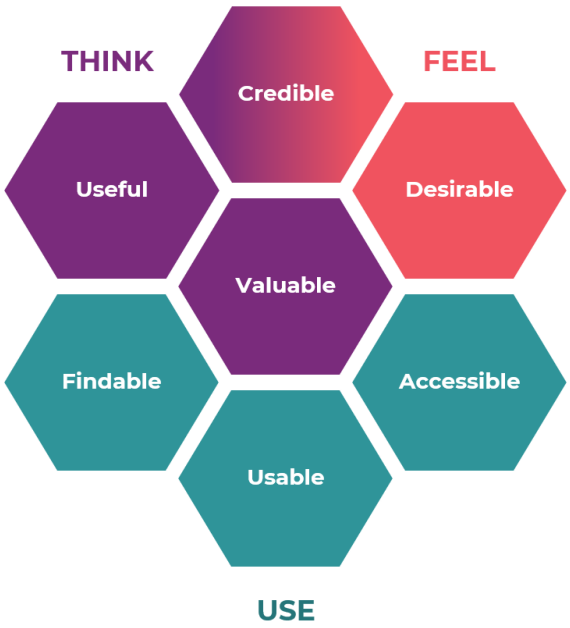


Figure 2.7: Karagianni (2018) proposed addendum to the honeycomb.

- **Think:** What do users think about the product? Is it useful? Is it valuable? Do they find it credible?
- **Feel:** How do people feel about the product? Do they find it desirable? Also, do they feel it is credible?
- **Use:** When it comes to actually using the product, it is findable, accessible and usable?

In addition to being interesting for integrating the honeycomb elements more comprehensively, this contribution is valuable because it allows us to glimpse the two prominent dimensions that make up UX: hedonism and pragmatism.

2.3.2 *Hedonism & Pragmatism*

As we have already seen, the Morville honeycomb allows us to understand the complexity of the elements that make up UX, the dynamics between them and the richness of each one. A more abstract classification of the various UX components distributes them between hedonic elements and pragmatic elements (Hassenzahl, 2007; Hassenzahl et al., 2000).

The hedonic components refer to the preferences, convictions, sensations, and conclusions of users that arise from the anticipated or episodic usage of a system, product, or service. The pragmatic components come from the features of the assessed system, such as functionality, interactive behaviour, supporting capabilities, usability, and performance (ISO, 2010).

In this manner we can classify the qualities of products and services accordingly (Hassenzahl, 2007; Hassenzahl et al., 2000):

- **Pragmatic qualities:** Related to practicality and functionality.
  - **Manipulation:** Refers to the functionality and how that functionality is accessed, i.e., the usability. At a very basic level, can it do what it needs to do? A consequence of pragmatic qualities is satisfaction. Examples of attributes that are typically assigned to websites (and software in general) are supporting, useful, clear and controllable. The purpose should be clear and the user should understand how to use it.
- **Hedonic qualities:** Related to the psychological needs and emotional experience of the user.
  - **Stimulation:** Users want to be stimulated in order to enjoy their experience with a product. Rarely used functions can stimulate the user and satisfy the human urge for personal development and more skills. Digital experiences can provide insights and surprises, e.g., if after a period of time a feature hasn't yet been used, the software could inform the user via a quick tip.
  - **Identification:** The human need for expressing ourselves through objects to control how you want to be perceived by others. We all have a desire to communicate our identity to others and we do this through the things we own and the things we use. They help us to express ourselves; who we are, what we care about and who we aspire to be. This is why people enjoy using personalisation on sites such as Twitter. Changing our background wallpaper and header image, helps us to express ourselves.
  - **Evocation:** Which memories and feelings does the experience evoke? Evocation refers to the symbolic meanings

that the experience has on our memories and our background. The visual aesthetics of a product may remind you of a past experience. For example, a travel website with a background image of a beach, might bring back memories of a past holiday and all the feelings (most likely highly positive) associated with that experience. As we all have different experiences in our lives, what we feel when we look at an identical website will be unique to us, the individual.

So far, we have only mentioned usability and UX in favourable terms. However, it exists also through the negative feelings that a person may feel towards an artefact. For example, Marshall (2019) gives us a small sample from a UX test conducted for a convenience store (see Figure 2.8).

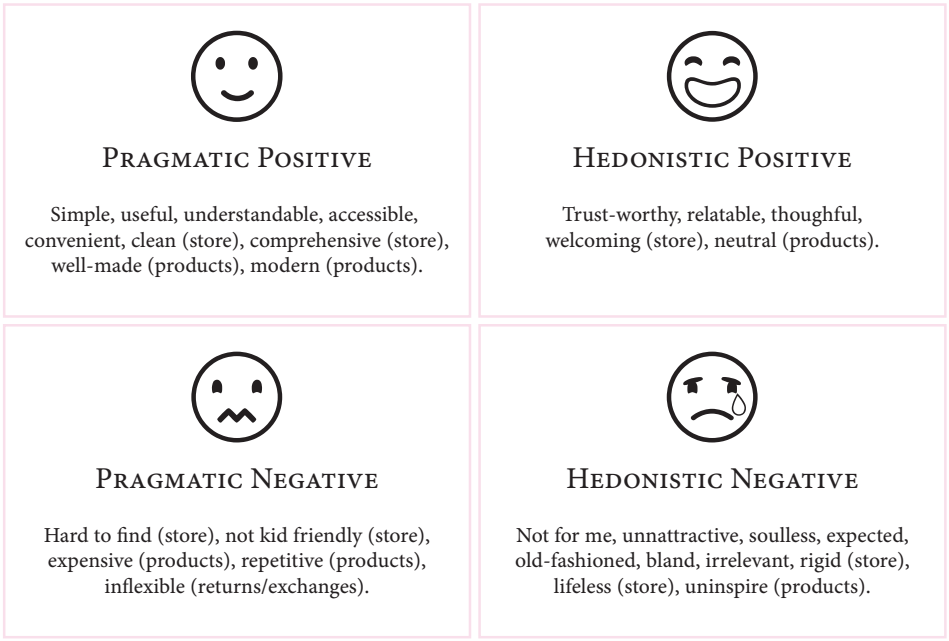


Figure 2.8: The results from the test can be classified as positive and negative in the hedonic and pragmatic UX model (Marshall, 2019).

While the core of UX is the current experience of usage, this is not enough to completely cover all the relevant issues that can be studied. UX is a highly dynamic concept, since it changes continuously when interacting with an artifact (Lallemand et al., 2015). People can have diverse and very different experiences before, during, and

after interacting with a product (Roto et al., 2011). Consequently, it is a critical design aspect to be able to measure the UX of an artifact at multiple times (Karapanos et al., 2010).

2.3.3 *UX & Time*

We can explain the concept of UX over time through four periods (see Figure 2.9). Each period is dynamic and can be viewed as an iterative process within and among those stages (Roto et al., 2011):

- **Anticipated User eXperience (AUX):** Obtained before the use of an artifact from imagination, expectations, and existing experiences.
- **Momentary User eXperience (MUX):** Perceived during the usage period of an artifact.
- **Episodic User eXperience (EUX):** Conceived after the use of an artifact through reflections of the experience.
- **Cumulative User eXperience (CUX):** Determined over time by the recollection of multiple periods of use.

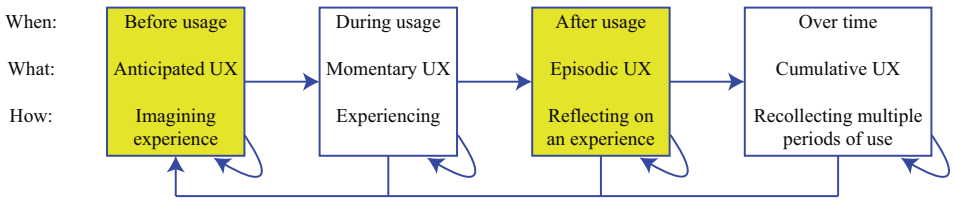


Figure 2.9: UX over time (Roto et al., 2011). In this work we focus on AUX and EUX (coloured in yellow).

Periods are essential because user responses may be different, e.g., when measuring momentary UX, it can result in a visceral response from the user. While if UX is measured some time after the use of an artifact, the user can remember more positive things and suppress the negative ones (Kujala et al., 2011). In this way, a study that considers more than one period could be more enriching.

While EUX is simply the experience that is obtained after having used a system, product or service (Winckler et al., 2016), AUX has to

do with attitudes and experiences that the user assumes to happen when envisioning using an artifact (Yogasara et al., 2011). Thus, the goal of an AUX assessment is recognising whether a determined idea offers the type of UX anticipated by developers for potential users Stone et al., 2005. Making AUX trials has been established worthily, even if there are not many research works on this subject (Bargas-Avila & Hornbæk, 2011; Karapanos et al., 2009; Roto et al., 2011; Vermeeren et al., 2010).

An example that can illustrate the various stages in the time of UX is a video game. When a person thinks to buy a game, expectations are created, e.g., how good it will be, the mechanics that it will implement, what is expected of the story and character development, i.e., AUX, because even if the opportunity has not yet been had to play, the possibilities are there. MUX would be when the person is just playing, and there they are experiencing first-hand what the game offers them, all the emotions they go through. When they reflect, forming an opinion after having finished the game, that is EUX, because once having completed the challenge or the story, they can create a judgment in hindsight. The value and contrast between each period are evident, e.g., the hype can be high, they can expect great things from the game, only to be disappointed once they are playing, because the product does not meet all the illusions of the players. However, once time has passed, that negative impression may change, and the game is remembered as not so bad or even good. Collecting and comparing all these stages is CUX.

As we already mentioned, UX and usability are not the same, but the latter can be considered as an element of the former. Following this line of thinking, another important element of UX is consistency.

2.4 CONSISTENCY

Consistency states that presentation and prompts should share as much as possible common features and refer to a common task, including using the same terminology across different inputs and outputs (Reeves et al., 2004). Several studies have shown that consistency is a crucial factor for multi-device experience, but they have also argued that it is a challenge for developers, since maintaining consistency of a multi-device system is an open problem (de Oliveira & da Rocha, 2007; Nichols, 2006; Pyla et al., 2006; Rowland et al., 2015). Consistency is important because it reduces the learn-

ing curve and helps eliminate confusion, in addition to reducing production costs (Grosjean, 2018; Nikolov, 2017; E. Wong, 2018).

Microsoft is an excellent case to exemplify the importance of consistency. Windows 10 and Office, in its most recent versions, are two of the most important products of the company; It is notorious that both GUIs are a design statement since they follow the same layout. In both software products, we can see that their toolbars have a similar design, i.e., the grouping, positioning, and labelling of buttons and commands is identical. This is intended to allow users to focus on their productivity, without the need to learn a new tool panel for each software they use. For this reason, Microsoft developed a series of tools, including an API and design guidelines that are integrated into a framework called Ribbon (Microsoft, 2018), so that this design discourse propagates to all applications developed by third parties (see Figure 2.10).

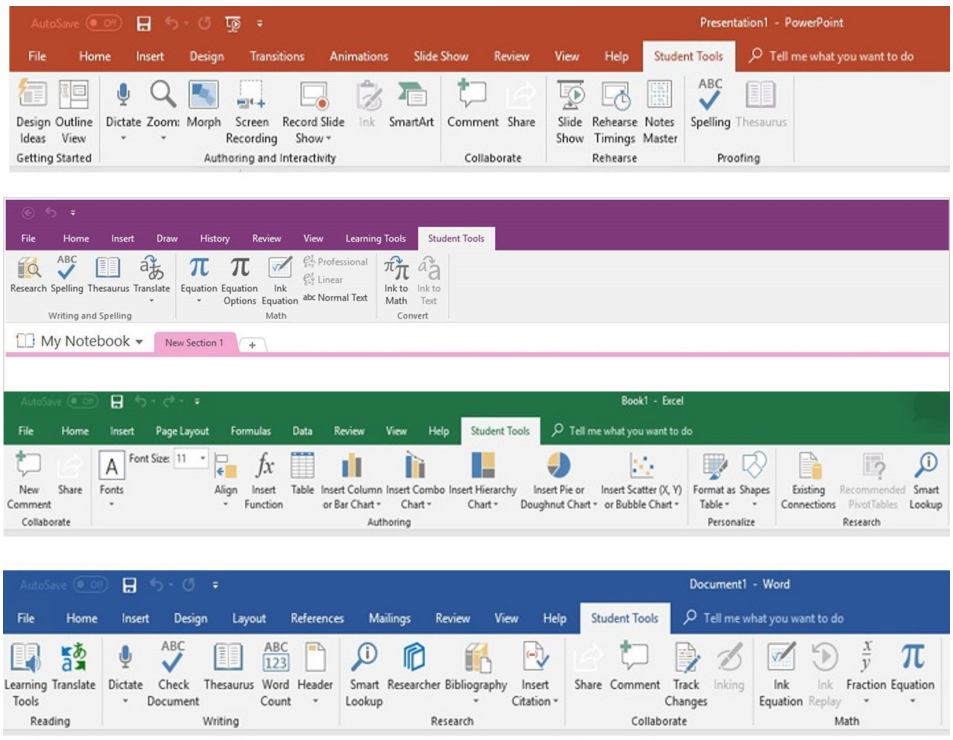


Figure 2.10: Consistent GUIs in Microsoft software.

In this way, we can talk about three approaches to the design of applications which, although authors like Coutaz and Calvary

(2008) and Vanderdonckt (2010) studied years ago, Levin (2014) summarises them in her 3C framework:

- **Consistent design approach:** Each device acts as a solo player, creating the entire experience on its own.
- **Continuous design approach:** Multiple devices handle different pieces sequentially, advancing the user toward a common goal.
- **Complementary design approach:** Multiple devices play together as an ensemble to create the experience.

This framework presents a series of challenges, since it involves, among other things, the fragmentation of the GUI and business logic. Thus the task of the developers is to preserve a positive UX among all the devices (see Figure 2.11).

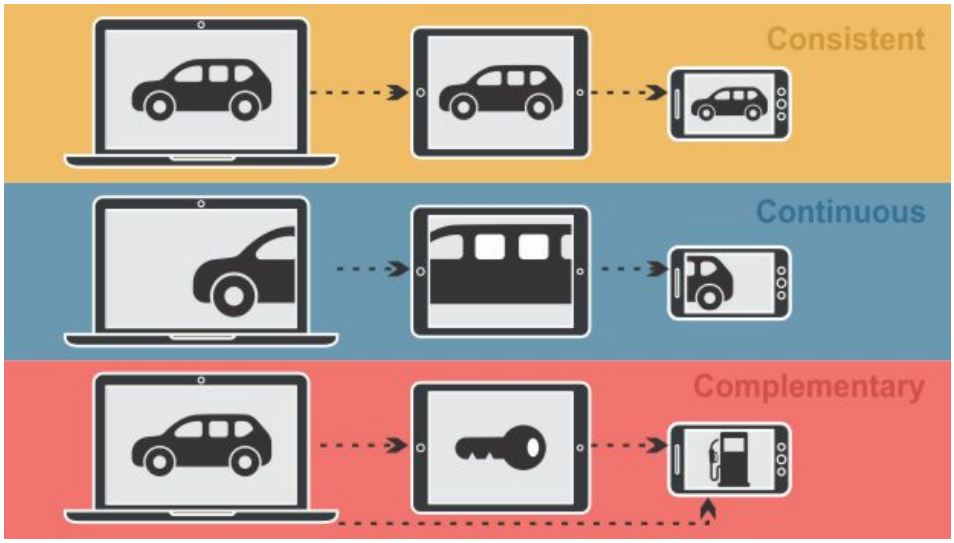


Figure 2.11: Levin (2014) 3C framework.

By adding consistency elements to the design of multi-device environments, usability is improved, and the possibility of a scenario with negative UX is reduced (Anić, 2018; Gaffney, 2018).

Having explained the concepts to be assessed, i.e., UX, usability and consistency, it is necessary to talk about evaluate them.

2.5 EVALUATION

The importance of usability and UX evaluations has been well established in state of the art (Ensina et al., 2019; Parlangeli et al., 1999; Sandars, 2010; Svaigen & Martimiano, 2018). The term "evaluation" is broad, as there are various tools and methods for conducting usability and UX evaluations.

For example, the general outline for an evaluation with surveys can be resumed in the following points (Geisen & Romano Bergstrom, 2017a):

1. Decide what aspect of survey to test.
2. Review for potential usability/UX problems.
3. Identify testing focus and concerns.
4. Determine where to conduct tests and what equipment to use.
5. Determine number and type of participants.
6. Choose testing approach and develop testing protocol.
7. Identify measurements to collect.
8. Recruit and schedule participants.
9. Conduct usability/UX tests.
10. Record observations, participant comments, and usability/UX metrics.
11. Debrief with observers.
12. Interpret data and diagnose problems.
13. Determine what to fix and how to fix it.
14. Report or present findings to stakeholders.
15. Repeat as needed.

As can be seen, evaluations are iterative in nature, to better understand this, we must explore the different types of tests that can be applied at various stages of development.

2.5.1 *Types of testing*

Usability/UX tests can be classified into three broad categories (Geisen & Romano Bergstrom, 2017b; Rubin & Chisnell, 2008):

- **Exploratory/formative testing.** This are tests that are done in the early stages of development. At this stage, most of the work is conceptual. They are used to test high-level designs before solving more complex and precise details. Identifying problems at this stage equates to saving multiple hours of design and programming.

  Marquis et al. (1998) suggest that at early stages of testing, the primary emphasis should be on evaluating the "interface design, arranging appropriate work sequences, and clarifying the meaning of words, icons, widgets, and other major features."

  In the formative tests, the focus of attention is the users, e.g., who are they? What tasks will they perform with the product? What do they think of the concept in general, and how does it compare to their mental model?

- **Assessment/summative testing.** Although this testing can happen at any point in the development cycle, it is usually done in development's early or middle stages, when prototypes exist for at least parts of the artifact. It evaluates users' actual behaviours how well people can actually use the product to complete a goal. It typically includes quantitative metrics as well as qualitative comments. Can provide insight on the high-level design or approach as well as a design's implementation.

  Assessment testing is usually conducted over several rounds with improvements made between rounds. Subsequent rounds evaluate the improvements or new aspects of the artifact as they are being developed. The quantitative metrics are tracked and compared across rounds, with the expectation that they will improve.

- **Verification/validation testing.** This usually occurs at the end of the development process just before the pilot test. The goal is testing how well the entire process works. The results of testing will be used to fine-tune and improve an existing design.

Not every project needs testing at every stage, while some projects will require multiple rounds of testing throughout (see Figure 2.12). The complexity of the project, budget, and schedule all factor into the amount of testing the development needs (Geisen & Romano Bergstrom, 2017b).

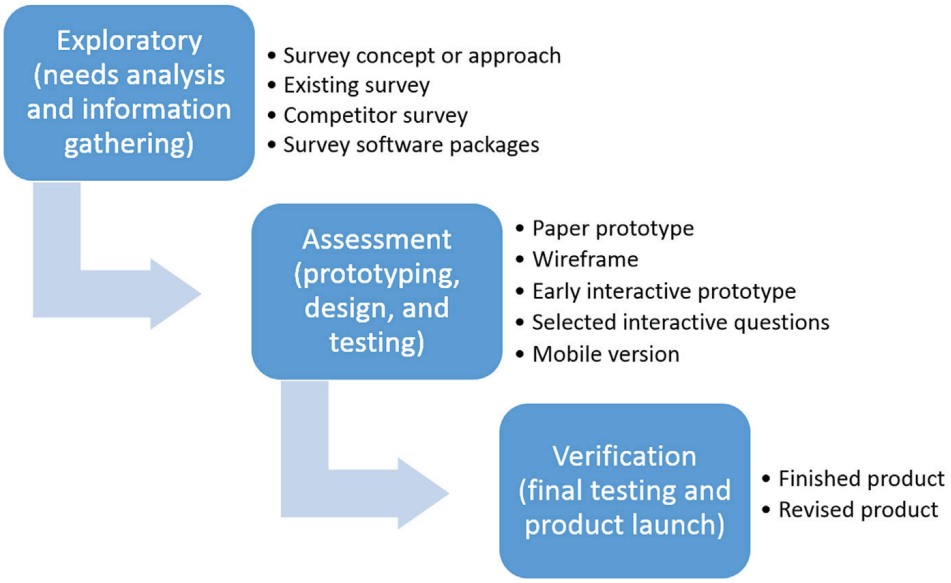


Figure 2.12: Usability/UX testing throughout the development cycle (Geisen & Romano Bergstrom, 2017b).

Different types of tests require participants, for this, we have to take into account various aspects according to what we want to test.

2.5.2 *Participants*

How many people we recruit often comes down to budget—how much time, money, and resources we have for the study. It is okay if one can afford to conduct only a quick study with a handful of users in a couple of rounds. Regardless of the total number of participants to recruit, we must recruit study participants who represent potential respondents—the people who would actually complete the test in the real world. To do this, we must determine what the target population is for the survey: Who should be included (within scope), and who should be excluded (out of scope) (Geisen & Romano Bergstrom, 2017c).

Nielsen and Landauer (1993) found that the number of unique usability problems found in a usability test conducted with n users can be predicted using Equation 2.1:

$$X = (N(1 - (1 - L)^n)) \tag{2.1}$$

Where X is the total unique usability problems, N is the number of problems known, L is the proportion of unique usability problems discovered by a single participant, and n is the number of participants.

Analysing the number of usability issues found across a large number of projects, Nielsen and Landauer found that, on average, the value of L was 0.31. That is, the average participant identified 31% of all usability issues identified in a given round of testing. The plot of the formula above with a value of $L = 0.31$ is shown in Figure 2.13 (Geisen & Romano Bergstrom, 2017c; Nielsen & Landauer, 1993).

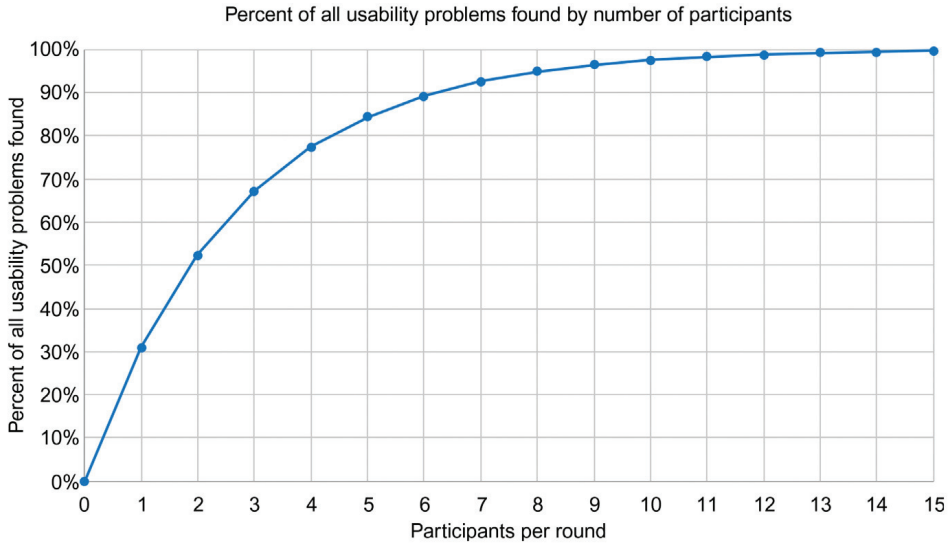


Figure 2.13: Percent of all usability problems found by number of participants (Geisen & Romano Bergstrom, 2017c; Nielsen & Landauer, 1993).

2.5.3 *User Centred Design Toolkit*

To follow a user-centred design, designers have a toolkit at their disposal according to the needs of the project. Although these tools are not the only ones that exist, they are the most recurrent in the design-evaluation process (see Figure 2.14) (Barnum, 2011b):

- **Interviews** — which can be structured, with a planned set of questions, or semi-structured, with some core questions that can start the conversation.
- **Shadowing a user for a day** — in which the evaluator follows the user around to understand "a day in the life" of the user.
- **Critical-incident technique** — which is used in situations where one can't observe people doing their job because it involves privileged information or is dangerous or it doesn't happen very often. Instead, one ask them to describe the situation or show how they do something in the situation.
- **Scenarios and role-playing activities** — in which the evaluator asks their target user to step into a situation and walk through what happens. This technique can be used in place of, or along with, the critical-incident technique. In some cases, one may want to play the part of the customer in the role play.
- **Card sorting** — a tool that is generally used early in development to learn users' preferences for and understanding of the information architecture of the artifact, as well as their understanding of the terminology. This activity can be done in person or remotely, using a web-based application.
- **Participatory design** — a development strategy that involves potential users in the design process. In some cases, these users are asked to review a product in development and provide feedback; in other cases, they are actively involved in generating design concepts.
- **Heuristic evaluation** — also called an expert evaluation, is an assessment or inspection of a product made by experts. Typically, this means usability experts, but it can also mean double expertise in usability and the product domain. Heuristics are

a set of general rules or principles used by experts to inspect a user interface for violations of these rules. It should be noted that a heuristic evaluation, like any other individual method, does not represent a complete usability evaluation, i.e., various tools are needed at different stages of development to have a comprehensive evaluation. Heuristic evaluations generally represent the first phase of testing a product or service.

- **Cognitive walkthrough** — another type of inspection, in which a team member, standing in for the user, walks through a prototype of the product to identify issues that affect ease of learning and related issues.
- **On-site usability testing** — testing that is done after the product has been released, to validate the usability in the user's environment. This type of testing is called field testing.
- **Server log data analysis** — an automated tool that runs behind the scenes and around the clock. This tool provides an analysis of a website and can generate a lot of data, such as pages visited, customer drop-offs, fluctuations in the volume of traffic, and so forth.
- **Longitudinal study** — testing that takes place over time through repeated contact with users.

It should be noted that although some of these tools explicitly mention usability, they are also valid for UX, only the points that are observed change.

Having explained so far what is measured (usability, UX and consistency) and how it is measured (evaluations), it only remains to explain where they are measured.

2.6 CONTEXT OF EVALUATION

This section explores the areas we focus on for our assessments, namely: social networks, chatbots, and home security systems.

2.6.1 *Social Networks*

The popularity of social networks has increased in recent years (Carta et al., 2020), especially due to the pandemic caused by COVID-

19 (Király et al., 2020; Wiederhold, 2020). However, they are not a new topic, much less unknown. Social networks have been studied by Computer Science researchers for a long time and from different angles, that is why we can find several definitions in the state of the art (L.-S. Chen & Chang, 2010; El Morr & Eftychiou, 2017; Lee et al., 2003; Preece et al., 2004; Y. Wang & Li, 2016). Among all, we adopted the one by boyd danah m. and Ellison (2007) "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site," because it denotes the main elements of social networks and their interaction.

Sociability and usability are vital factors for any social network. Sociability refers, of course, to the contact and exchange of information among users, whereas usability enables technology to allow those exchanges (V. H. H. Chen & Duh, 2009; Preece, 2000).

A result of sociability is participation. Therefore, several studies have been carried out in order to understand the motives of individuals to engage in a social network (Jacobsen et al., 2017; Nov & Ye, 2010; Tella & Babatunde, 2017; Zhou, 2011). We believe that the progressing of any social network heavily relies on: 1) the collaboration among its users to create contents and make contributions to the community (Lamprecht et al., 2016), and 2) the user interaction with businesses, organizations, colleagues, family members, and friends to create together their production and consumption experience and meet their necessities (Fragidis et al., 2010; Mai & Olsen, 2015; McCormick, 2010).

While it is true that UX is a crucial factor for interaction among users on any digital platform, it is not the only aspect to consider. Social networks are a complex phenomenon. Therefore, it is useless to oversimplify them and try to study them from a single front (Ling et al., 2005). For example, since the eighties, Grudin studied why collaborative work applications fail (Grudin, 1988). A perfect example that there is no formula for success is that of Google+. Despite having elements of good design, it never succeeded and ended up closing (Talin, 2019). Although UX is not the only aspect that should concern developers, it is essential to help make interactions among members of a social network as seamlessly as possible.

2.6.2 *Chatbots*

In recent years, there has been a growing interest in chatbots, which are programs that use Natural Language Processing (NLP) to interact with humans under different contexts (Shawar & Atwell, 2007). According to *Business Insider*, consumers are expected to spend up to $ 142 billion using these systems in 2024 (Intelligence, 2021). Chatbots can be found in various environments, e.g., business (Ravi, 2018), tourism (Dian Sano et al., 2018), FAQ (Ranoliya et al., 2017), procedures (Agus Santoso et al., 2018) and recommendations (Argal et al., 2018). Usually, chatbots work by searching for keywords, phrases or examples that they have stored in their knowledge bases, intending to offer information about products or services, activities or places, within social networks or websites (Ranoliya et al., 2017).

Social distancing, caused by the SARS-CoV-2 pandemic, has highlighted the vital importance of digital media (De' et al., 2020). In this way, chatbots have gained considerable notoriety, as they have been developed as tools to combat the pandemic itself (Miner et al., 2020), to provide customer service (Hao, 2020), as well as in many other commercial and governmental fields (Vergadia, 2020). This rampant demand for chatbots is not only because they are a tool that can offer immediate communication and automation of specific tasks, but also to the different technologies that allow rapid developments since functional chatbots can be obtained in a matter of hours (Luo et al., 2020).

Thus, in the education sector, we can find chatbots as a means to provide information about courses, procedures, and school services (Shaw, 2012). Nowadays, students receive a significant part of their education through online information, such as class topics, homework, and practices. For this reason, chatbots can provide valuable help in the teaching/learning process (Molnár & Szüts, 2018). Also, the development and use of chatbots begin to be of great interest to schools and universities (D-LABS, 2019; Talin, 2019).

2.6.3 *Home Security Systems*

Artificial Intelligence (AI), and Internet of Things (IoT) are very discussed topics today, although they are not new in Computer Science. Phenomena such as cheaper technology (Grosjean, 2018), and indus-

trial automation (Acemoglu & Restrepo, 2019), have caused many questions and problems to arise in both areas.

A particular branch of IoT with a significant growth is that of home security systems. This type of systems generally consists of a camera that streams video over the Internet, microphones, speakers, and a cloud platform from which users can remotely monitor their homes. With the promise of increasing quality of life, and improving the security of their properties, many users have adopted these systems (AlHammadi et al., 2019). However, they are controversial, since much has been investigated from the perspective of information security (Dey & Hossain, 2019), privacy (El-Moussa, 2018), and social psychology (Klobas et al., 2019).

A significant problem that is particularly emerging in IoT home surveillance systems is that users feel besieged in their own home (Gaffney, 2018). By integrating AI algorithms (e.g., detection of human forms) this type of systems sends alerts to the users' mobile devices, every time a movement is detected, creating a false sense of insecurity. Although crime levels in the USA have gone down (Nikolov, 2017), the perception of citizens does not match that data (Anić, 2018). Thus, users do not have clear information about when it is an actual alert, and when it is an error or a situation that does not require any measure. This lack of consistency can lead to situations of stress, anxiety, and unnecessary vigil (E. Wong, 2018).

All of the above poses a challenge for HCI researchers because the privacy notifications and settings of these systems must be exceptionally clear and convenient if they are to be used in real life (Zheng et al., 2018).

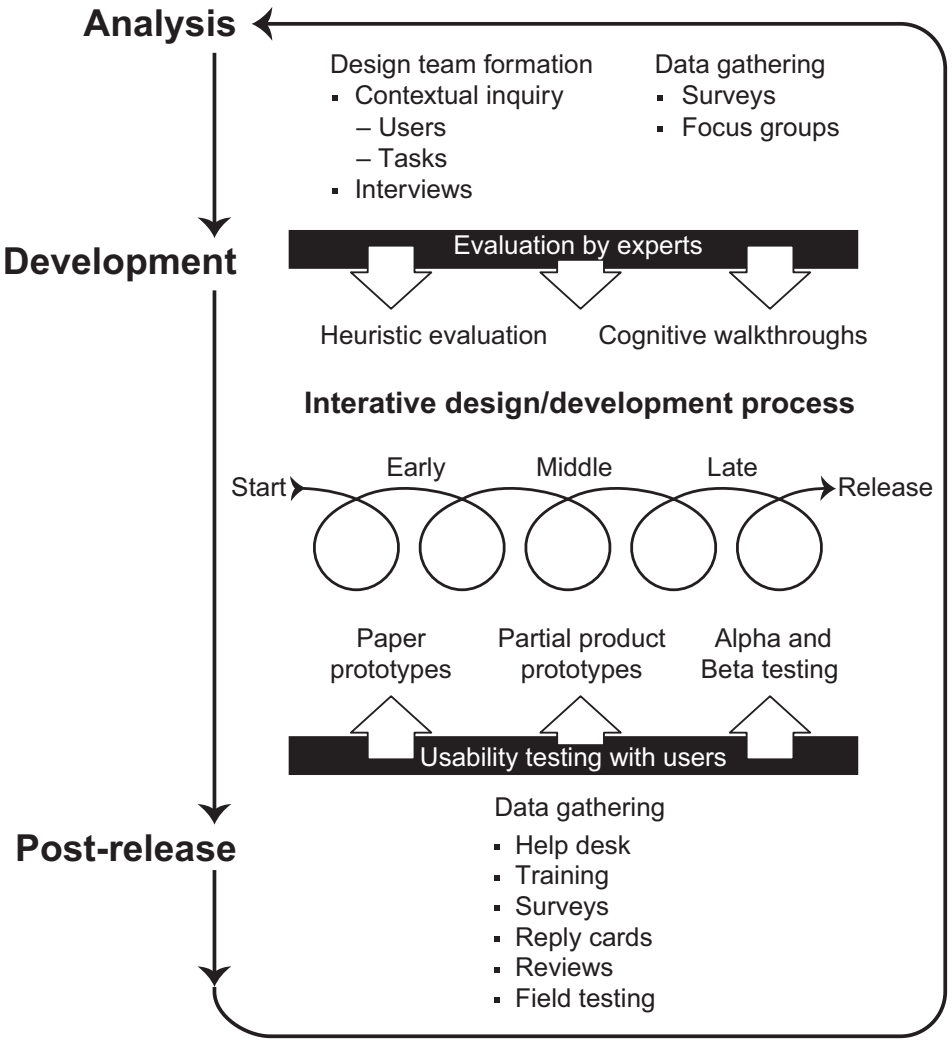


Figure 2.14: Information and user and expert feedback flow into product development with a user-centred design process (Barnum, 2011b).

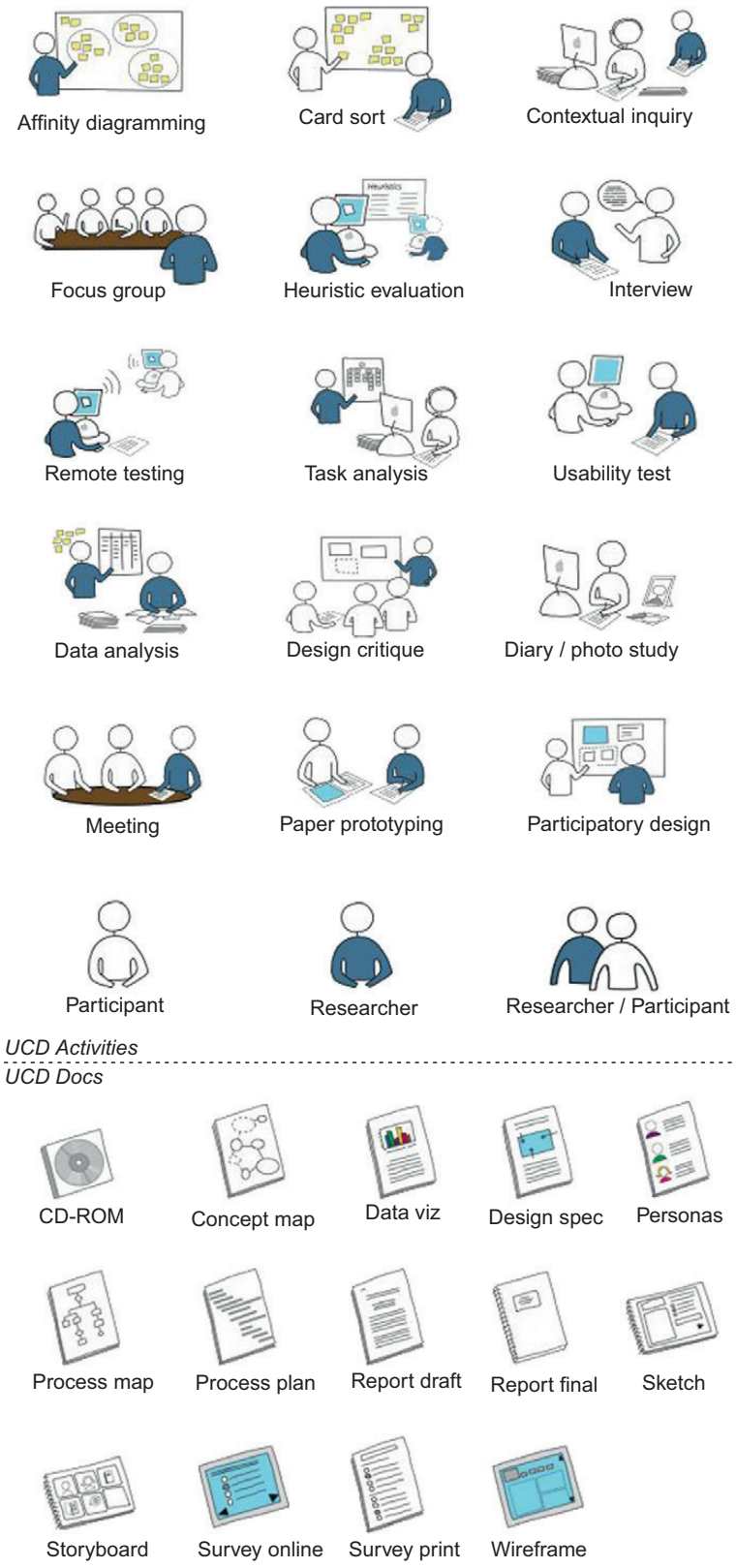


Figure 2.15: User Centred Design Toolkit (Barnum, 2011b).

3

RELATED WORK

This chapter compiles some works related to our proposals. Like our contributions, they are divided into three orientations: Task, User and Heuristics.

3.1 TASK ORIENTED

In this section, we present a brief review of some outstanding works that involve AUX and EUX. We classify them into these two groups because it seems that this is the trend in most UX work. The former group are researchers who study popular systems in the market and then propose theories (see Section 3.1.1). The latter group are those who, after studying theoretical works, use their knowledge to propose changes in practical systems (see Section 3.1.2). We consider that our work has a hybrid approach, trying to bring together the best of both paths.

3.1.1 *From Practice to Theory*

Practice is vital, as it allows collecting people's opinions and reactions. As they did Aladwan et al. (2019), who designed a framework through review searches and constructed a prototype that describes user anticipations and experiences, using instructional fitness applications. The main limitation of this work is the difficulty in unravelling ambiguous user reviews.

Although, in general, qualitative evaluations are complicated to analyse because they precisely lend themselves to ambiguities, they are an indispensable resource if the investigation is about transferring real-world interactions to a virtual environment. Such is the case of Moser et al. (2014) that organised workshops for children around the world. Through various types of activities, they managed to gather children's expectations and idealizations regarding games. Although they detailed the way to capture AUX,

they did not make comparisons, nor propose elements for the design of GUIs.

The works of Margetis et al. (2013), and E. Zhang et al. (2018) also fall into this area of gathering the users' know-how. As the former ones created an Augmented Reality (AR) system that facilitates reading and writing in books without being invasive to users. In addition to a heuristic evaluation, there is no evidence of AUX evaluation, only of EUX after testing the prototype. Whereas the latter authors designed a card game that encourages the practice of people who are learning a foreign language. Even though in their design they did an AUX study, there are no contrasts with EUX.

User expectations are also gathered when new environments are studied. For example, Kukka et al. (2017) investigated the integration of Facebook content in three-dimensional applications. They created design guidelines based on the problems they could identify in this kind of environment. Being a preliminary investigation, they not compared AUX vs EUX. Another example is Wurhofer et al. (2015) that examined in the context of UX motorists. Through a study of cumulative UX, they compared expectations against the real experiences of drivers. Despite this is a study of UX over time, it does not include GUIs.

3.1.2 *From Theory to Practice*

Theory is essential because it identifies and proposes elements that can be used to design and evaluate systems. Such is the case of Magin et al. (2015) that described possible factors that cause a negative UX using apps. Through a prototype app, AUX and EUX were measured by the participants. They concluded that the lack of usability causes negative emotions. Similarly, Sato et al. (2012) reported a series of elements used in multi-agent systems that can possibly be applied in Communities of Practice (CoP). Though the impact that these elements would have on UX can be deduced, they did not evaluate UX.

3.2 USER ORIENTED

The focus of this section is chatbots as tools to support the educational process. User oriented design is important to meet the needs of everyone involved in the process.

Basogain et al. (2017) reviewed the limitations of the current education systems, in particular, in the area of mathematics. They discussed a set of fundamental changes in curriculum and teaching methodology. Finally, they examined the role of e-learning as an integral part of the transition from traditional education systems to modern systems. While this work does not propose any work with chatbots, it makes clear that digital tools are valuable for the teaching-learning process. An example of these digital tools is the one proposed by Fonte et al. (2016), they implemented a system consisting of two parts, an Android application and a server platform. The Android application implements a chatbot which interacts with both the student and the server. The objective for the system was to enable the student to carry out several actions related to their studies like consult exam questions, receive recommendations about learning materials, ask questions about a course, and check their assessed exams. Although the architecture and characteristics of the system are presented, they did not perform tests with end-users.

An important branch of e-learning tools are chatbots, for this, Cunningham-Nelson et al. (2019) made an exploratory literature review of both chatbots in general use and chatbots in education. Two preliminary chatbots applications are presented; a Frequently Asked Questions (FAQ) chatbot for answering commonly asked student questions, and a short response quiz chatbot designed to facilitate and provide automated feedback based on student responses. They concluded that the chatbots provide a promising area for potential application in the future of education, as they have the capability of streamlining and personalising components of learning.

One example of chatbot with user oriented design is the one by Benotti et al. (2014), they designed their system to foster engagement while teaching basic Computer Science concepts such as variables, conditionals, and finite state automata, among others to high school students. The tests they performed show that a chatbot can be a valuable tool to interest and help students with school issues. Another one is presented by Clarizia et al. (2018), they

developed a chatbot architecture to manage communication and furnish the right answers to students. Their proposed a system can detect questions and gives the answers to students, thanks to the use of NLP techniques and the ontologies of a domain. Although tests were done with students, they consisted in determining the correctness of the chatbot's responses, and not of the experience or usefulness it had.

3.3 HEURISTIC ORIENTED

This section compiles some works that represent the importance of our heuristic proposals, as well as the gap that we try to fill with them.

3.3.1 *Consistency*

Marcus (1995) was a pioneer in the description of good practices to develop GUIs. He claims that the *organisation*, *economisation*, and *communication* principles help GUI design. The highlights are his four elements of consistency: 1) *internal:* applying the same rules for all elements within the GUI, 2) *external:* following existing conventions, 3) *real-world:* following real-world experience, and 4) *no-consistency:* when to deviating from the norm.

In a more recent period, and following this train of thought, we have a notable pair of works. After having interviewed 29 professionals in the area of interactive environments, Dong et al. (2016) identified three key challenges that have prevented designers and developers from building usable multi-device systems: 1) the difficulty in designing interactions between devices, 2) the complexity of adapting GUIs to different platform standards, and 3) the lack of tools and methods for testing multi-device UX. The other one by Woodrow (2016) defines and contextualises three critical concepts for usability in multi-device systems: 1) *composition:* distribution of functionality, 2) *consistency:* what elements should be consistent across which aspects, and 3) *continuity:* a clear indication of switching interactions. He makes a call for more active involvement by both the systems engineering and engineering management communities in advancing methods and approaches for interusability,

i.e., interactions spanning multiple devices with different capabilities.

Similarly, there are specialised studies on multi-device systems. Meskens et al. (2010) presented a set of techniques to design and manage GUIs for multiple devices integrated into *Jelly*, a single multi-device GUI design environment. *Jelly* allows designers to copy widgets from one device design canvas to another, while preserving the consistency of their content across devices using linked editing. O'Leary et al. (2017) argue that designers of multi-device UX need tools to better address situated contexts of use, early in their design process through ideation and reflection. To address this need, they created and tested a reusable design kit that contains scenarios, cards, and a framework for understanding tradeoffs of multi-device innovations in realistic contexts of use.

3.3.2 *Home Security Systems*

An interesting work that shows us an overview of security systems is done by Mäkinen (2016), she examined why and how home surveillance systems are used and what the meanings and implications of these systems are to the residents. Through a series of interviews, she discovered that being under surveillance, especially in the privacy of one's own home, can evoke positive and negative feelings simultaneously. This is an exploratory study where possible solutions to the problems raised are not provided. In the same topic, Urquhart and Rodden (2017) presented a series of critical challenges to consider for the regulation of domestic IoT. They argue that novel regulatory strategies can emerge through a better understanding of the relationships and interactions between designers, end-users and technology. This is a discussion/position paper with no experiments.

With the popularisation of smart homes, several studies have emerged in this regard, we highlight a couple of them. Zeng et al. (2017) conducted semi-structured interviews with fifteen people living in smart homes to learn about how they use their smart homes, and to understand their security and privacy-related attitudes, expectations, and actions. Although their interviews provide guidelines for future work, they are only general aspects of those that recommend a thorough investigation. Shehan and Edwards (2007) discussed a range of usability issues with home networking,

as well as the sources of many of these issues. They contend that these problems will not disappear over time as the networking industry matures, but rather are due to structural usability flaws inherent in the design of existing network infrastructure, devices, and protocols. While this study does not address security systems, it is a vision of what HCI can bring to Do It Yourself (DIY) systems.

With all these problems, there are some proposals to face them, like the one from Alshamari (2016), he explored the differences between usability factors and aspects related to security and privacy. He developed some basic guidelines for reducing the gap between usability and security, as well as frameworks and some models for the same objective. However, his study is theoretical, and no tests of any kind were made.

3.3.3 *Conversational Systems*

Usability in chatbots is a topic that has started to become popular in recent years, but there are still many challenges in the area. This is how Valtolina et al. (2020) presented a study highlighting the benefits of developing a conceptual model based on a conversational interaction style to allow users to communicate with a system in a familiar way that works for them. Although they evaluated various applications in health and home automation and identified open problems, their analysis is relatively superficial since they used well-known general-purpose questionnaires in state of art. Likewise, Ren et al. (2019) conducted a review of the state of the art that deals with usability evaluation techniques in chatbots. They identified various elements to consider for a satisfactory evaluation within the classic components of usability: effectiveness, e.g., completeness of tasks and certainty. Efficiency, e.g., time to complete a task and mental effort. And satisfaction, e.g., ease of use and context-dependent questions.

One of the most important areas is how to study usability in chatbots, for this Kocaballi et al. (2019) studied various UX questionnaires to assess their coverage when evaluating chatbots. First, they found that many studies used the concepts of UX and usability interchangeably, skewing the results. Afterwards, they studied how many and which dimensions the questionnaires covered. They concluded that, although several questionnaires are used to evaluate a chatbot, it is not enough to carry out a complete evaluation since

not all the relevant dimensions for conversational systems were identified.

In such a way that there are works that make clear the lack of evaluation tools in chatbots, e.g., Ding et al. (2019) developed a panel that discussed how traditional means of interaction (e.g., websites) are migrating in chatbots and all that this represents, e.g., research focused on the user, usability tests, information architecture and technical development skills. The authors argue that a heuristic evaluation is a simple but effective tool for conducting usability evaluations in chatbots. And the case of Holmes et al. (2019), that evaluated, with various tools, a chatbot that helps with weight management. Through the application of two questionnaires, interviews and tasks, they detailed the level of convergence of the questionnaires, concluding that non-traditional methods are required to evaluate the usability of a chatbot since the questionnaires did not consider all the possible dimensions to be evaluated.

And of course, we can't ignore the context. Chatbots in education, for example, not only need usability evaluations, but also need to verify that they are the correct tool. For instance, Yin et al. (2020) investigated the impact chatbots have on college students' learning, motivation, and performance. They evaluated, in one session, two groups from an introductory computing class: the first had a traditional session, i.e., a face-to-face class, and the second group used a chatbot. The results showed that both groups obtained a similar performance, showing that chatbots can be integrated into a class since they help motivation and student performance. Furthermore, Rafael et al. (2019) conducted a study involving two student groups from the Chilean tax system. One group did active learning activities in a study session, while the other group used a chatbot. Thanks to the results of an open-ended questionnaire, which they applied before and after the session, they discovered that the group that used the chatbot performed better.

Figure 3.1 represents the relationship between the related works and the orientations of our proposals, as it outlines motivations and limitations from all the studies.

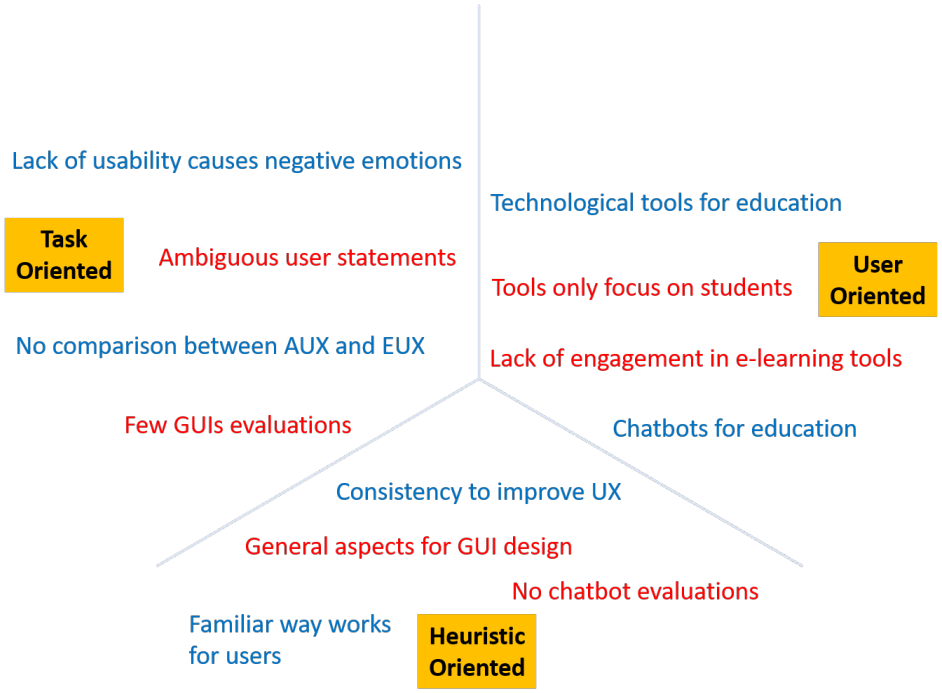


Figure 3.1: Motivations (blue) and limitations (red) from the related works.

4

RESEARCH METHODOLOGY

This chapter focuses on the methodologies we use during the development of our proposals.

4.1 DESIGN SCIENCE RESEARCH METHODOLOGY

As a guide to carry out our research, we use the Design Science Research Methodology (DSRM) process model by Peffers et al. (2007). This methodology was selected because it has been used in works that are under the same UX and usability study spectrum. For example, Carey and Helfert (2015) used this methodology to develop and validate their interactive evaluation instrument, whose goal is to improve the process for mobile service innovation. Strohmann et al. (2019) followed DSRM to create recommendations for the representation and interaction design of virtual in-vehicle assistants. Lastly, Kumar and Chand (2018) used this methodology to design an app that provides remote students with a learning support.

The DSRM iterative process consists of a research entry point and six stages (Peffers et al., 2007):

1. **Problem identification and motivation:** Define the specific research problem and justify the value of a solution.
2. **Define the objectives for a solution:** Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible.
3. **Design and development:** Create the artifact. Such artifacts are potentially constructs, models, methods, instantiations or new properties of technical, social, and/or informational resources.
4. **Demonstration:** Demonstrate the use of the artifact to solve one or more instances of the problem.
5. **Evaluation:** Observe and measure how well the artifact supports a solution to the problem.

6. **Communication:** Communicate the problem and its importance, the artifact, its utility and novelty, the rigour of its design, and its effectiveness to researchers and other relevant audiences such as practising professionals, when appropriate.

As for the initiation point, it could be (Peffers et al., 2007):

- **Problem-centred:** It is about solving a specific and well-defined problem.
- **Objective-centred:** It denotes the achievement of a goal, generally it is about improving a previous solution.
- **Design-and-development-centred:** It marks the start-up of a plan already developed for a specific scenario.
- **Client/content-centred:** It specifies the case of a project initiated by a client.

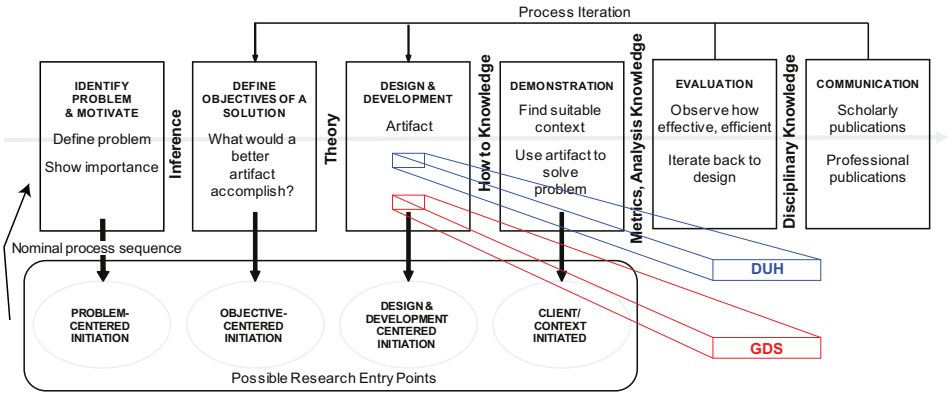


Figure 4.1: The six stages of the DSRM (Peffers et al., 2007).

The flexible nature of this methodology makes it possible to only use certain stages depending on the research needs, i.e., it is not necessary to always have to go through stages 1-6, nor is it necessary to always start from step 1 (Peffers et al., 2007).

The DSRM served us as a higher level methodology, i.e., as can be seen in the *Design & Development* stage of Figure 4.1, we use less abstract methodologies focused on more concrete results to achieve the desired artifact in each case.

4.2 GOOGLE DESIGN SPRINT

Google Design Sprint (GDS) is a time-constrained, five-phase process that uses design thinking with the aim of reducing the risk when bringing a new product, service or a feature to the market (Banfield et al., 2015):

1. **Understand:** Participants evaluate the problem they are trying to solve, the personas they are designing for, and the form factor they are going to use.
2. **Diverge:** Participants are encouraged to let go of all their presumptions and engage in different activities to generate as many ideas as they can, regardless of how feasible or far-fetched they are.
3. **Decide:** Through different activities, participants decide which ideas to pursue further.
4. **Prototype:** Participants rapidly sketch, design and prototype their ideas, concentrating on User Interface (UI) flow.
5. **Validate:** Participants put their product in front of users, test and are encouraged to show and tell when possible.

The process aims to help teams to clearly define goals, validating assumptions and deciding on a product roadmap before starting development (see Figure 4.2) (Banfield et al., 2015).

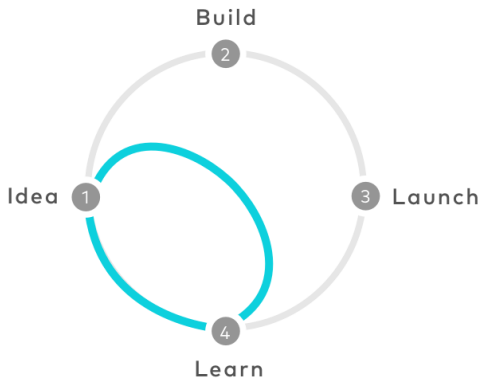


Figure 4.2: The sprint gives teams a shortcut to learning without building and launching (Banfield et al., 2015).

Although it is not a research methodology *per se*, we decided to adopt it because it is a way to create valuable products, i.e., that they are not only usable and aesthetically pleasing but that generate a change of skills and a way of thinking (Sari & Tedjasaputra, 2017), as it is that we finally look for with our proposal. This technique, in particular, seeks to reach a viable solution within five days, this is something we decided to modify because we were interested in user-oriented design focus, but we considered that the time would constrict us.

4.3 DEVELOPING USABILITY HEURISTICS

Developing Usability Heuristics (DUH) is a methodology to create heuristics proposed by Quiñones et al. (2016). It consists of six stages:

1. **Exploratory stage:** Gather bibliography related to the main research topic.
2. **Descriptive stage:** Highlight the most important characteristics of the information previously collected.
3. **Correlational stage:** Identify the characteristics that usability heuristics must-have for the particular system.
4. **Explanatory stage:** Formally specify the proposed set of heuristics.
5. **Validation stage:** Contrast the new heuristics against the traditional heuristics through experiments.
6. **Refinement stage:** Based on the feedback obtained in the previous stage.

This methodology is also iterative and sets a precedent for the creation and validation of usability heuristics, as this is often an informal process (Quiñones et al., 2016).

5

TASK ORIENTED

This Chapter contains our proposal that contrasts between AUX and EUX, in here we follow the DSRM as a research methodology (see Section 4.1). Section 5.1 is about our problem at hand. In 5.2, we present the hypothesis that serves as the objective of our study. Section 5.3 explains our evaluation method, while Section 5.4 explains how to apply it. Finally, Section 5.5 shows the evaluation of our method.

5.1 IDENTIFY PROBLEM

Social networks have problems in the two areas that comprise them: technological, i.e., the platform that supports them. And social, i.e., misinformation problems, lack of motive, and guidance (Koh et al., 2007). User-tools can help to solve the problems in these both areas (see Figure 5.1), which are vital in a successful social network (Apostolou et al., 2017; Hummel & Lechner, 2002; Iriberri & Leroy, 2009; Preece, 2001).

User-tools are groups of widgets that make up the GUI of a social network, in order to allow users to perform tasks and communicate with each other, e.g., friend lists, newsfeeds, chats, and publishing menus. The granularity of user-tools is dictated by activities, i.e., a specific set of widgets, that allows solving a specific activity, conforms a user-tool.

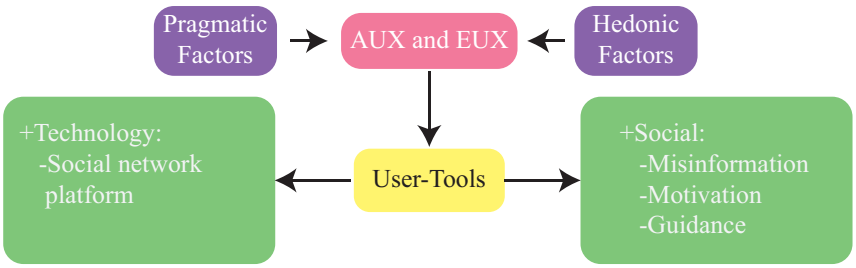


Figure 5.1: Design elements and influence of user-tools.

5.2 DEFINE OBJECTIVES OF A SOLUTION

As we have been mentioning, user-tools are the elements that allow interaction among users on a social network, so its design should be a primary issue. For this, our task focuses on contrasting AUX and EUX, since with this we hope to identify which dimensions of UX have the most significant weight in each period.

Identifying these contrast elements could help improve the design of the user-tools. Thus, we propose the following hypothesis for our study:

> *There is no significant difference in perceived UX between the prototypes imagined by the participants and the actual social networks performing the same tasks.*

5.3 DESIGN & DEVELOPMENT

Here, we describe each step of our assessment method (see Figure 5.2). To demonstrate how our proposal works, we take the basic case when one person uses a chat to make contact with another person:

- **Set Goals:** This step is about the objectives that developers need to achieve, e.g., a chat must allow users to communicate effectively with each other.
- **Identify Tasks:** It refers to the stages that the user has to follow with the aim of attaining the aforementioned objectives, e.g., a user has to recognize the receiver of the message, display the direct message option or window, compose the message and finally send it.
- **Identify User-tools:** This step involves determining which user-tools are available to accomplish the previously identified tasks, e.g., avatars, user profiles, lists, buttons, commands, and text boxes.
- **Assess AUX:** It concerns an AUX evaluation over the prototyped artifact. This stage can be done with various tools, e.g., low-fidelity prototypes (Virzi, 1989; Walker et al., 2002), or techniques such as The Wizard of Oz (Davis et al., 2007; Maulsby

et al., 1993). Nevertheless, the important thing is to stimulate the creativity of participants, so that we can obtain their idealizations and expectations. To know what aspects should be taken into account at this stage, we rely on the bases proposed by Yogasara et al. (2011):

  - **Intended Use:** It is about the practical connotation of each user-tool, e.g., the functioning of a chat from the user's point of view.
  - **Positive Anticipated Emotion:** It refers to agreeable feelings that the user expects undergoing as a result of the interaction with a user-tool, e.g., satisfaction after sending a message, happiness when the answer comes, generally pleased for not receiving errors or any other type of alert.
  - **Desired Product Characteristics:** As for this aspect, we accommodated the principles suggested by Morville (2005) to our case of study. These principles specify that a user-tool must be worthy, functional, helpful, attractive, attainable, honest, and discoverable.
  - **User Characteristics:** It concerns the mental and physical faculties of users, e.g., developing a generic chat does not imply the same endeavour that developing one intended for children or for seniors, since each group has specific needs.
  - **Experiential Knowledge:** We need to know the background of users, because they rely on their experience to gather information, then compare and contrast, e.g., a user might ask whether the new chat is more suitable than the one provided by Facebook.
  - **Favourable Existing Characteristics:** This aspect is about the properties that users have identified in the past as assertive in comparable tools, e.g., a user could think that they enjoy the chat from another platform thanks to the response time, availability, and ease of use.

- **Assess EUX:** This step involves conducting an EUX assessment over the developed artifact. For this step, we need at least a mid-fidelity prototype (Coyette et al., 2007; D-LABS, 2019), i.e., something so that participants can already experience

the tool on a PC or a mobile device. However, to make the comparison of results achievable, it is vital to evaluate all the aspects taken into account for the AUX assessment, e.g., if NASA TLX questionnaire (Hart & Staveland, 1988) was used in the AUX assessment, it is necessary to reapply it, this time for EUX, being careful to measure similar parts or functionalities between both stages.

- **Compare Results:** Once AUX and EUX assessments were carried out, the results have to be contrasted, so that developers can make resolutions on the design of user-tools, placing side by side the idealizations of users and reality, and examining whether their propositions were developed or not, e.g., compare the evaluations of the NASA TLX questionnaire of the prototype and developed chat.

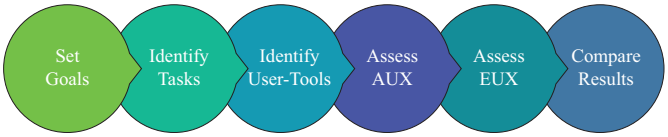


Figure 5.2: Steps of the AUX and EUX assessment method.

5.4 DEMONSTRATION

To carry out our tests, we use basic materials. For the development of prototypes, we have stationery such as sheets of paper, pens, pencils, and markers of various colours. Whereas for social media tests, we used a 15-inch laptop with internet access, and Firefox as a web browser. For each social network, we created a new user profile.

An essential factor that can compromise the validity and reliability of a study is improvisation. Choosing the wrong instrument invalidates the results, no matter how rigorous a study's proposed methodology was (Hernández-Sampieri & Torres, 2018; Kothari, 2004). That is why we weigh in on the various factors that could affect our tests. AttrakDiff, since its original proposal in 2003 (Hassenzahl et al., 2003), has been used in multiple tests to measure UX based on its pragmatic and hedonic factors (Hassenzahl, 2007). In each study,

experts have used this tool, and it has been tested for validity and reliability in different contexts (Braun, 2020; Díaz-Oreiro et al., 2019; Hassenzahl & Monk, 2010; Isleifsdottir & Larusdottir, 2008; Klaassen et al., 2013; Lallemand & Koenig, 2020; Ribeiro & Providência, 2021; Takahashi & Nebe, 2019), it has been translated into various languages (Lallemand et al., 2015), and it has been modified to suit the specific needs of particular experiments (Isomursu et al., 2020). Besides, it is simple to answer and does not represent a burden for participants (Walsh et al., 2014). All these results made us choose AttrakDiff as a valid tool to study UX.

The AttrakDiff full questionnaire is composed of 28 semantic pairs, i.e., pairs of words that make a strong contrast to each other (e.g., good-bad). Through these semantic pairs, the questionnaire measures the following aspects (Hu et al., 2013):

- **Pragmatic Quality:** It refers to the perceived quality of manipulation, i.e., effectiveness and efficiency of use.
- **Hedonic Quality - Identity:** It indicates the user's self-identification with the artifact.
- **Hedonic Quality - Stimulation:** It means the human need for individual development, i.e., improvement of knowledge and skills.
- **Attractiveness:** It reports the overall worth of an artifact based on perceived quality.

The hedonic and pragmatic dimensions are autonomous of each other and provide evenly to the UX evaluation (Hassenzahl, 2007). We use a printed version, in English, of the questionnaire available on the official website of the tool [1]. All participants had the same materials at their disposal.

Since we try to study the user-tools of social networks, and we have one independent variable with two factors, prototypes and social networks, our tests follow a basic design (Lazar et al., 2017a). Moreover, since we had only one group of participants who were exposed to both factors, our tests have a within-group design (Lazar et al., 2017a).

1 http://attrakdiff.de/index-en.html

Our only dependent variable is UX, of course, but since it is a latent variable and therefore cannot be measured directly (Sauro, 2016), we have AtrakDiff, which with its four dimensions helps us measure the UX perceived by our participants (see Section 5.4).

Finally, our control variables are the environment where we carried out the tests, since all the participants were exposed to the same conditions, e.g., materials, noise and light levels, desk, chair, and room. The characteristics of our participants were also controlled (see Section 5.5). Table 5.1 summarises the variables of our tests. The method for conducting our tests has been widely used by various authors in similar contexts (Aula et al., 2010; Bevan et al., 2016; Chin & Fu, 2010; Merz et al., 2016).

Table 5.1: Variables of our Study.

Independent Variable	Dependent Variable	Control Variables
User-tools (prototypes and social networks)	UX (Pragmatic Quality, Hedonic Quality - Identity, Hedonic Quality - Stimulation, Attractiveness)	Ambient, and Participants

5.5 EVALUATION

We used an opportunistic sample to recruit our participants, given that they are all members of our department. All participants gave their informed consent for inclusion before they participated in the study. Besides, the study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of our department.

Our testing group was composed of 20 participants (five of whom were females), whose average age was 28.15, and the maximum and minimum ones were 38 and 20, respectively. We made the decision of limiting their age to a range between 20 and 40 years, in order to prevent our results from being biased by participants with particular needs, e.g., oversimplify the language and instructions used or make the fonts of the GUIs larger. Although we know that it is a rather small sample, it is within the average for this kind of tests (Díaz-Oreiro et al., 2019).

Participants were selected because of their familiarity with social networks. We think that people unconnected to such platforms

constrain their potential to perform the assigned tasks, causing an invalidating impact on our study. Moreover, we believe that better results will be obtained if participants have experience with social networks.

We carried out the AUX and EUX assessment of user-tools in a peaceful ambiance to limit outer sources of noise in our study. Each volunteer individually participates in the testing sessions, which were conducted by a moderator *in situ*.

As the first step of our tests, participants filled a questionnaire about their demographic information and former contact with social networks. Afterwards, participants performed the tasks and assessments.

Each session had a length of around 40 minutes. We run the tests in 20 days, i.e., one participant per day. All tests were done around 10 a.m.; we did this to try to have a similar state in each participant.

The results of the aforementioned questionnaire reported that YouTube is proven to be the most used platform by our participants with 100% of usage. As for Facebook, it got a moderated use with 47%, and Reddit was the least used with 2%. Therefore, we decided to use these three platforms to asses EUX.

First of all, we said that our **goal** is to improve the design of user-tools through the contrast of AUX and EUX. To achieve this, we devised the following three **tasks** that represent common activities within social networks. Participants would have to complete each one twice, one for AUX, and one for EUX, during the trial:

1. **Message:** Transmit a private message to another user.
2. **Publication:** Share multimedia.
3. **Search:** Look for somebody or for a certain theme.

To **identify** the required **user-tools** for accomplishing each task, we analysed different ways, e.g., giving them user-tools made up of paper cut-outs. Nevertheless, if we provided participants with a predetermined set of user-tools, they would have prejudices, i.e., we would obtain very similar results between each participant, including the possibility of identical prototypes, consequently limiting their feedback significantly. Thus, for the AUX assessment step, the best alternative was that each participant created their own user-tools.

The next two steps of our method are the AUX and EUX **assessments** of user-tools:

- **Prototypes construction:** First, we asked participants to imagine that they took the role of a Web designer with the aim of creating a novel GUI for a social network. Afterwards, relying on their experience, they had to create three paper prototypes, corresponding to the three tasks previously defined. Participants had to draw GUI elements to solve the tasks, just as if they were designing a website GUI. In our pilot tests, we obtained prototypes similar to the one depicted in Figure 5.3(a), so we resolved to design a canvas to make it easier for the participants to create their prototypes. Figures 5.3(b)-(d) show random samples of prototypes in our actual tests. When they concluded the construction and description of each prototype, participants had to assess them with the AttrakDiff questionnaire. Therefore, this stage allowed participants to explain their decisions about how they conceive the behaviour of the GUI, the rationale behind their designs, and the user-tools required to accomplish each task. In this manner, we assess the AUX of user-tools.

- **Tasks using online social networks:** Once the three prototypes and their assessments were concluded, we asked participants to carry out the same three tasks, but now using online social networks. Hence, on Reddit, participants transmitted a private message to another user; on Facebook, they shared multimedia, and on YouTube, they sought for a somebody or for a certain topic. Like in the previous stage, after finishing each task, they had to assess it through the AttrakDiff questionnaire. Just like that, we assess the EUX of user-tools.

In this way, and taking into account that each participant made six evaluations, we finished with 120 questionnaires: 60 corresponding to AUX and 60 to EUX.

This test has a hybrid nature between formative and summative (see Section 2.5.1). By prototyping, participants carry out a formative test, since they will help us assess how well the fundamental elements of a GUI work. The summative nature is in the comparison of social networks with prototypes (Joyce, 2019; Sauro, 2010a).

5.5.1 *Results*

Seven semantic pairs correspond to each dimension of AttrakDiff. The ratings go from one to seven, and the higher, the better. Table 5.2 contains the results from the 120 questionnaires, the means (μ) and the standard deviations (σ) of each dimension for the three tasks.

Table 5.2: AttrakDiff Dimensions Results.

			Pragmatic Quality	Identity	Stimulation	Attractiveness
Message	AUX	μ	5.65	4.75	3.42	5.17
		σ	0.29	0.95	0.53	0.29
	EUX	μ	3.22	3.50	3.52	3.30
		σ	0.49	0.44	0.51	0.31
Publication	AUX	μ	5.57	4.77	3.70	5.10
		σ	0.44	0.69	0.45	0.48
	EUX	μ	5.97	5.35	3.99	5.74
		σ	0.23	1.08	1.19	0.29
Search	AUX	μ	5.51	4.67	2.98	4.96
		σ	0.54	1.02	0.57	0.52
	EUX	μ	6.23	5.42	3.75	6.02
		σ	0.45	1.05	1.31	0.26

Figure 5.4 is the graphical representation of the results. For all plots, the X-axis contains the four dimensions of AttrakDiff, and the Y-axis measures their averages. As the legend indicates, clear bars are the measurements of AUX, while dark bars represent the results of EUX in our three tasks: Figure 5.4(a) contrasts the results for Messages, Figure 5.4(b) does the same for Publications, and Figure 5.4(c) for Searches.

AttrakDiff offers three graphs for each test. Figures 5.5, 5.7, and 5.9 are the portfolio graphs, in which the values of hedonic quality are represented on the vertical axis (bottom means a low value). The horizontal axis represents the value of the pragmatic quality (left means a low value). Depending on the values of the dimension, the tested product will lie in one or more "character-regions".

Figures 5.6, 5.8, and 5.10 feature diagrams of average values for the four dimensions. While Figures 5.11, 5.12, and 5.13 show the average values for each pair of words in the questionnaire. These two types of charts have a range that goes from -3 to 3, which represents the scale of seven steps that each semantic differential has (higher is better).

In assessing these results, we also look at the reliability scores for the different dimensions. Table 5.3 shows the Cronbach's alpha values for the AttrakDiff dimensions in each task ($\alpha level = 0.05$).

Table 5.3: AttrakDiff Dimensions Reliability Analysis (Cronbach's alpha values).

	Message		Publication		Search	
Dimension	AUX (0.82)	EUX (0.87)	AUX (0.83)	EUX (0.83)	AUX (0.78)	EUX (0.83)
Pragmatic Quality	0.79	0.87	0.80	0.62	0.83	0.70
Identity	0.56	0.65	0.53	0.67	0.62	0.57
Stimulation	0.92	0.83	0.94	0.76	0.86	0.77
Attractiveness	0.81	0.93	0.80	0.86	0.76	0.93

Table 5.4: p values for paired-samples *t*-tests (comparisons between AUX and EUX in each dimension).

Dimension	Message	Publication	Search
Pragmatic Quality	$2.79 \times 10^{-6*}$	0.18	0.01^{*}
Identity	$3.08 \times 10^{-5*}$	0.05^{*}	0.003^{*}
Stimulation	0.82	0.53	0.08
Attractiveness	$6.53 \times 10^{-6*}$	0.06	0.0005^{*}

$^{*}p \leqslant 0.05$ significant

To contrast the results of the tests, and given that the design we have is within-groups with an independent variable of two factors, the statistical analysis we performed was a paired-samples *t*-test (Lazar et al., 2017b). In this way, we determined whether there are significant differences between the means of each dimension of

AttrakDiff in the AUX and EUX tests for each task (see Table 5.4). To obtain all the statistical analyses, we use the R language.

5.5.2 *Discussion*

Table 5.2 clearly shows that the paper prototypes were better evaluated than their counterpart in Reddit. Moreover, Figure 5.4(a) reveals something similar. The prototypes for messages were the only ones where the assessment of AUX exceeded that of EUX. This is likely because, for most participants, this was their first time using Reddit. It can also be attributed to Reddit offering a negative UX, since it was not so easy for participants to use their previous experiences on a new platform.

Even though participants were free to design their user-tools at their convenience, based on their experience, real social networks gave them more satisfying experiences. Figure 5.4(b) and 5.4(c) show that, indeed, all the dimensions were superior in social networks, although it is interesting that there is a difference, but not that much. An intriguing observation is that the participants were quite incisive in criticising their prototypes, i.e., they complained that they did not do a good job, because they did not have the experience or knowledge necessary to design a GUI.

In general, we can say that the reliability of data is good, since most of the dimensions obtained good results (> 0.7) as can be seen in Table 5.3. The result that stands out the most is that of the *Hedonic Quality - Identity* dimension, as none of the tests was significant. This could come to mean that AttrakDiff has a weakness to measure the *Identity* dimension. Of course, we would need more evidence to verify or refute that assumption.

Table 5.4 suggests which results of the *t*-test with paired-samples allows us to reject our null hypothesis. The comparison between AUX and EUX of the Messages task were significant in the dimensions of *Pragmatic Quality*, *Identity*, and *Attractiveness*. For the Publications task, only *Identity* was significant, while for the Search task, *Pragmatic Quality*, *Identity*, and *Attractiveness* were significant. These significant dimensions indicate that, in these tests, we can refute the null hypothesis, because there is a significant difference between the UX perceived by the participants between the prototypes and the social networks. It is interesting to note that the only dimension that was consistently not significant in any task was *Stimulation*.

According to Aladwan et al. (2019), when users of fitness applications were physically stressed by exercise and tried to use said apps with no avail, their stress increased, as their expectations were not met. This makes sense with our findings, since it is likely that, in an altered state of mind, users will need to rely on pragmatic elements that are familiar to them. Something similar happens with the tests carried out by Kukka et al. (2017), Margetis et al. (2013), Wurhofer et al. (2015), and E. Zhang et al. (2018), as their participants focused on interactions that they considered safe, when they found themselves in an unfamiliar environment.

Magin et al. (2015) studied the possible sources of negative emotions in UX, e.g., anger, sadness, and confusion. They determined that a significant part came from instrumental elements, i.e., usability, which agrees with our findings, since users expect things like that a button is active under certain circumstances or that a selected item can be removed, i.e., practical tasks.

The work by Moser et al. (2014) is interesting, because the expectations they measured came from children. It seems that their imagination was more oriented to hedonic aspects, mainly self-identification, since they cared that the games reflected their personality and decisions. It is striking because it goes against our findings: perhaps the AUX perceived by children has more weight in the hedonic factors, which could indicate a future path of investigation.

An exciting result we obtained was that the *Stimulation* means were not significant, as it could indicate that participants thought about basic user-tools to make their prototypes and found similarly essential elements in social networks. Now, we know that if we want to draw more reliable conclusions from this, we will need to do more research. However, we could speculate that the experience and imagination of the participants are limited to the essential elements that are commonly found in all GUIs, i.e., they prefer to play it safe. Users are looking for security rather than looking for new experiences when testing new GUIs, so *Stimulation* could become a more decisive factor when they are already familiar with GUIs.

Such behaviour could also indicate that user expectations are more grounded in pragmatic aspects than in hedonic ones. This could have significant implications. For example, it would imply that, when creating new GUIs, designers have to pay more attention to including basic user-tools that allow users to efficiently complete tasks, since user expectations would be mainly focused on practical

aspects, e.g., that they imagine a button, its action, but not how it looks.

The results presented in this work could have been affected by the sampling of our participants. Given that each evaluation took around 40 minutes, having a random sample would represent a significant challenge. Our participants did not receive any kind of incentive.

Similarly, the limitations of the within-groups design make it difficult to control the effects of learning and fatigue. We try to alleviate this by offering a comfortable and relaxed environment for the participants and reiterate them that they were helping us to evaluate the systems, and that we were not evaluating them (Lazar et al., 2017a).

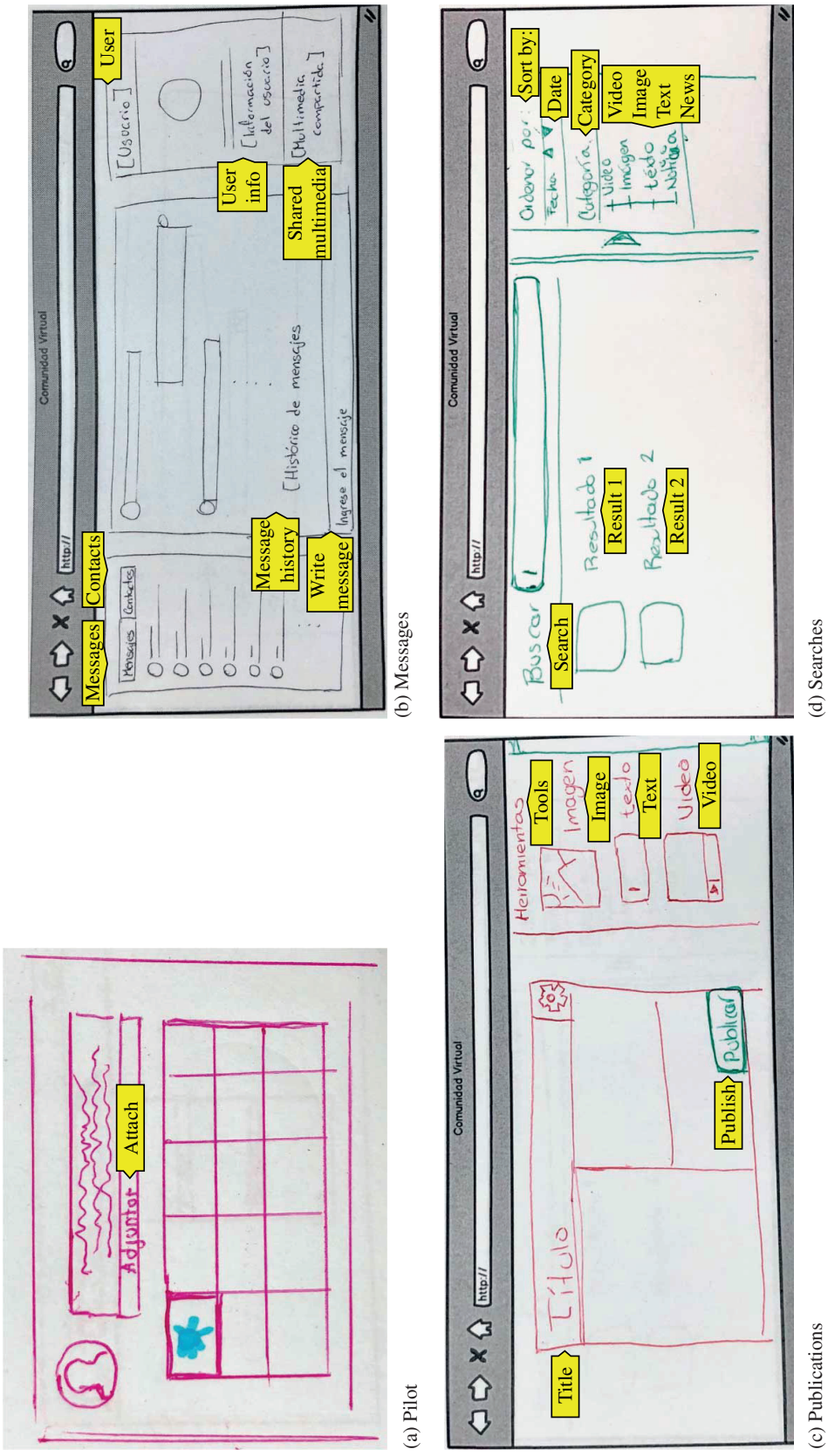


Figure 5.3: Samples of prototypes from the pilot tests (a) and from the actual tests (b)-(d).

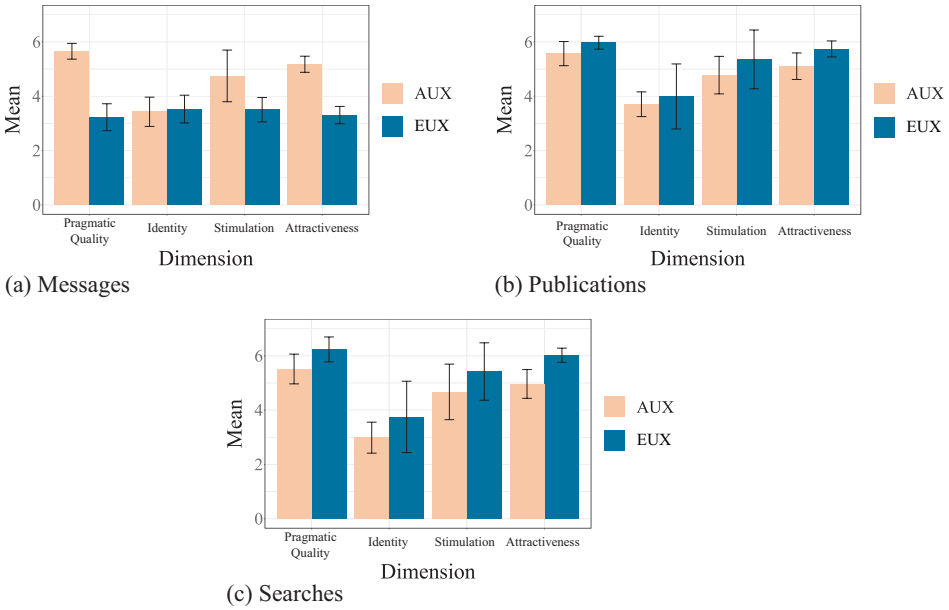


Figure 5.4: AttrakDiff results for Messages (a), Publications (b), Searches (c).

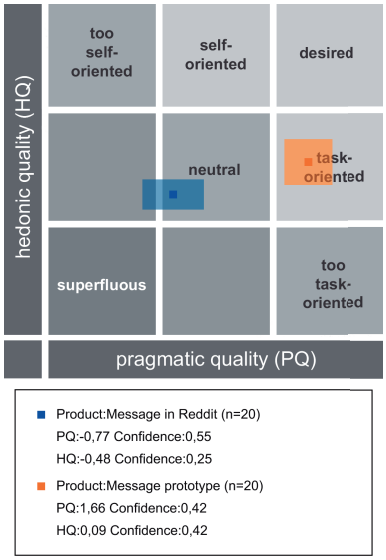


Figure 5.5: Reddit obtained rather low grades in both dimensions, while the prototypes are located in the region "task-oriented" meaning that there is room for improvement. Therefore, Reddit user tools did not precisely meet the expectations of participants.

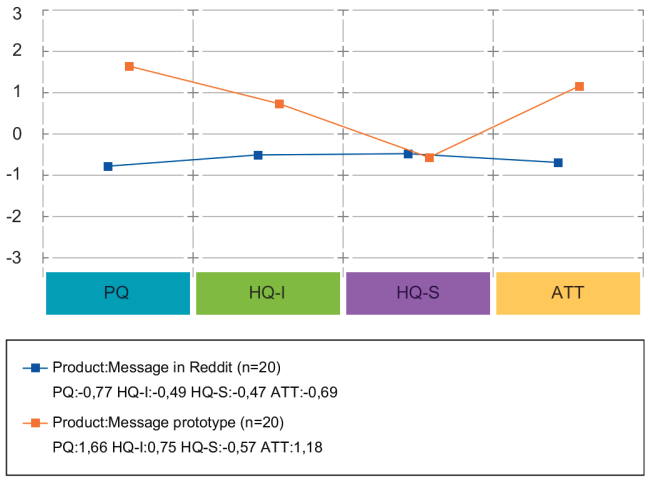


Figure 5.6: In each dimension, the participants evaluated their prototypes better than Reddit, except in HQ-S .

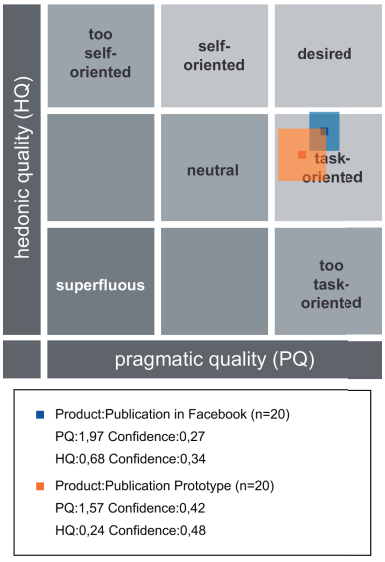


Figure 5.7: Although quite close, Facebook obtained better results than the prototypes. In both cases, changes would have to be made to arrive at the "desired" region.

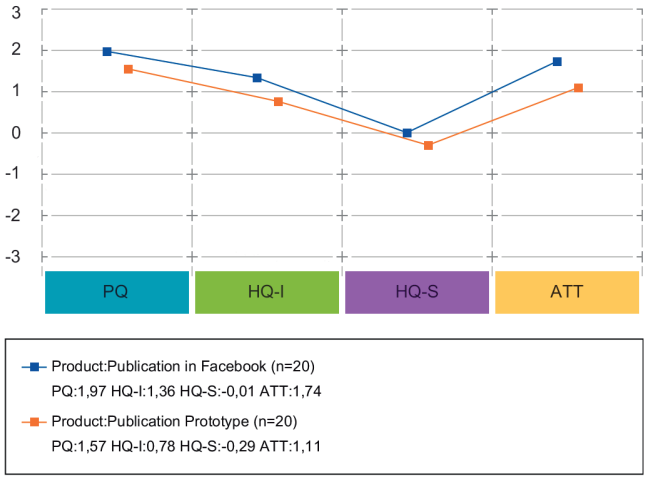


Figure 5.8: Facebook came out slightly better evaluated than the prototypes.

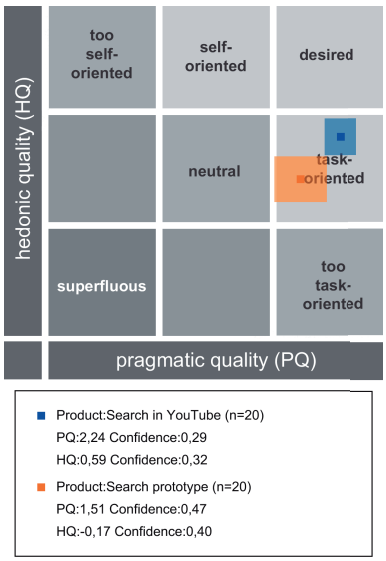


Figure 5.9: In the search task, both evaluations are in the "task-oriented" region. However, YouTube got slightly better results.

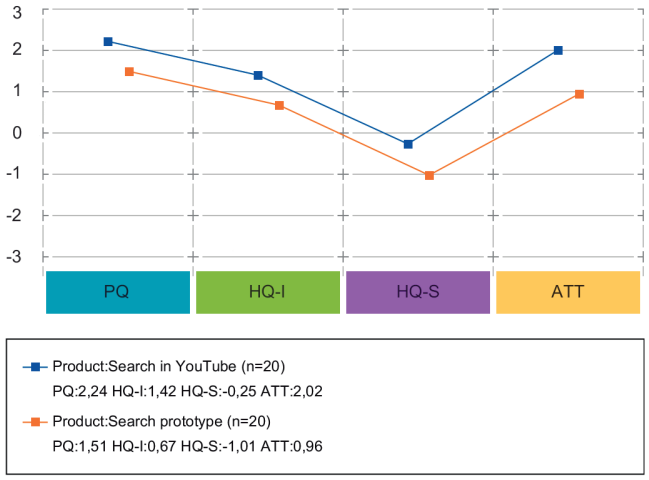


Figure 5.10: Search evaluations are quite similar; YouTube has a little advantage over the prototypes.

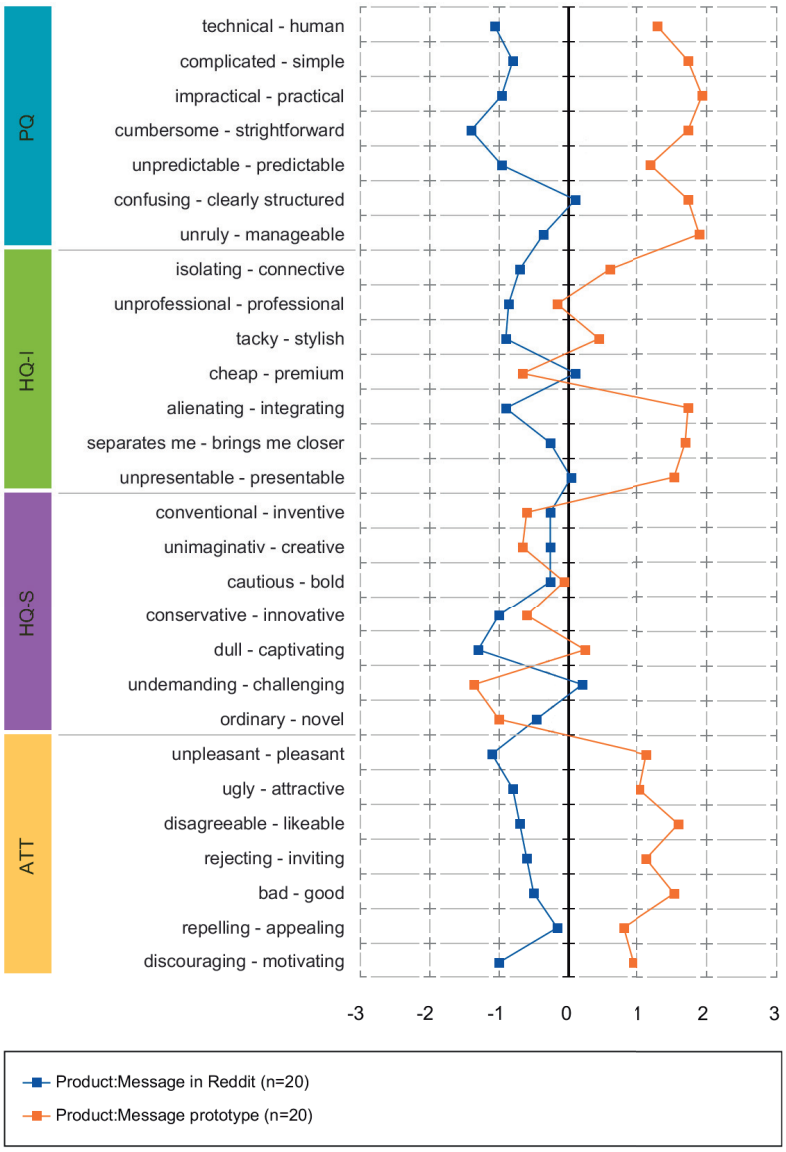


Figure 5.11: The most remarkable differences we can observe are that the participants rated their prototypes as *straightforward*, *integrating*, and *pleasant*.

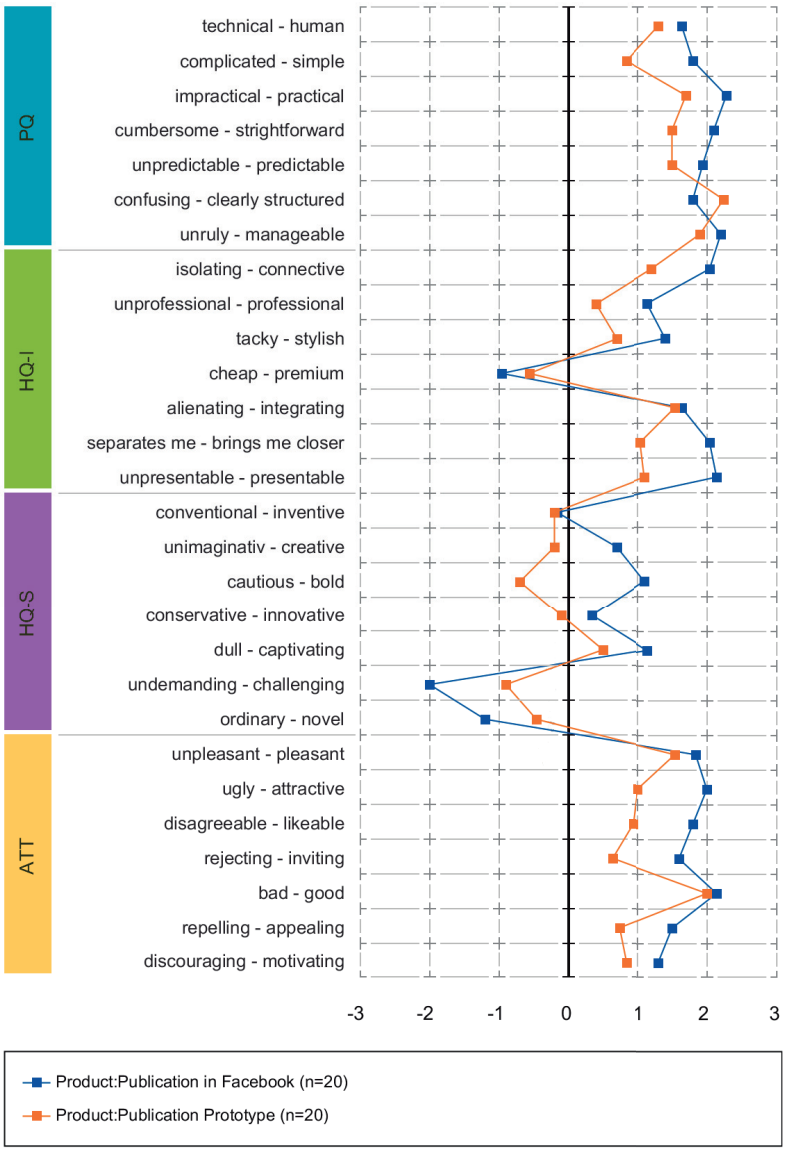


Figure 5.12: In each semantic differential, the participants evaluated similarly, but we can observe the differences in *brings me closer*, *presentable*, and *bold*.

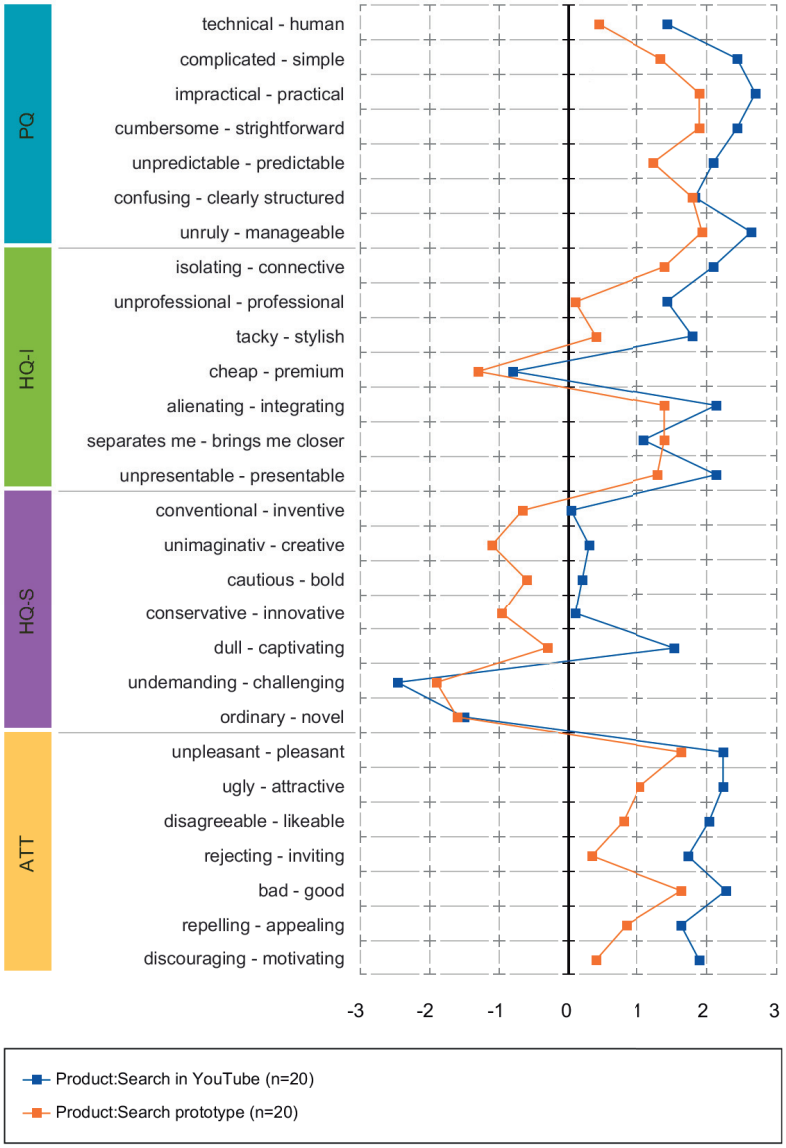


Figure 5.13: YouTube and the prototypes get very close ratings, even so, we can notice differences in *professional*, *stylish*, and *captivating*.

6

USER ORIENTED

This chapter shows the development of a chatbot prototype oriented to the learning/teaching process. We focus on gathering the needs and expectations of users. All the work described here is the first iteration of the GDS methodology (see Section 4.2). Section 6.1 talks about the problem that concerns this test. Section 6.2 exposes the various solution alternatives and the one that was finally decided to develop. This development is reflected in Section 6.3 and, finally, Section 6.4 is the evaluation of our prototype.

6.1 UNDERSTAND

The primary objective of this chatbot is to serve as an extra-school tool and, at the same time, as an intermediary between teachers and students: advising them, monitoring them, and facilitating communication between them. In this way, students could express themselves more freely, as the chatbot serves as a bridge with their teachers and other personnel involved in this process, e.g., social workers, psychologists, pedagogues, prefects, and administrative staff.

The proposed chatbot was designed and implemented as a Web application, using a text-based user interface. We define several profiles, e.g., teacher, student, and administrative staff, to interact with the chatbot, since each one has specific functions in the teaching/learning process. For instance, in the case of the student profile, the chatbot gives suggestions for their classes, as well as exam dates and project deadline reminders. As for the teacher profile, the chatbot allows the teachers to receive student questions, as well as to suggest exercises and complementary material for reinforcing some specific topics. In this way, students can have several sources and different ways to understand a topic and enrich their knowledge.

The collaboration between a middle school and our university took place in the context of a special project between both institutions. The objective of this project was to create a technological tool so that students could have additional support in their classes, a

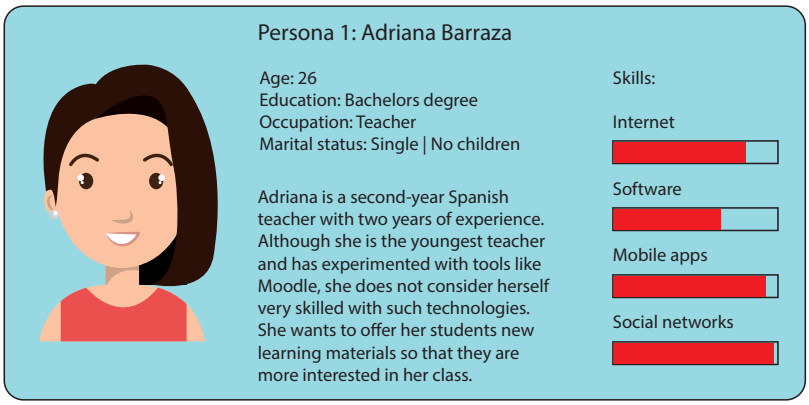
means of information on procedures, as well as a means of communication between them, teachers and administrative staff.

After spending some days familiarising ourselves with the most common processes that students have to perform, conducting interviews with students, teachers and administrative staff, we create three personas who represent these three critical profiles of end-users of the system (see Fig. 6.1).

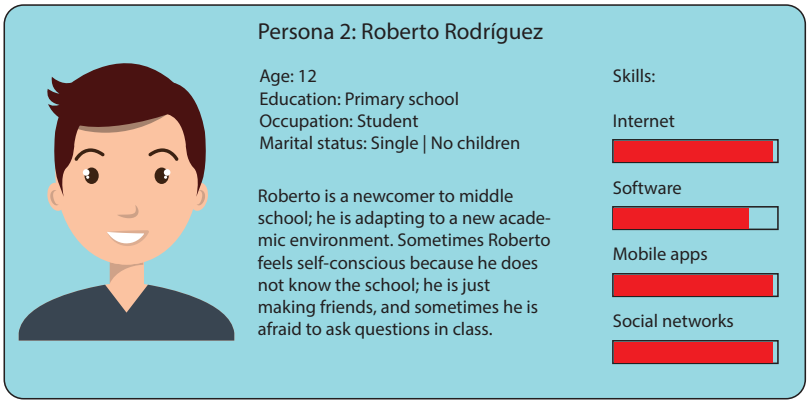
The second part of this stage was to refine our interviews to discover the requirements and characteristics of each user profile (Keijzer-Broers & de Reuver, 2016; Southall et al., 2019), and then apply this knowledge in user stories format (Cohn, 2004) to each persona (see Tables 6.1, 6.2, and 6.3).

Table 6.1: User stories for persona Adriana.

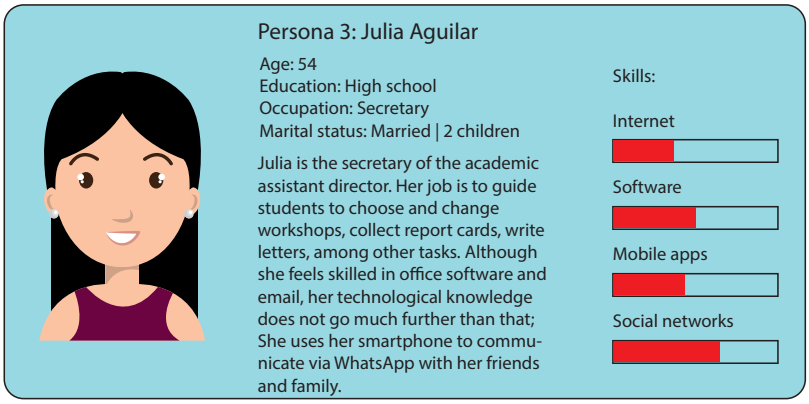
Requirements	Must-have	Nice to have
Functional	I want to be able to see all my information classified by class and grade. I want to be able to upload extra material for my classes. I want to be able to see the questions of my students.	I would like it to allow me to post information notices periodically for my students. I wish it to allow me to organise activities in and out of class.
User interaction	I want it to be simple to use. I want to be able to access the system from my home and smartphone.	I would like it to have a register of each student, in order to write down the strengths and weaknesses of each one, so I could help them better.
Social context	I want it to help me communicate better with my students. I want it to be a tool to improve my classes.	I would like it to improve the work between teachers and administrative staff.



(a) Adriana - Teacher



(b) Roberto - Student



(c) Julia - Secretary

Figure 6.1: The three personas we develop through interviews play the leading profiles of the system: teacher 6.1a, student 6.1b, and administrative staff 6.1c.

Table 6.2: User stories for persona Roberto.

Requirements	Must-have	Nice to have
Functional	I want it to know the schedule of my classes and exams and other important events. I want it to have additional material and explanations to the topics seen in class. I want to be able to send messages to my teachers so they can answer my questions.	I wish I could check my grades. I would like the system to explain to me complete topics of classes.
User interaction	I want it to be simple to use. I want to be able to access the system from my home and smartphone.	I wish I could send my homework through the system.
Social context	I want it to help me communicate better with my teachers. I want it to help me improve in class.	I wish I could share my problems and concerns.

Based on the personas and their user stories, we were able to discuss multiple scenarios. Fortunately, the needs of the end-users agreed and complemented each other, so we finally chose a scenario that would guide the design of the system: *The school needs a highly available web application that is easy to use for both experienced users and those struggling with technology. Response times need to be fast and, in some cases, immediate, so some processes need to be automated. The system must recognise the three profiles of the personas created*.

6.2 DIVERGE & DECIDE

Taking into account the requirements of the users and the guide scenario, and relying on various brainstorm techniques (e.g., mind maps and storyboards) each participant designed multiple solutions for the scenario in question. All proposals were presented without considering the possible limitations or criticisms and were widely

Table 6.3: User stories for persona Julia.

Requirements	Must-have	Nice to have
Functional	I want students to receive reminders about important dates, e.g., enrolment. I want them to download formats for the various procedures. I want the system to answer the FAQs.	I would like the system to guide the students in each step of their paperwork.
User interaction	I want it to be simple to use. I want to be able to access the system from my home and smartphone.	I would like the system to direct students with the corresponding person and office for each paperwork.
Social context	I want the system to help us get closer to the students.	I would like the system to help me to provide the necessary attention to those students who need it most.

discussed. Table 6.4 summarises some of the ideas we reached, their corresponding verdict and the reason for it.

To finally choose the solution proposal, each participant chose their favourite proposal and the reasons for their choice. After a debate between the alternatives, it was chosen by a joint agreement that a chatbot would be the most viable option.

6.3 PROTOTYPE

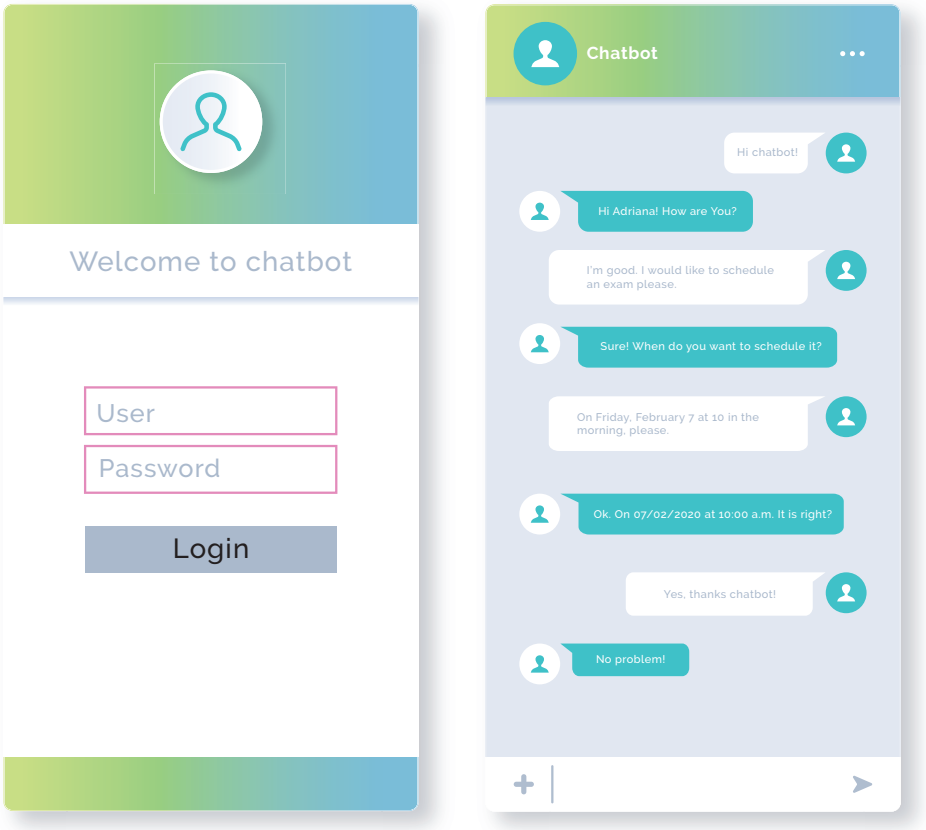
After choosing a chatbot as the best option, we decided to make a quick implementation, because we wanted to know to what extent our solution met the requirements of the end-users, the flexibility of the chosen technology, as well as giving us an outline of how it

Table 6.4: Objective solving proposals.

Proposal	Verdict	Reasons
Use Moodle	Rejected	Some teachers commented that they tried to implement it in their class but lacked the training and motivation to keep the platform up to date.
Use Gradelink (or similar)	Rejected	Although initially it can be expensive, a prefabricated and tested system usually gives good results. However, it would require training, interaction with students is usually limited, and as with Moodle, motivation is a significant obstacle.
Custom system (such as Gradelink or similar)	Rejected	Developing a school administration system is a huge task. While the school can save on licenses, development and maintenance can be very expensive. To make sure that development is on track, multiple tests are needed throughout the process, and this also increases the cost.
Use social networks	Rejected	Although social networks allow us to communicate very quickly, as that is one of their objectives, they would probably fall short of fulfilling the persona's requirements. Besides, there is an inherent issue of privacy and information security.
Chatbot	**Accepted**	The main challenge of developing a chatbot is its training, i.e., that the answers it provides are consistent with what was asked, in order to have a meaningful conversation. A chatbot is a good option since it will always be available, anyone can use it, many processes can be automated, and it allows all users to participate equally.

would be to develop the system altogether. Had we tried a functional or paper prototype, we would probably have ended up with an incomplete picture of whether a chatbot was the best solution or not.

In this way, using Angular, Node.js and Dialogflow, we develop an initial version of the chatbot with the functionalities that meet the main requirements of the end-users (see Fig. 6.2).



(a) Login (b) Scheduling an exam

Figure 6.2: The chatbot is a web application. We create a user-friendly interface for mobile devices.

Until now, the chatbot has the following features:

- There are three types of profiles: teacher, student, and administrative staff. At the moment, the registration and classification of all users are done manually by the chatbot administrator.

- The teacher profile contains their assigned groups and classes, as well as a list of students classified by those areas.
- The student profile contains the classes, teachers and assigned schedules.
- At the moment, the profile of the administrative staff is the same as that of the teacher.
- Teachers and administrative staff can schedule events with mandatory date and time.
- Students can ask for previously scheduled events.
- The chatbot can make automatic event reminders.
- Teachers can tell the chatbot to save links (e.g., websites, files, YouTube videos) and assign them to a subject so that students can then ask for them.
- If a student asked for material from a subject and the chatbot found nothing, it tells the teacher that the student asked for that material.
- The teacher can send a message to all students who belong to the same group.
- Teachers and students can send files to each other, for example, to send homework.
- The chatbot can respond to basic conversions on subjects and personal matters.
- The chatbot is trained to answer the FAQs of the essential administrative procedures.

6.4 VALIDATE

In this section, we describe the first set of tests conducted with end-users. We evaluated the workload perceived by the participants after performing some tasks. We gathered eight end-users: two teachers (one male-36, one female-28) and six students (three male, three female, between 14 and 15 years old) from middle school. It is essential to mention that all of them were using a chatbot for

the first time. To accomplish our tests, we first let them explore the user interface of the chatbot for a few minutes, in order to get acquaintance with the different kinds of widgets and their corresponding functionalities. Then, according to their profile, we asked them to perform the tasks listed below.
For both profiles:

- Asking a specific person for some material in the form of a file or of a simple answer to a question.
- Sending a file to a given person or group.
- Receiving a file from a specific person.

For students:

- Establishing a personal or academic conversation with the chatbot.
- Receiving some suggestion from the chatbot about a personal or academic question.
- Asking for information about a course, e.g., doubts or questions.

For teachers:

- Adding information about a course, in order to be shared with the concerned students, e.g., extra-instructional material, project schedules, and reminders of homework deadlines or quiz dates.
- Receiving some warning from the chatbot about the instructional materials that a given student is needing.

The nature of this evaluation is absolutely formative (see Section 2.5.1), not only because it occurs with a prototype at the beginning of a development but also involves the role of the users in the test.

6.4.1 *Results*

To measure the workload perceived by our eight users of the chatbot, we used the *NASA Task Load indeX* (NASA-TLX) (Hart & Staveland, 1988). This is a subjective assessment tool that uses six scales (also called dimensions) to measure the workload perceived by an individual or a group in the execution of some task. The test has two parts. In the former, the users evaluated a task in the following scales: *mental demand*, *physical demand*, *temporal demand*, *performance*, *frustration*, and *effort*.

For each scale, the user selected a value in the interval $[0, ..., 100] \subset \mathbb{N}$ with ticks of 5 units, giving 21 possible qualifications.

On the left of Figures 6.3 and 6.4, the given ratings for the six scales are shown. It is important to mention that we measured the whole workload for the set of tasks described above, so a big value means a lot of workload, which is considered as bad. On the other hand, a small value close to zero means a small workload, which is considered as good.

In the latter part, each user assigns a custom weight for the six scales, which is used to adjust the previous qualifications, as shown on the right of Figures 6.3 and 6.4.

6.4.2 *Discussion*

The average adjusted qualifications for the six scales are shown in Table 6.5. The scale with the smallest average rating was *Physical Demand*, followed by *Frustration*. It is interesting to observe that practically all users gave very little importance to the physical aspect, while interacting with the user interface of the chatbot. The ratings in this scale were small *per se*, but also the given weight was zero for almost all users. The scale with the biggest average rating was *Performance*.

Finally, we calculated the pondered rating from the adjusted ratings in the six scales. Figure 6.5 shows the pondered ratings given by each participant. The average of pondered ratings for the perceived workload given by the eight users is 23.67. The minimum pondered rating was 14 and the maximum was 36.33, in the range [0,100]. These qualifications are in the "good" region of the scale, i.e., around the first quarter, and give us an approximate idea about

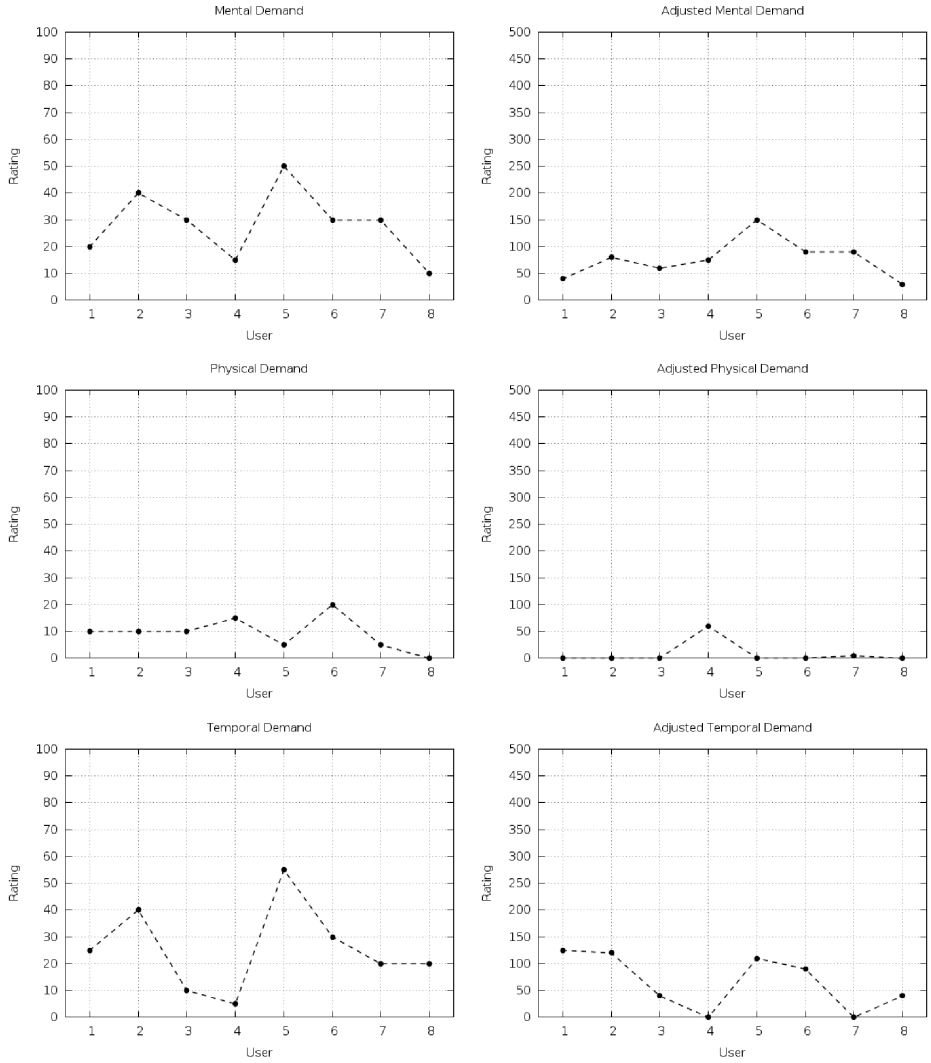


Figure 6.3: Mental, physical, and temporal demand perceived by each user (left) and adjusted by the assigned weight (right).

how good was the perception of the participants about the burden of performing tasks while using our chatbot.

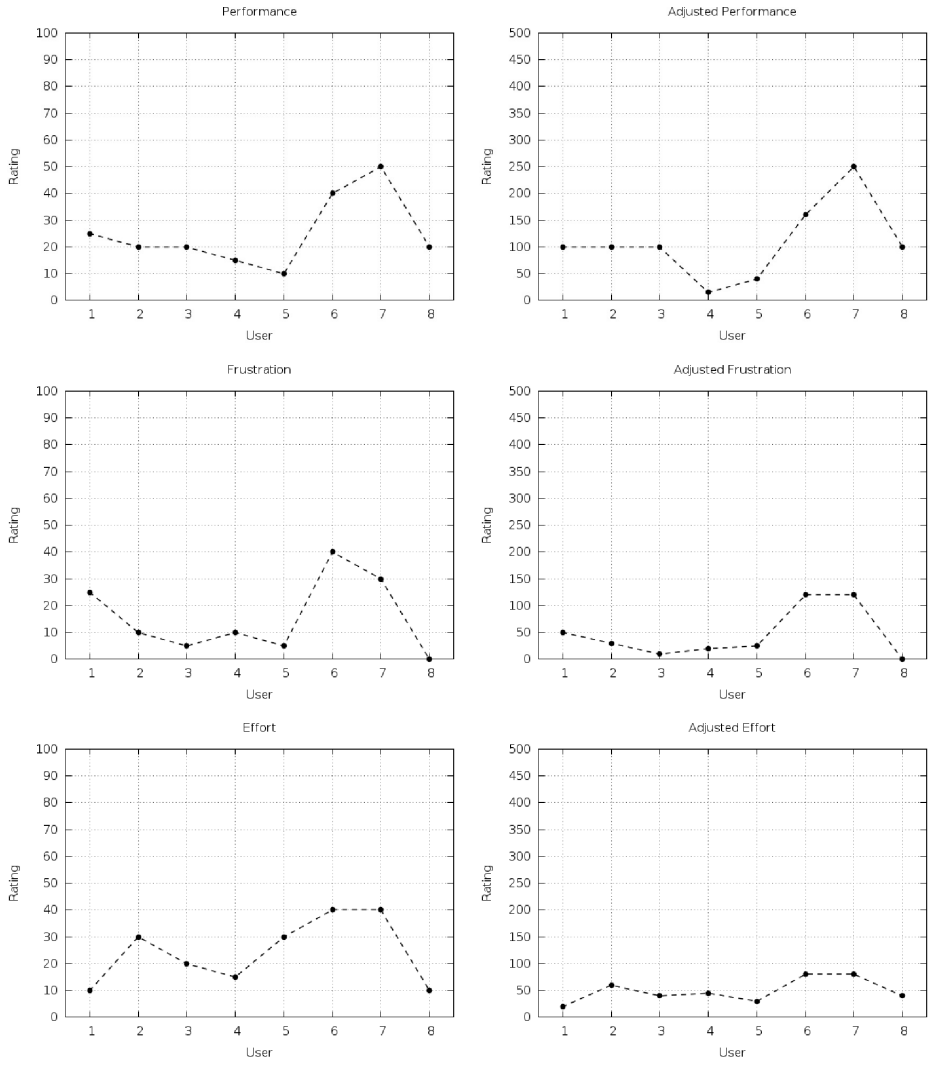


Figure 6.4: Performance, frustration, and effort per user (left) and adjusted by the assigned weight (right).

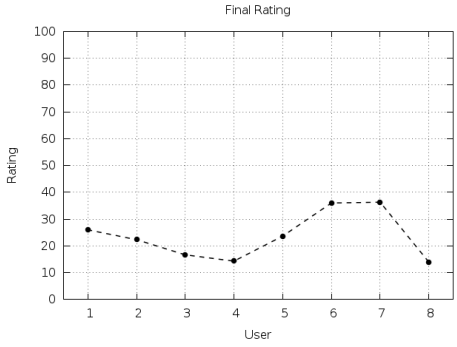


Figure 6.5: Weighted rating per user.

Table 6.5: Average adjusted ratings.

Mental Demand	76.875
Physical Demand	8.125
Temporal Demand	65.625
Performance	108.125
Frustration	46.875
Effort	49.375

7

CONVERSATIONAL HEURISTIC ORIENTED

This Chapter contains the development and validation of our usability heuristics to evaluate conversational systems so that the structure corresponds to the development of the DUH methodology (see Section 4.3): Section 7.1 deals with exploration in state of the art. Then, Sections 7.2 and 7.3 set out, respectively, the most relevant points of the exploration and how they relate to our proposal. Next, Section 7.4 contains our conversational heuristics, validated and refined in Sections 7.5 and 7.6, respectively. Finally, Section 7.7 represents the Case Study that helped us validate this set of heuristics.

The intrinsic nature of heuristic evaluations is formative (see Section 2.5.1), as it is evaluated in the early stages of development using well-known principles or empirically identified good practices (Joyce, 2019).

7.1 EXPLORATORY STAGE

An extensive search for papers was carried out in the most recognised digital scientific libraries, such as Scopus, Web of Science, Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect and Springer Link.

7.2 DESCRIPTIVE STAGE

Chatbots have been around for decades. The oldest is probably "ELIZA", dating from 1966 (Weizenbaum, 1966). However, the technological foundations of chatbots have evolved rapidly, so an evaluation standard has not yet been established. From the search carried out in the previous stage, three evaluation categories can be established for chatbots: 1) those works that focus on technical or theoretical aspects, i.e., that have to do with the implementation (Sedoc et al., 2019) or with NLP (Kuksenok & Martyniv, 2019); 2) evaluations involving user satisfaction (Segura et al., 2019); and 3) some hybrid proposals that take elements from the first two classifications (Qiu

et al., 2017). Given the nature of this proposal, only papers from the second category were selected.

It should be noted that no usability heuristics were found for chatbots, neither other evaluation tool or method designed specifically for this class of systems. However, some relevant characteristics that were found are:

- The chatbot's responses are classified as correct or incorrect, according to the user's criteria or the evaluator (Segura et al., 2019).
- Many studies present results from questionnaires *ad hoc*, which only allow for success or failure ratings (Kazi et al., 2012).
- The usability characteristics that are most measured are efficiency, satisfaction, effectiveness, ease of use, performance, frustration, difficulty and mental effort (Guerino & Valentim, 2020).

7.3 CORRELATIONAL STAGE

As already mentioned, usability is classically founded on the axes of effectiveness, efficiency and satisfaction (Bevan et al., 2016). According to the exploration carried out in the first stage and the conclusions of the descriptive stage, it is estimated that the criteria summarised by Ren et al. (2019), in each usability axis, are those necessary to evaluate chatbots:

- **Efficiency:** task completion, accuracy and recovery.
- **Efficiency:** time to finish tasks, mental effort and communication effort.
- **Satisfaction:** ease of use, context-dependent questions, complexity control, physical discomfort, pleasure, desire to use it again, and learning ability.

This work was chosen because: 1) it compiles the essential characteristics to be evaluated in each usability axis, and 2) it represents the vision of multiple investigations, given that it is the result of a survey.

7.4 EXPLANATORY STAGE

According to the criteria mentioned in the previous stage, the following five heuristics are proposed. For each one we present, definition, examples, benefits (what is expected when the heuristic is fulfilled), as well as its relationship with usability axes (effectiveness, efficiency and satisfaction) and the rationale of each one, i.e., the fundamental reason of its existence based on the characteristics found by each usability axis:

H1 **Completeness:** Refers to the ability and flexibility of the chatbot to understand user input and help them solve their problems.

Example: If a user wants to create a reminder, the chatbot requests all relevant data (e.g., date, time and subject).

Benefit: Create interactions that help users to perform their tasks.

Usability axis: Efficiency.

Rationale: In order to finish user tasks accurately, the chatbot needs to identify all context-dependent data.

H2 **Context:** Indicates the ability of the chatbot to switch context, i.e., when conversing with a user, how easy is it to keep the conversation going? How quickly it adapts to user changes?

Example: When in a conversation the user used the word "tasks" and then used the word "homework" under the same context, the chatbot adapts.

Benefit: Offer the user a flexible interaction with the chatbot.

Usability axis: Efficiency.

Rationale: Reducing the user's mental and conversational effort is essential since good communication allows completing tasks in less time.

H3 **Naturalness:** Refers to whether the user can perceive that they are conversing with a computer or with a human.

Example: If the chatbot cannot understand the user, it offers to direct them with a person.

Benefit: Never confuse the user, as adverse effects can be caused, such as the uncanny valley Ciechanowski et al., 2019.

Usability axis: Satisfaction.

Rationale: By reducing the complexity of the conversations, the pleasure of using the chatbot increases.

H4 **Learning:** Denotes the ability of the chatbot to learn new inputs and interactions, as well as to offer alternatives when it did not understand the user.

Example: If the user requested a registration form, but the chatbot did not understand, it offers to contact the person in charge of school services.

Benefit: Whether by automatic or manual training, learning will be essential for the chatbot to remain useful.

Usability axis: Satisfaction.

Rationale: It is essential to constantly increase the learning capacity of the chatbot to make it a dynamic and handy tool.

H5 **Functionality:** Expresses the ease of use of each of the functions that make up the chatbot.

Example: The chatbot offers a calendar widget when the user enters a date.

Benefit: Avoid over or underdevelopments in widgets or functions that have already been studied before.

Usability axis: Efficiency.

Rationale: The reduction of mental effort and possible input errors are necessary factors to achieve a natural workflow.

7.5 VALIDATION STAGE

The validation was carried out with the help of ten usability experts (three female and seven male, between 30 and 50 years old). The experts were chosen for their experience conducting usability and UX evaluations. All the experts are university professors with postgraduate studies in some branch of Computer Science; two of them belong to CINVESTAV-IPN. Their experience comes from both industry and academia (between 5 and 25 years). It should be noted that none of them is directly related to this research, with the obvious exception that they were volunteers to participate in this study.

For this stage, the proposal was validated by contrasting it with the most popular set of heuristics, i.e., the one proposed by Nielsen and Molich (1990) which, although it was designed for software inspection, has found a place in evaluations of various types of products and services (Barnum, 2011b). The original set consisted of nine heuristics. However, for this evaluation, the new set consisting of 10 heuristics was used (Nielsen, 2020). We chose this set of heuristics because they are the best known and most used in studies of this kind (Mathis et al., 2021; Momenipour et al., 2021; Oulasvirta & Hornbæk, 2021; Quiñones & Rusu, 2017). In this way, we wanted our heuristics to be compared against solid, well-ingrained knowledge.

The experts were divided into two groups. One group used our heuristics, while the other used those of Nielsen and Molich. Both groups evaluated a chatbot for education called *Ask Frank*. This chatbot is implemented within the Facebook Messenger platform and answers simple questions about Mathematics, Science and History (Smutny & Schreiberova, 2020).

The organisation of five experts per group is because Nielsen recommends integrating three to five experts in an evaluation to obtain the maximum cost-benefit (Nielsen, 1992).

Figure 7.1 shows the expert associations of the problems with the heuristics. As can be seen, it is considered that the group that worked with the Nielsen and Molich set was 100% accurate. This is due both to the heuristics' clarity and the familiarity that the experts have with them. The group that evaluated using the heuristics for chatbots obtained similarly promising results since its problem-heuristic associations were judged to be accurate in 84.22% of the

cases. Seen another way, it is weighed that they were wrong by 15.78%, which indicates that the heuristics are understandable.

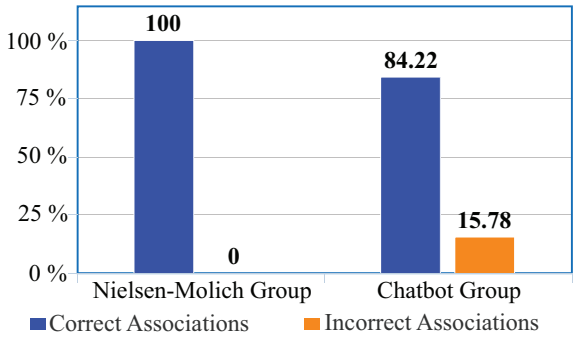


Figure 7.1: Percentages of correct and incorrect associations according to each group of heuristics.

Another indication that our heuristics are good to find problems is the numbers that resulted from the group that worked with our proposal. As shown in Figure 7.2, the experts who evaluated with these heuristics were able to find 19 problems, while the group that evaluated with the Nielsen and Molich set found 8. There were only four problems in the intersection.

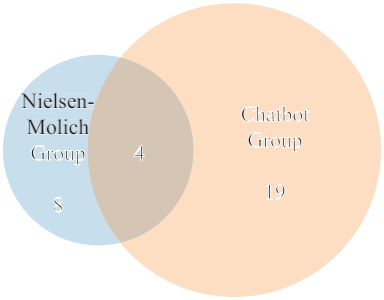


Figure 7.2: Comparison of problems identified by each group of heuristics.

7.6 REFINING STAGE

Thanks to the feedback obtained from the experts, it was decided to refine the heuristic "H1 Completeness" since it could be confused with "H2 Context". In this way, the heuristic was modified as shown below:

H1 **Completeness:** Refers to the ability of the chatbot to obtain, invariably, all the data necessary to solve a task.

7.7 CASE STUDY: EDUCATIONAL CHATBOT

Once the proposed heuristics have been explained as well as their origins and validation, the experts evaluated a product in this Section. The methodology followed for this evaluation was set out by Barnum (2011b) and successfully replicated in various studies (Chuan et al., 2015; Kumar et al., 2020; J. Zhang et al., 2003).

As the object of evaluation, previous work was chosen, a chatbot that serves as an extracurricular tool for middle school students (Mendoza et al., 2020). This chatbot is a tool that has various functionalities for teachers, students and administrative staff; users interact with it through text, in Mexican Spanish. The chatbot can store and make reminders of important dates (e.g., deadlines for submitting assignments, exams and registrations), offering material to students (that teachers had previously saved), allowing file sharing, answering frequently asked questions about administrative processes, among other functions. The chatbot is a web application developed with *AngularJS*, *Firebase*, *Dialogflow*, and *NodeJS*.

For this evaluation, we had the help of the same five experts who worked with our heuristics in the Validation Stage (see Section 7.5) since we wanted to take advantage of the experience they acquired in that stage. First, they were allowed to get a little familiar with the chatbot, operation, and user interface. Afterwards, they were asked to perform some simple tasks (e.g., ask the chatbot a question, schedule an exam and ask for a grade) and then do a quick evaluation, i.e., to say the possible errors they had found using the proposed heuristics. The latter served to strengthen the meaning of the heuristics and, in this way, obtain better results. The experts were instructed to evaluate the chatbot for two days, detect problems, and make a list of violated heuristics.

Once all the experts had evaluated the chatbot and prepared their lists, these were consolidated into a master list of problems. This master list was given to each expert so that, individually, they could assign a rating according to the severity of the problem and note which heuristics were violated in each case. Thus, the scores were averaged, and the results are presented in Table 7.1. Ratings were

assigned the same way as in the case of DistroPaint, Spotify and Home Security Systems (see Section 8.2.2).

7.7.1 *Results*

With the help of our heuristics, the experts found a total of 16 problems (an average of 3.2 problems per expert). The severity of the problems had a mean of 2.54. The master list consists of 10 unique problems, with a total of 20 heuristic violations. With seven problems found, **H4 - Learning** was the heuristic with the highest number of violations, followed by **H3 - Naturalness** with five violations. In contrast, with only two violations, **H1 - Completeness** was the heuristic with the fewest problems detected (see Figure 7.3).

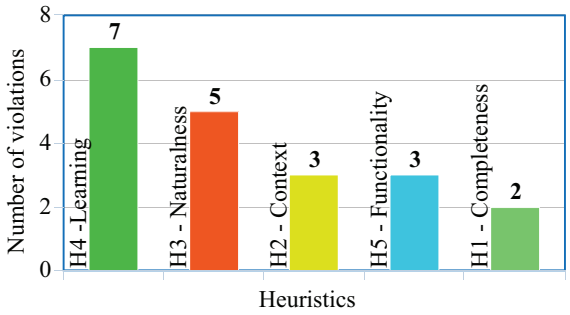


Figure 7.3: Heuristic violations for chatbots.

Regarding the severity of the problems detected, it can be seen in Figure 7.4 that severity level 3 - "Major problem" was the most frequent with 42%. On the contrary, it is noted that 0 - "Not a problem" scored 0%.

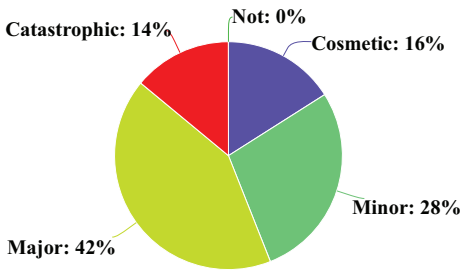


Figure 7.4: Severity of violations.

Table 7.1: Usability Problems and their Severity in an Educational Chatbot.

† Completeness (H1), Context (H2), Naturalness (H3), Learning (H4), Functionality (H5)

Problem	Heuristics†	Severity
I have to put the full date in a default format (DD/MM/YY) to create a reminder; there is no flexibility.	H1, H3, H4	2.4
It needs some widgets (e.g., file previews) as interacting with only text gets a bit monotonous.	H4	1.2
Some tasks require too many attempts for the chatbot to understand what the user meant.	H2, H3, H4	4.0
I do not have the confidence to provide my personal information.	H3, H4	2.8
Assigning grades, student by student, is a cumbersome task.	H5	3.2
The chatbot lacks "personality", i.e., it is an answering machine, and its responses are very formal.	H2, H3	1.6
More initiatives should be offered, e.g., if I ask for a grade, it should offer to see the rest of my grades.	H4, H5	2.6
It should be able to integrate with other applications (e.g., *iCal*).	H5	1.6
Sometimes it is difficult to change the context; when a task is being carried out, the chatbot loses the context of the previous task.	H1, H4	3.2
There is no persistence, i.e., if I want to get to a certain point in a task, there are no shortcuts; it is like always having the whole same conversation over and over again.	H2, H3, H4	2.8

7.7.2 *Discussion*

It is not a coincidence that the heuristic with the highest number of violations was **H4 - Learning**, since it is directly related to the possibility of completing a task; If the chatbot took too many attempts to understand the user input, this protocol becomes exhausting, as the process of starting a task becomes more expensive in every way than the task itself. Possibly, one way to mitigate this, in addition to improving chatbot training, is to offer links that lead to starting events, i.e., that the chatbot has the initiative to start tasks that are related to user input.

It is interesting to note that **H2 - Context** and **H5 - Functionality** got the same number of violations. From the evaluators' comments, it can be deduced that this tie of qualifications may have a certain correlation; When users do not know if the chatbot is still in the same context as them or that the answers provided do not provide adequate feedback, it creates a distrust of continuing with the task, especially if it deals with private information. Finally, that **H1 - Completeness** was the least violated indicates that when the chatbot successfully helped users with their tasks, it did so because it knew what data it needed to do so. The chatbot seems to be a suitable tool for the school aid scenario.

Regarding the severity of the violations, it can be argued that the evaluators found problems that should be corrected, since most of the ratings, i.e., 42%, fell within the classification of "Major problem", so it is essential to fix them to improve the usability of the system. From what can be seen, a large part of these problems is related to communication flows, i.e., since the text is the only form of interaction offered by the chatbot, it limits many tasks that have to be done constantly (e.g., assign grades to students). As some experts have already pointed out, this could be solved with the implementation of widgets that offer the user a shortcut to the desired task without the need to go through the same conversation over and over again.

Although the "Catastrophic Problem" rating obtained 14% of the total ratings, it can be considered a relatively high percentage. The problems detected by the experts reveal that those that fall into this level of severity have to do with the training and attempts of the chatbot, i.e., with aspects that concern the processing and

understanding of natural language. The only foreseeable solution is to improve chatbot training.

The rest of the problems obtained 44%, i.e., the combination of "minor" and "cosmetic" problems. This may result from the more significant problems because if the interaction flows fail at a certain point, this will most likely cause more failures. While it is true that some would be solved with the implementation of certain features that the experts missed, it is suspected that most of these problems will be solved once the larger ones are dealt with.

In general, it can be said that most of the chatbot's functionalities need considerable improvement.

An inherent disadvantage of heuristic evaluations is cognitive bias (Administration, 2013). This means that the quality of the evaluations is subject to the experience of the experts; the problems they encounter are conditional on what they already have prejudged as "good" usability. However, the following steps were taken to try to mitigate the biases in the tests:

- A standard evaluation methodology was adopted.
- We had the help of five experts, all with extensive experience in the field of usability and UX.
- All experts had the same information as a starting point.

Some qualitative indicators that the results are valid are the quantity and characteristics of the problems encountered. On the one hand, there were repeated problems among experts, that is why, out of 16 problems in total, we ended up with a master list of only 10. Which indicates that they understood the heuristics and similarly applied them. On the other hand, the problems are exact and well-identified; they concur with the purpose of the proposed heuristics, i.e., they are not general or vague reasoning that could coincide coincidentally.

8

CONSISTENCY HEURISTIC ORIENTED

This Chapter contains the case studies that allowed us to validate a set of consistency heuristics that we had previously developed. Section 8.1 presents our set of heuristics. Next, Section 8.2 shows DistroPaint, a prototype that we developed following our heuristics. Section 8.3 contains the evaluation of Spotify. Finally, Section 8.4 deals with the evaluation of Home Security Systems.

As we stated in Chapter 7, the heuristic evaluations are formative.

8.1 HEURISTICS

With the review of various works in the state of the art, and taking into account the challenges discovered and common characteristics of each one, we present our five heuristics to maintain consistency in multi-device systems (Sánchez-Adame, 2016):

- **Honesty:** Interaction widgets have to do what they say and behave expectedly. An honest GUI has the purpose of reinforcing the user's decision to use the system. When the widgets are confusing, misleading, or even suspicious, users' confidence will begin to wane.
- **Functional Cores:** These are indivisible sets of widgets. The elements that constitute a Functional Core form a semantic field, out of their field they lose meaning. The granularity level of interaction for a Functional Core depends on the utility of a particular set of widgets.
- **Multimodality:** Capability of multi-device systems to use different means of interaction whenever the execution context changes. In general, it is desirable that regardless of the input and output modalities, the user can achieve the same result.
- **Usability Limitations:** When multimodality scenarios exist, it is possible that situations of limited usability could be reached. When the interaction environment changes and its context is

transformed, the environment can restrict the user's interaction with the system.

- **Traceability:** Denotes the situation in which users can observe and, in some cases, modify the evolution of the GUI over time.

8.2 CASE STUDY: DISTROPAINT

In order to demonstrate the proposed consistency heuristics, we developed DistroPaint, a prototype application that integrates them. We decide to create a basic graphics editor, which provides several tools that can be distributed on several devices (PC, phone, and tablet). This section describes our proof of concept (see Section 8.2.1) and the expert analysis carried out (see Section 8.2.2) based on the works by Andrade et al. (2015), Grice et al. (2013), and Schmettow et al. (2017).

8.2.1 *DistroPaint*

DistroPaint is a Web application for basic graphic design. The user can access the application from a PC, a phone, and a tablet. They can distribute the GUI from the PC to the mobile devices, e.g., the colour pallet can be displayed on the phone, while the drawing tools are being shown on the tablet (see Figure 8.1). The user can configure the GUI at any moment. Below we list how our heuristics are reflected in the implementation of DistroPaint:

- **Honesty:** The part where DistroPaint's honesty stands out most is its presence system (see Figure 8.2), since it informs the user about the availability of their devices. The Honesty at this point is critical, because it allows the user to make decisions (to distribute, or not) according to the state of their interactive environment.

- **Functional Cores:** The main way of interaction in our application is the toolbox (see Figure 8.3), so we choose it as the main element for the Distributed User Interface (DUI). The decision of how to divide the elements could seem trivial, e.g., each tool (brush, eraser and line) could be distributed individually among several devices, however, this could be a risky option,

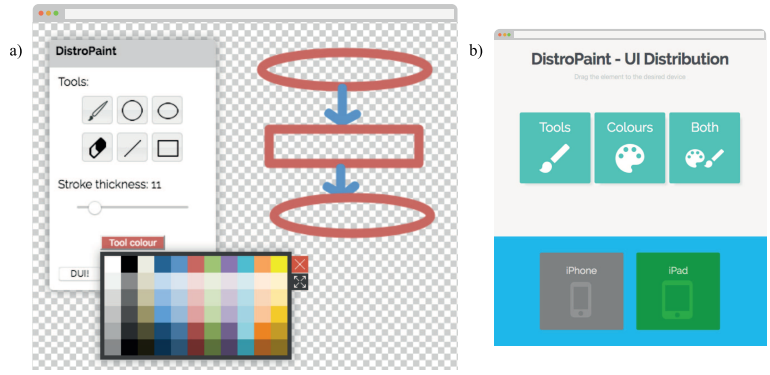


Figure 8.1: Predominant GUIs of DistroPaint on a PC web browser: (a) GUI of the graphical editor, and (b) the distribution menu for the widgets.

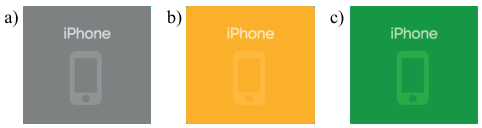


Figure 8.2: Presence system: (a) a grey box means that the device is unreachable; (b) an orange box indicates that the device is connected but it can not receive widgets; and (c) a green box expresses that the device is ready to receive widgets.

since it would bring very few benefits to the cost of generating confusion and increasing the system requirements.

So we decide that the tools and the slider for the stroke thickness should form a semantic field. In the same way, another field would be occupied by the colour palette, thus, we have two Functional Cores as result.

- **Multimodality:** The element for the change of context that has more repercussion in our application is the change of platform. No matter whether a user uses one element of the toolbox from the PC (by clicking with a mouse) or from a mobile device (by touching with a finger), DistroPaint has to respond seamlessly (see Figure 8.4).

- **Usability Limitations:** We create a synthetic limitation in our prototype (see Figure 8.5). We decide that both of our Functional Cores have to be available for both the phone and the

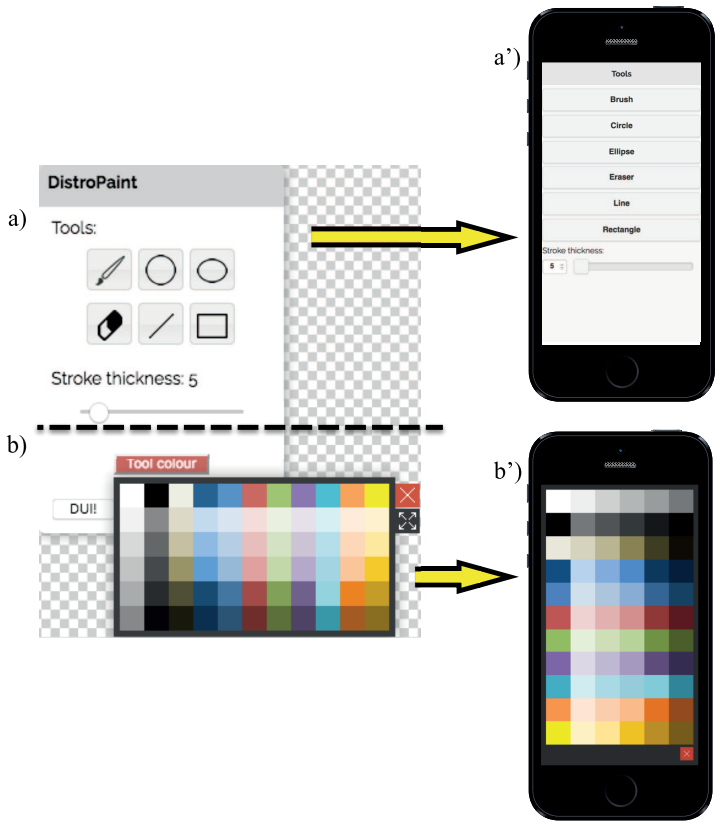


Figure 8.3: Functional Cores division for the toolbox: (a) tools core, (b) colours core, and their respective mobile formats (a') and (b').

tablet, but only the tablet can display both at the same time. Although this can also be achievable for the phone, we want to demonstrate that despite the capabilities of the devices (in this case, the difference in screen sizes), it is desirable to offer alternatives, so users can accomplish theirs tasks in one way or another.

- **Traceability:** Besides the already explained presence system, DistroPaint also gives feedback to the users about where the widgets are being distributed and also maintains synchronised all the values for all the widgets from the toolbox, no matter from where or when the user changes such values (see Figure 8.6).

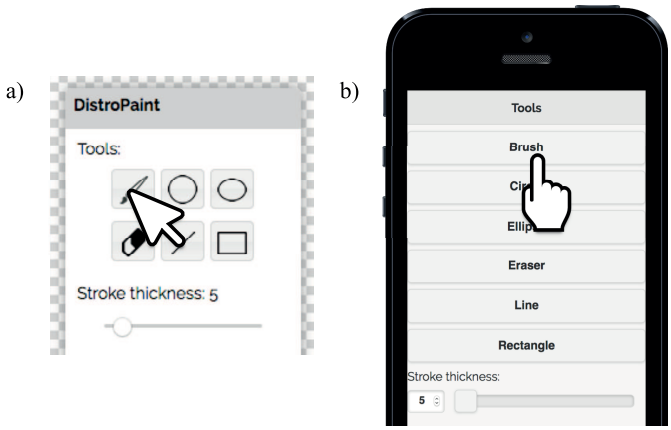


Figure 8.4: DistroPaint allows interaction through: (a) a mouse, and (b) with a finger; with both modalities the user can obtain the same result.

8.2.2 *Evaluation and Results*

The evaluation has been worked out with the help of five UX experts. We chose the experts for their experience applying usability tests, and because they are familiar with the topics of our research. All the experts are university professors and have postgraduate studies; two of them belong to our university. Their experience comes from both work in industry and research centres. It should be noted that none is related to this work in addition to their participation in the evaluation.

Before starting the evaluation, we gathered and explained to the experts each of our heuristics, their purpose, and discussed some examples so that everyone had a similar starting point. Each expert drafted a list of problems and violations of the heuristics that we propose. Once the evaluators have identified potential consistency problems, the individual lists have been consolidated into a single master list. The master list was then given back to the evaluators who independently have assessed the severity of each violation. The ratings from the individual evaluators are then averaged, and we present the results in Table 8.1. For the rating, we adapted the severity classification proposed by J. Zhang et al. (2003):

0 - Not a consistency problem at all.

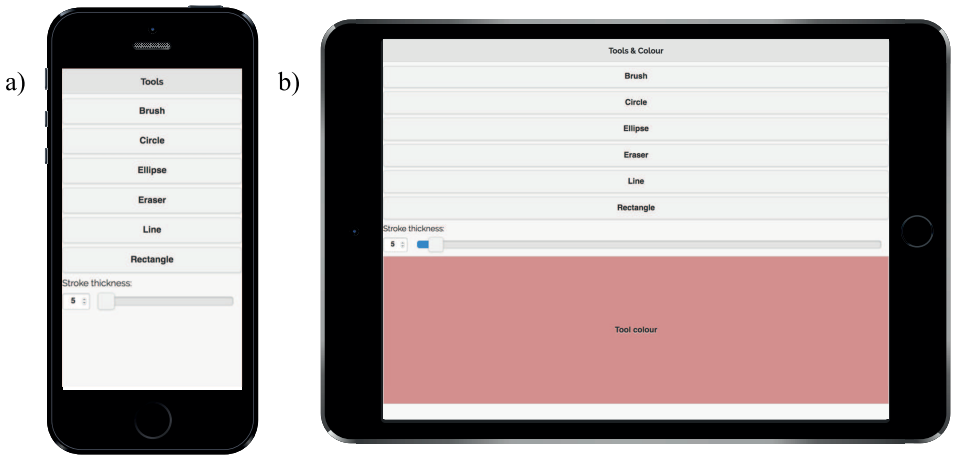


Figure 8.5: Functional Cores can be seen: (a) one at a time on the phone; (b) both of them at the same time on the tablet. The reason to do this is that the tablet has a bigger screen, thus, it can display more widgets.

1 - Cosmetic problem only. No need to be fixed unless extra time is available.

2 - Minor consistency problem. Fixing this, should be given a low priority.

3 - Major consistency problem. Important to fix, should be given a high priority.

4 - Consistency catastrophe. Imperative to fix this before the product can be released.

Evaluators found a total of 23 usability problems using our heuristics (a mean of 4.6 problems per evaluator). The severity rating of problems had an average of 2.42. For the master list, a total of 10 problems were evaluated and heuristics were violated 18 times (see Figure 8.7). Honesty and Traceability were the two most frequently violated heuristics, 6 and 4 times, respectively. In contrast, the heuristic with less detected problems was Functional Cores with 2 violations.

With respect to the severity of the problems detected, we can see in Figure 8.8 that severity level 3 - "Major consistency problem" was the most frequent with 36%, closely followed by severity level 2 - "Minor consistency problem" with 24% of occurrence. On the

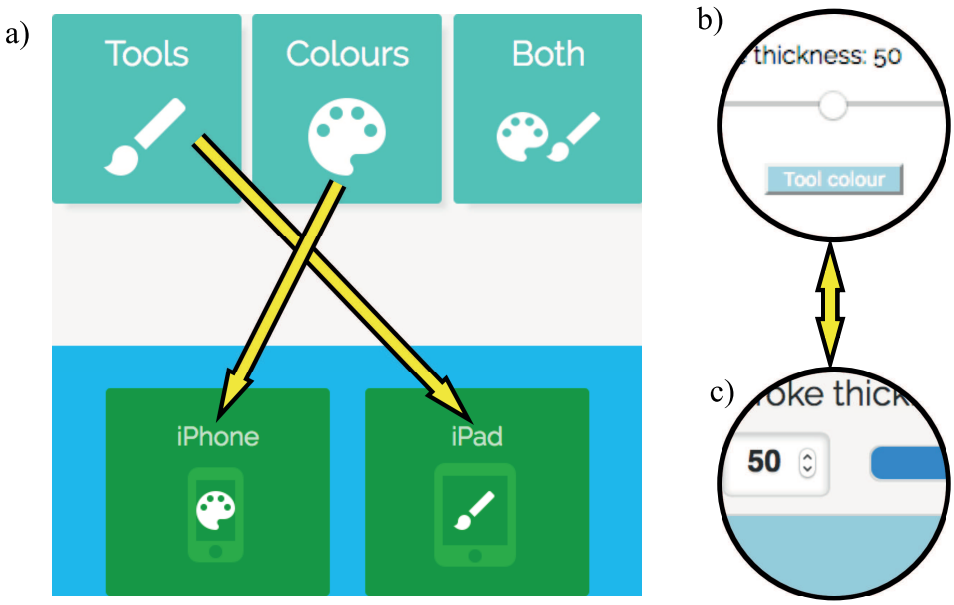


Figure 8.6: (a) As part of the presence system, the user knows where the widgets are. When the user makes a change in a widget, the system automatically reflects such a change in all the GUIs, e.g., tool, stroke thickness, and colour are synchronised between: (b) the PC and (c) the tablet.

contrary, we can notice that the lowest classification 0 - "Not a consistency problem at all" got 8%.

8.2.3 *Discussion*

In general, we can say that DistroPaint has many aspects in which to improve because several problems with severe qualifications were identified. Nevertheless, the evaluation was fruitful, as various problems could be discussed, as well as scenarios that, if neglected, could cause conflicts in the future. So our heuristics were advantageous in identifying particular conflicts in this specific case.

That Honesty was the heuristic with the highest number of violations is an exciting aspect. Perhaps improving those weaknesses of design, the violation of the other heuristics disappears, or its qualification is reduced because Honesty brings with it a better workflow and a more solid GUI.

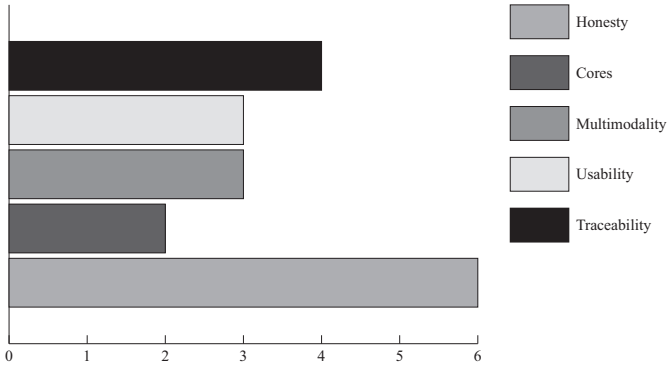


Figure 8.7: Heuristics violations in DistroPaint.

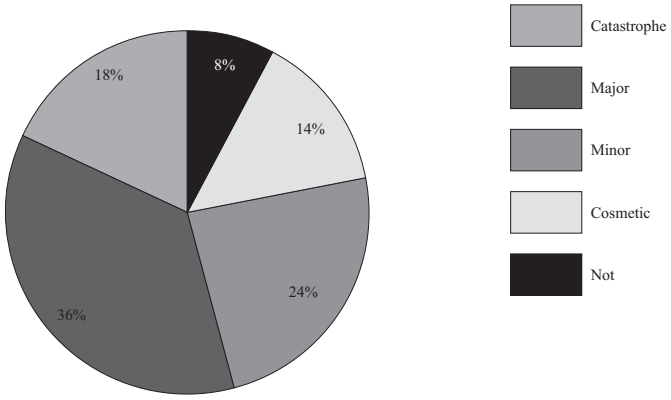


Figure 8.8: Severity rating of consistency problems found in DistroPaint.

The experts concurred that the heuristics could be a useful tool to detect consistency problems. However, they also acknowledged that in order to be more effective, they have to be refined and detailed.

8.3 CASE STUDY: SPOTIFY

In this section, we focus on Spotify (Spotify, 2018), which was chosen as a case study because it is a well known commercial application, and many users around the world use it.

Spotify is a cross-platform application for playing music via streaming. It allows users to play individual songs as well as playback by artist, album, or playlists created by other Spotify users. Data is streamed from both servers and a peer-to-peer network. There are clients for Mac OS and Windows along with several smart-

phone platforms and other devices, like video games consoles and Internet-connected speakers (see Figure 8.9).

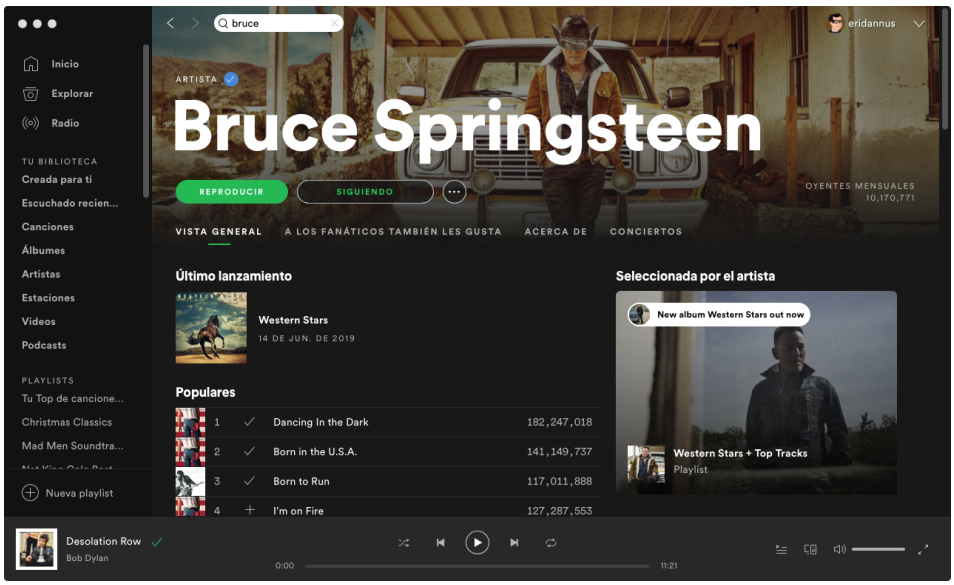


Figure 8.9: Spotify PC GUI.

The GUI is similar to those found in other desktop music players. Spotify offers the possibility of listening to music on devices that users have linked to their account. This is achieved through a list of devices that can appear on the desktop or mobile clients. Thus, the user can play songs from one device and control the playback from a different one (Kreitz & Niemela, 2010).

This evaluation was carried out under the same procedure as in the case of DistroPaint (see Section 8.2.2).

8.3.1 *Results*

A total of 10 problems were detected, and heuristics were violated 23 times. *Traceability* and *Multimodality* were the two most frequently violated heuristics, six and seven times, respectively (see Figure 8.10). In contrast, the heuristics with less detected problems were *Usability Limitations* and *Functional Cores* with four and one violations respectively.

Concerning the severity of the problems detected, we can see in Figure 8.11 that severity level 2 - "Minor consistency problem" was the most frequent with 30%, closely followed by severity level

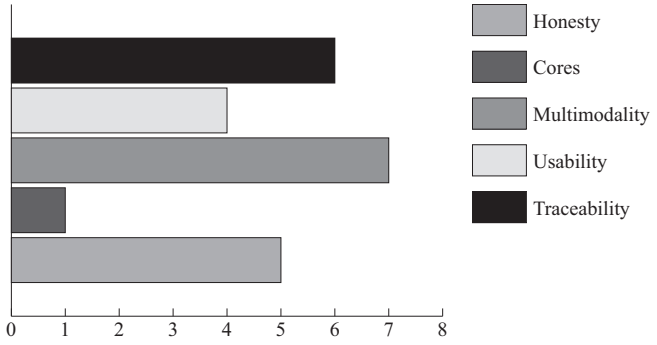


Figure 8.10: Consistency violations in Spotify.

1 - "Cosmetic problem only" with 24% of occurrence. On the contrary, we can notice that the severest classification 4 - "Consistency catastrophe" just got 10%.

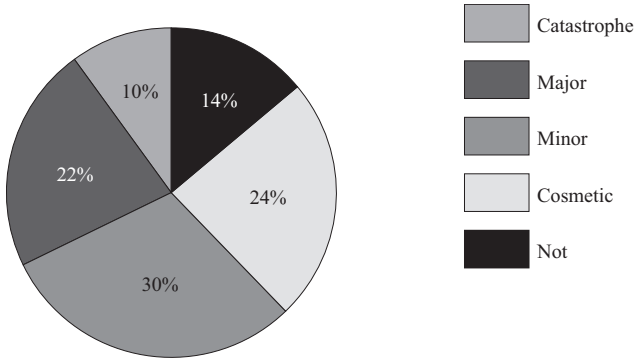


Figure 8.11: Severity rating of consistency problems found in Spotify.

8.3.2 *Discussion*

In general, we can say that Spotify got positive evaluations because the most severe classifications were few. The evaluation was also fruitful, as several problems could be discussed, as well as scenarios that, if neglected, could cause conflicts in the future.

Something that we could emphasise is that the heuristic of *Functional Cores* was only violated once. This remark could tell us that the widgets were well designed and that they fulfil their function, homogeneously, through the devices. Contrarily, *Multimodality* was the heuristic that more violations accumulated; this did not represent a

surprise, because the more devices an application encompasses, the harder it will be to replicate the functionalities in each one.

8.4 CASE STUDY: HOME SECURITY SYSTEMS

We chose three home security systems: *Ring Video Doorbell 1*, *Nest Hello*, and *Eufy Doorbell* because of their popularity in the market. The three systems are similar to each other; All three are video intercom systems that have WiFi connectivity, video streaming, two-way audio, motion alerts, and they are integrated into their proprietary security system that is controlled from a mobile app (see Figure 8.12).

(a) Ring (b) Nest (c) Eufy

Figure 8.12: Ring 8.12a, Nest 8.12b, and Eufy 8.12c are wall mounted devices.

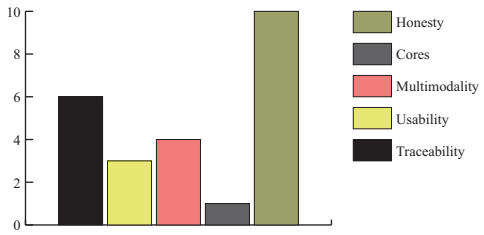
This evaluation was carried out under the same procedure as in the case of DistroPaint and Spotify (see Section 8.2.2). The experts took the systems home and tested each one for a week. We ask them to keep a diary of their experiences, noting, among other things, the problems they encountered, the characteristics they liked, and the possible failures they might experience. They were always taking into account our heuristics.

8.4.1 *Results*

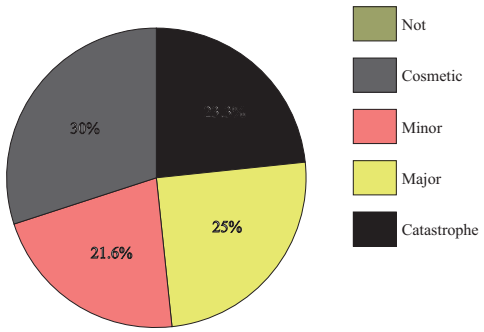
Evaluators found a total of 15 consistency problems using our heuristics (a mean of 3 problems per evaluator). The severity rating of problems had an average of 2.42 (3.1 for Ring, 1.7 for Nest, and 2.5 for Eufy). For the master list, a total of 10 problems were evaluated

and heuristics were violated 24 times. Honesty and Traceability were the two most frequently violated heuristics, 10 and 6 times, respectively. In contrast, the heuristic with less detected problems was Functional Cores with 1 violation (see Figure 8.13a).

With respect to the severity of the problems detected, we can see in Figure 8.13b that severity level 1 - "Cosmetic problem only" was the most frequent with 30%. On the contrary, we can notice that the lowest classification 0 - "Not a consistency problem at all" got 0%.



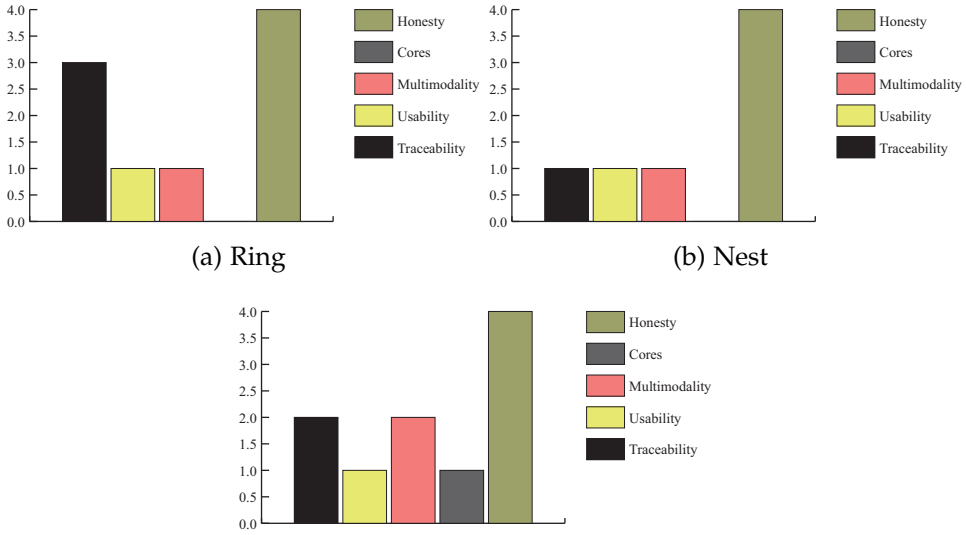
(a) Heuristics violations



(b) Severity rating

Figure 8.13: General results of our heuristic evaluation.

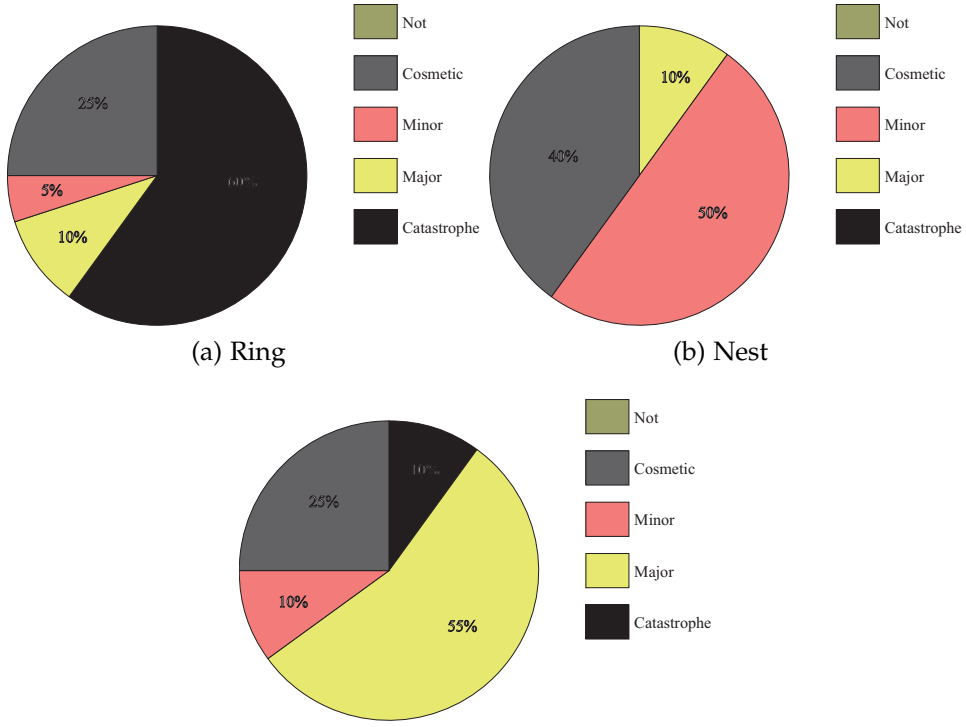
For the individual evaluations of the systems, it is notorious that the heuristic in which more problems were consistently found was Honesty, while Functional Cores was only violated once in the case of Eufy (see Figure 8.14). Interestingly, severity ratings vary diametrically in all systems. For example, the most severe "catastrophe" rating occupies 60% in the case of Ring, while in Nest nothing was rated in that range, and Eufy only obtained 10% (see Fig 8.15).



(a) Ring (b) Nest

(c) Eufy

Figure 8.14: Heuristics violations in Ring 8.14a, Nest 8.14b, and Eufy 8.14c.



(a) Ring (b) Nest

(c) Eufy

Figure 8.15: Severity ratings in Ring 8.15a, Nest 8.15b, and Eufy 8.15c.

8.4.2 *Discussion*

In general, we can say that the worst-rated application was Ring since it was the one that obtained a good part of its ratings in the most severe range. On the contrary, Nest obtained the best marks, since 50% of its problems were classified as minor.

It is no coincidence that the heuristic with the highest number of violations was Honesty, as it relates to the problem that afflicted all our evaluators in all systems, false alarms and the excessive amount of notifications. This is a severe problem because according to our evaluators, it was the leading cause that sometimes they felt anxious, and the sensation of the siege came. The systems tried to mitigate this by configuring the sensors and recognising faces, but none of these measures was helpful.

It is a complex challenge. A good part of the solution lies in improving the AI of these types of systems, but it is not the only thing that could be done. Improving the design of the doorbells by themselves as well as the way to install them could help people living in a particularly busy street, this could lead to more effective handling of notifications. More transparent controls and configurations would also be of great help so that users could choose the settings that best fit their environment and thus obtain a positive UX.

Table 8.1: Consistency problems and its rating in DistroPaint.

† Honesty (H), Functional Cores (F), Multimodality (M), Usability Limitations (U), Traceability (T)

Place	Problem	Heuristics†	Severity
Tools	On the PC, the buttons of the drawing tools contain icons, while in the mobile widget they are texts.	F	3.8
	The buttons on the mobile widget for the drawing tools are too small when viewed on the phone.	F, M	3.6
	If the user reloads the main page of DistroPaint or the distribution menu, all changes and configurations will be lost without previous warning.	H, T	2.2
	There is no feedback on the actual selected drawing tool among the devices.	H, T	1.4
	The buttons of the toolbox, in the main page of DistroPaint, are too small on the phone; also, the toolbox is too big, reducing the space available for the canvas.	M, U	3.2
Loading screen	Without previous explanation, the loading screen might confuse some users.	H	1.2
Distribution	The distribution menu is only accessible through the PC.	H, M, U	2.2
	Without previous explanation, the colours of the presence system might be unintelligible.	H, T	2.4
	Without previous explanation, the user has no way to know why the widget "Both" cannot be distributed into the phone.	H, U	3
	If a user closes the tab in a mobile device while this has a widget designated, such designation is not lost, but the user does not have a clear feedback of this.	T	1.2

Table 8.2: Consistency problems and its rating in Spotify.

† Honesty (H), Functional Cores (F), Multimodality (M), Usability Limitations (U), Traceability (T)

Place	Problem	Heuristics†	Severity
Devices list	For a device to appear in the list, it has to be unlocked and with the client in the foreground	H, T	1.8
	Only the available devices are displayed, if there is none, the list cannot be displayed	H, T	1
	In the PC application, there is no settings option, just the icon next to the volume. The user has to click on it to find other devices on their network	H, T, M	2.4
	If a user wants to see a history of the devices with access to their account, to consult or withdraw the permission, they have to do it in the web version, there are no alternatives	H, M	1.6
	Sometimes, devices are not shown if they are not on the same WiFi network, in some cases they do. However, it is specified that all elements of the interactive space have to be on the same network.	H, M, U	2.6
Login	To associate a device with an account, the user has to log in, so in devices such as televisions where the keyboard is not as intuitive as in a PC, this interaction can become cumbersome	F, U	1.6
Native clients	Not a problem *per se*; being all the GUIs native clients, many consistency problems are reduced; the development of each client carries a cost. Also, the user has a rather closed environment	M	0
Devices	For some devices (e.g., speakers) there is not a syncing process to manually add them. It either connects or does not	M, T, U	3.6
	Some devices need a dongle to be compatible with Spotify, thus the user must be sure that their system is powered up and turned to the correct input before listening to any music	M, T, U	3.4
	Devices with no GUI (e.g., speakers) will need a mobile device as a remote control	M, T	1

Table 8.3: Consistency problems and its rating in Home Security Systems.

† Honesty (H), Functional Cores (F), Multimodality (M), Usability Limitations (U), Traceability (T)

System	Problem	Heuristics†	Severity
Ring	It does not fulfil its doorbell function at all; As the sound comes from the device itself, it can only be heard outside, and sometimes notifications to my phone arrived long after the person had ringed (up to 10 minutes later).	H, M, T	4
	The system never alerted me that the batteries were running out. I only knew it when in a long time, I did not receive any notification and went to check.	H, U, T	4
	No matter how you set the sensitivity of the camera to start recording, it began to do so only when a person was very close to the door, or when they were leaving.	H, T	3.2
Nest	Possibly it has the most sensitive sensor of the three systems, although I put it to a minimum, notifications of movement were too many, reaching the point of being exasperating.	H, U	2.2
	Facial recognition can be a useful feature, but it was wrong a couple of times since it identified a stranger as if he were a relative.	H, T	1.4
	On several occasions I could watch the video stream from my phone without any problem, however, when consulting that clip stored in the cloud, people appeared and disappeared suddenly, it was clear that the video was cut, I could not know what the problem was.	H, M	2.2
Eufy	When you get a notification that there is activity and you tap on the notification it takes you to the live view instead of what it recorded.	H, M, U	2.6
	I installed the application on my phone and also on my husband's. Only one person can log in to the service at the same time, i.e., we can both watch video streaming, but only one of us receives notifications when someone rings.	H, F, M, T	3
	Sometimes the applications notified me of motion alerts, especially at night, but the video stream showed nothing. This happened even if I deactivated said movement alerts.	H, T	3.4
All	Interactions with people who ring the doorbell can become awkward, and potentially dangerous, as one is speaking as if one were at home when it might not be so. As the person who rang now knows that the house is alone.	H	1

9 DISCUSSION

According to Johnson (2014a), our world's perception is biased, as we do not perceive what exists out there. This bias comes from the past (our experience), the present (the current context) and the future (our goals). These influences are also reflected in works such as the Theory of Planned Behaviour (Ajzen, 1991) that postulates that behaviour is the consequence of the attitude (behavioural beliefs), beliefs about the normative expectations of others (normative beliefs), and beliefs about the presence of factors that may facilitate or impede the behaviour performance (control beliefs) (see Figure 9.1).

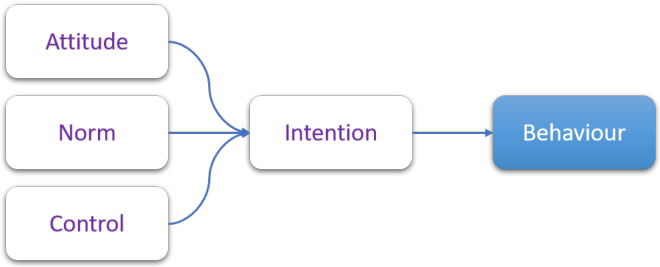


Figure 9.1: Simplified theory of planned behaviour (Ajzen, 1991; Nik, 2021).

In this way, beliefs and attitudes go together, leading to intention, that links directly to behaviour (see Figure 9.2). According to Yocco (2016), if we account for these factors in our designs, we should predict with relative accuracy what a user's end behaviour will be: the use of our artifact.

The implications of the behavioural study in GUIs design include concepts like guiding users to their goals, letting people use perception rather than calculation, and making the artifact familiar (Johnson, 2014b).

Making clear the aspects that influence the behaviour, we can discuss the aspects that impact the UX according to our research.

Section 2.3.3 already discussed the four main periods: AUX, MUX, EUX and CUX. This work concentrates on AUX, as it is the least

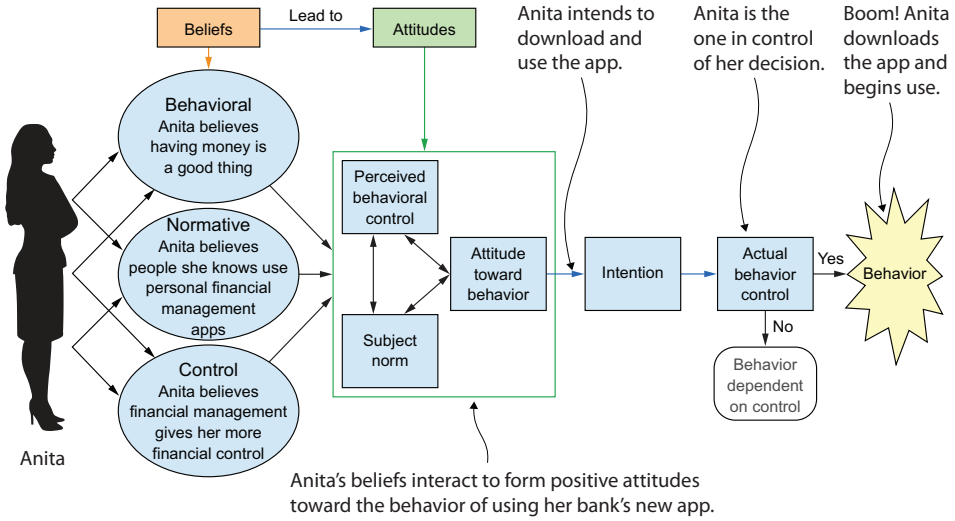


Figure 9.2: Components of planned behaviour. Beliefs lead to attitudes, which create intention and then behaviour, if the individual is in control of the behaviour (Yocco, 2016).

explored period of the four. We also contrasted AUX and EUX in one of our studies. Why don't we compare between other periods? Of course, this would have been more helpful, as it would probably provide a more illuminating picture, but the details of each period must be taken into account.

On the one hand, doing MUX evaluations are a complex task; They can be approached from the qualitative aspect, with techniques such as "Thinking aloud" (Soure et al., 2021) or from the quantitative view, with an eye-tracking contraption (de Souza et al., 2021). However, these techniques are difficult to analyse or require rather costly equipment.

On the other hand, CUX evaluations are expensive since participants are usually required to keep a diary of experiences (Tulaskar & Turunen, 2021) and attend interviews from time to time (e.g., weekly) (R. Y. Wong, 2021). For this, evidence is collected with photographs, audio and video. So the cost is not only in the materials used but also in the human resources (trained professionals) required to carry out the assessments. Not to mention the specific requirements of each period studied in the evaluations.

Now, choosing to study a particular period of UX is not enough; It is necessary to plan how to study it. To do this, as can be seen

in Figure 9.3, we adopted three approaches that allowed us to obtain information from all points of view that involve usability/UX evaluations.

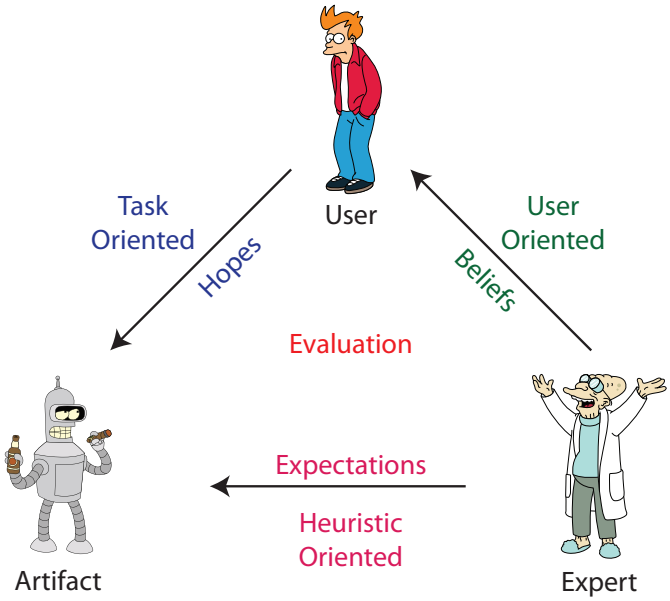


Figure 9.3: AUX study orientations and what we got from each one.

In our tasks evaluations, we study users' hopes because when they find themselves in unknown circumstances, they trust their previous experiences, hoping that this knowledge will help them. Furthermore, user oriented studies revealed their beliefs, what they considered suitable and desirable. Finally, the heuristics that we develop meet users' expectations and good practices found in the state of the art. Of course, these concepts, like UX itself, are dynamic and could all be obtained from a single evaluation.

An important detail to consider is the observations we had from our participants. Figure 9.3 shows the duality of users in evaluations since they can act as evaluators (Task Oriented) or as evaluatees (User Oriented). In the former role, users are more perceptive and demanding, as they trust that the artifact is well designed. In the latter one, users are more cautious, expressing their opinions less securely and taking more time to think (see Figure 9.4). This expands on what has been discussed in similar works (Følstad, 2017; Weinschenk, 2010).

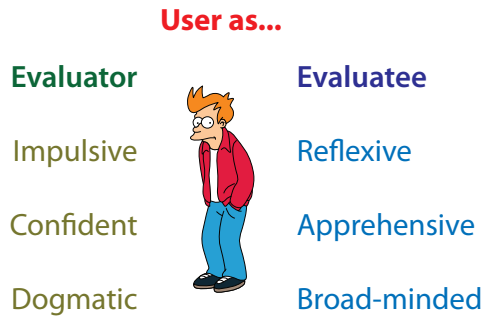


Figure 9.4: The duality of the user in usability/UX evaluations.

In this way, our discoveries have added value for understanding user behaviour (see Figure 9.5).

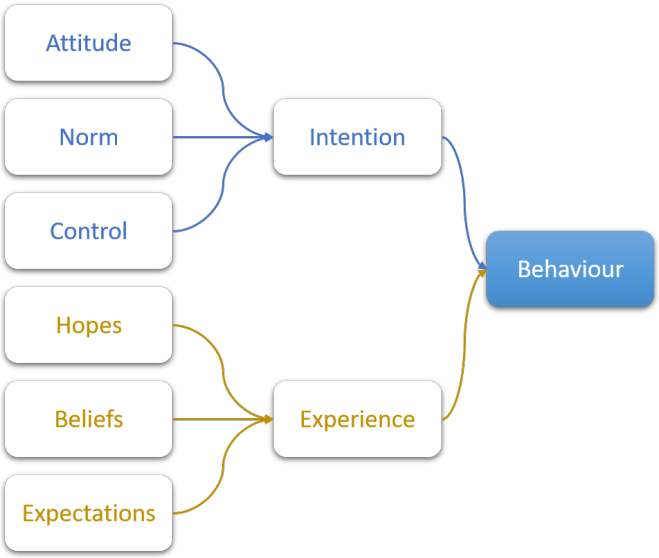


Figure 9.5: AUX elements can help study user behaviour.

Returning to the initial premise that we adapted from Le Corbusier, an artifact produces happiness when it is useful and beautiful. We can argue that the utility part is relatively simple because, as we already explained, an artifact is useful when usability and utility are combined (see Section 2.2). When an artifact is beautiful? That is more complex to discuss.

According to Hassenzahl (2004), beauty largely depends on identification, a hedonic attribute group, which captures the artifact's ability to communicate significant personal values to others. Perceived

usability and goodness are affected by the experience, whereas hedonic attributes and beauty remain stable over time. Overall, the nature of beauty is rather self-oriented than goal-oriented, whereas goodness relates to both.

Tractinsky (2004) agrees with Hassenzahl's observations, adding that the discussion of æsthetic stimuli is exceptionally complex since the perception of beauty is conditioned by previous experience and cultural aspects. However, he also maintains that usability is not equivalent to beauty, although it certainly does impact.

D. Norman (2004) argues that there are different levels of beauty: surface beauty (visceral), beauty in operation (behavioural), and beauty in depth (reflective). However, this view was readily disputed by art historians and psychologists. Indeed, these discussions are not new. Not for nothing did we introduce the definitions of Joyce and Aquinas. So no matter how new the works that contribute to this exciting topic (Märtin et al., 2021; X. Wang et al., 2021), there is still no unanimous consensus on when an artifact is beautiful or what *is* beauty whatsoever.

Finally, as we have already stated, AI can be a vital element for the future of HCI, and that is why it deserves a brief reflection. The most complete and complex point of view, in our opinion, is the one presented by Harper (2019), who concludes that the future is not AI, but rather the enablement of AI through HCI, i.e., not only the application of technology should be understood, but the individuals who use that technology should be recognised. It is also necessary to create methods that harmonise users and AI systems. We should not take too much for granted the insides of these systems.

Our personal view of the role HCI should play in the AI era is about ethics. We, HCI researchers, have a responsibility to alert and hold back (if possible) harmful AI advancement. For example, privacy is a fundamental issue today, and we must take with a grain of salt the results that come from algorithms that result from the predictive analysis of users (Julien, 2012) or recommendation systems (Zhao et al., 2021) because, in many cases, it is not known to what extent they need sensitive user data. Another example is the bias of supposedly neutral algorithms. These can have a devastating impact when used in mass facial recognition, primarily if it is a government implementation (Andrejevic & Selwyn, 2020; Raji et al., 2020).

This view may seem alarmist or even baseless. However, we offer two compelling arguments that support our point of view. The former one comes from The Centre for the Study of Existential Risk (CSER), a research centre at the University of Cambridge. The CSER has a whole area dedicated to AI risks and tries to create forums in specialised journals and conferences, with the support of professional societies such as IEEE and ACM, to expose this issue[1]. The latter argument comes from one of the most prestigious scientists in the area, Peter Norvig, co-author of the book "Artificial Intelligence: A Modern Approach". Norvig said that the most critical questions in AI today are user-centred, and he posed some inquiries aimed at students and researchers: *Whose interests are you serving? Are you being fair to everyone? Is anyone being left out? Is the data you collected inclusive, or is it biased?* (Lynch, 2021).

9.1 LIMITATIONS

The sample size and its representativeness render a constant challenge in all usability and UX studies. Ideally, all samples will be random, with archetypical target people and large enough to draw conclusions based on rigorous statistical analysis. However, as is well known in all fields dealing with human participants, the ideal scenarios are very far from reality (Baxter et al., 2015; Sauro, 2010b). In all cases, testing with human participants is restricted by the available budget (see Figure 9.6).

In our field, a general rule of thumb is that the more focused the context of an investigation, the smaller the sample can be used (see Figure 9.7). For example, Sauro and Lewis (2012a) mention that for summative studies, depending on the kind of evaluation carried out, groups of up to 26 people can be used to obtain levels of confidence of 99%. Similarly, formative studies can have up to 20 participants for 99% confidence (Sauro & Lewis, 2012b).

In addition to the number of participants, the form of recruitment is also essential. Despite the biases they can cause, non-probabilistic samples such as convenience are widely used in usability and UX studies (J.-S. Chen et al., 2021; Cheng et al., 2021; Kairy et al., 2021; Karani et al., 2021; Luctkar-Flude et al., 2021).

1 https://www.cser.ac.uk/research/risks-from-artificial-intelligence/

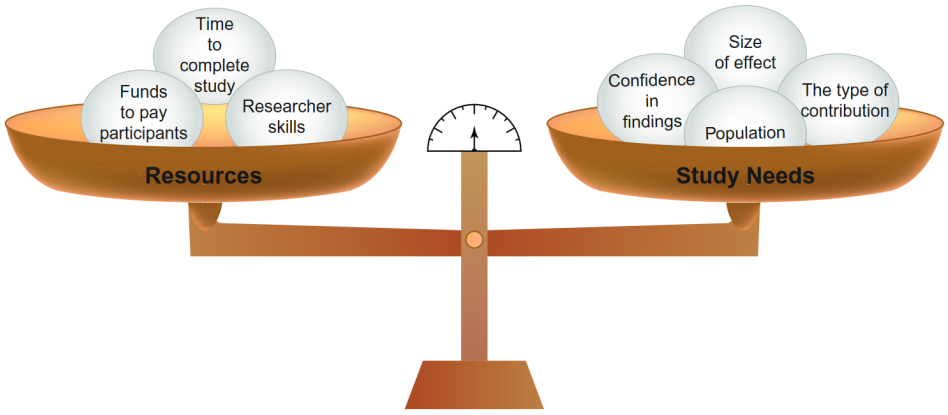


Figure 9.6: Weighing resources with study needs (Baxter et al., 2015).

In this way, we can see that our numbers of participants and sampling, although not free from bias, are based on standard practices in the state of the art. Specifically, we can comment on the limitations in the nature of our participants: we had a small number of women in our studies. Apparently, there are no significant differences between men and women in UX perception (Aufderhaar et al., 2019) and since none of our studies is focused on particular sex, we can classify this bias as minor.

The most notable limitation is the population of graduate students who participated in our contrast assessment (see Chapter 5). Their mostly pragmatic perception may be related to their education being strictly attached to STEM fields. However, thanks to related research (Kim et al., 2013), we can say that our findings are still valuable, as we could replicate the results in a different context, i.e., different participants, different time frames, and different objectives.

Regarding the tools we used in our evaluations, i.e., AttrakDiff and NASA-TLX, our limitation is in their development and application contexts. Hernández-Sampieri and Torres (2018) indicate that the improvisation of instruments generates few valid and reliable results. That is why we choose our tools with caution. As we have already explained, we decided to use them because they have been implemented in various evaluations over time and are recognised as valid and reliable instruments (Castro et al., 2021; I. Díaz-Oreiro et al., 2021; Febiyani et al., 2021; Miyake, 2020; Müller et al., 2021; Ribeiro & Providência, 2021).

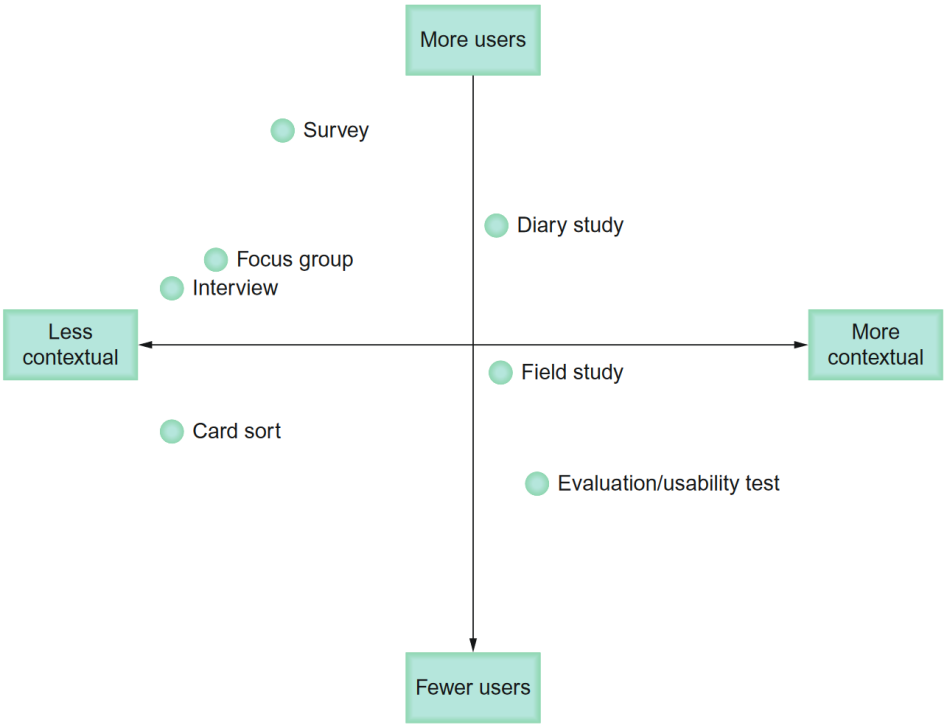


Figure 9.7: Graphical representation of number of participants required by context (Baxter et al., 2015).

For the particular case of our evaluation with NASA-TLX (see Section 6.4.1) we used a version in Spanish translated by the *Ministry of Labour and Social Affairs of the Government of Spain* (Díaz Ramiro et al., 2010) and the *Ministry of Labour and Social Welfare of the Government of Chile*[2]. A context validation is that the tool was used in a study involving university professors from northern Mexico (Jiménez & Thamar, 2019).

We apply our other tool, AttrakDiff (see Section 5.4) in English, as there are only additional versions in German and French (Lallemand et al., 2015). We can also talk about contextual validation since works have been published in Mexico that uses the tool (Iniguez-Carrillo et al., 2021).

We know that the application of these questionnaires can cause bias in our results. However, we do not know the existence of standardised tools for measuring usability or UX developed explicitly in or for Mexican populations.

2 https://ergomedia.isl.gob.cl/app_ergo/nasatlx/nasa-tlx.pdf

Finally, it is worth mentioning that there are various usability sets. We decided to use Nielsen's heuristics (Nielsen, 2020) because they are the best known and used in this type of evaluation. Other alternatives would have been those proposed by: Johnson (2008), D. A. Norman (1983), Shneiderman and Plaisant (2010) and Stone et al. (2005). The common element of all these proposals is that they are concise, direct and clear enough to cover a broad spectrum without losing their nature as heuristics. We found no evidence that any specific element in any of these sets affects the evaluation result, e.g., the number of elements in each set or the level of granularity of each heuristic.

10 CONCLUSIONS AND FUTURE WORK

When users choose an application not only for its functionality but for how it makes them feel and how well they can express themselves, it is clear that providing a positive UX is vital to the survival of any artifact. This is not an easy challenge because, although various evaluation methods and metrics exist, UX is still a relatively young field and full of exciting challenges. One of those challenges is exploring AUX because, as we have already mentioned, it is a little-explored area.

Despite our limitations in non-localised assessment tools and our opportunity sampling, the contribution of our studies expands the frontier of AUX knowledge. Knowing users' ideas, expectations, and previous experiences, either directly, i.e., through studies with end-users, or indirectly, i.e., knowledge compiled using heuristics, will allow developers to create products of higher value.

Regarding the epistemic value of our proposals, we can mention that now there is more evidence of how to work in AUX evaluations, how it is evaluated, who evaluates it, what is evaluated, and how the results are presented.

Below we present the conclusions and some ideas for the future of our particular studies.

10.1 TASK ORIENTED

UX evaluation is always valuable, regardless of the nature or purpose of the evaluated artifact. Our proposal compares the AUX and EUX of user-tools through daily tasks in social networks. Our tests revealed that our participants build their expectations with pragmatic criteria, i.e., hedonic and attractiveness aspects were secondary when they were building their prototypes.

Our research contributes to further increasing the understanding of UX, how perceived experiences are measured, and which factors are most relevant at a certain point in an evaluation or development. Our results quantitatively confirmed that AUX seems to be mainly composed of pragmatic aspects. The development of this idea could

lead to improving existing evaluation methods and the creation of new ones.

10.2 USER ORIENTED

To study the expectations that are underlying in the end-user requirements, we introduced a chatbot to support the teaching/learning process in a middle school environment. The work presented here is the first iteration of the GDS methodology that we adapted. This methodology allowed us to know the wishes and requirements of the end-users in the profiles of student, teacher, and administrative staff.

We know well that the tests and the population we used is scarce, but our purpose was to perceive the acceptance of end-users: observe whether the chatbot met their requirements, whether it has the possibility of being a tool for everyday use, and above all whether the features it has were the indicated ones to improve communication between stakeholders. We believe that this first iteration was helpful and will allow us to move forward with the development.

10.3 HEURISTIC ORIENTED

This section contains the learnings we got from our pair of heuristic sets.

10.3.1 *Consistency Heuristics*

Originally our coherence heuristics were intended to evaluate systems with a distributive nature, i.e., Distributed User Interfaces (DUIs) (Melchior et al., 2011) and Meta-UIs (Coutaz, 2007), which allow the user to send "chunks" of the same GUI to various devices. These developments pose very complex technical challenges, since we have to think about the application's design *per se* and what happens in other devices simultaneously. Thus, the design of interactions becomes multidimensional.

Multi-device applications present the challenge of configuring the available resources and their role in the environment. When the users control the application, it allows them to explore their environment, identify the tasks and services compatible with it, and

combine independent resources in a significant manner, in order to perform tasks and interact with services. Consistency is the element that maintains the users in a stable base, since it is the key to assist GUI distribution. Besides, it is an essential factor in maintaining a positive UX.

With the development and evaluation of DistroPaint, a proof of concept, we learned that our proposal could be used beyond their original conception. In this way, we chose Spotify as an excellent alternative to test our set of heuristics, not only because it is no longer an *ad hoc* context, but because it is a popular application whose users have well-rooted daily interactions, thus the challenge of finding problems would fall solely on our heuristics.

Since the test with Spotify was successful, we ventured to try our luck with systems somewhat further away from the original domain of our proposal: home security systems. From this test, we learned that our set of heuristics could be applied to evaluate IoT systems. This is exciting, as IoT offers engaging scenarios and challenges to integrate and improve our heuristics. It was also proven that our approach was the correct one, as we never wanted to focus only on GUIs but on maintaining consistency in highly interactive multimodal systems.

We consider that the challenge of identifying AUX factors in multimodal systems is quite complex since it may be the case that in the same interactive environment, there is a device that is very familiar to the user and one that is utterly unknown to the user.

10.3.2 *Conversational Heuristics*

We presented a novel set of usability heuristics for evaluating chatbots. Using a case study that included an educational chatbot and the help of five experts, these heuristics were put to test.

The case study revealed the problems caused by violating the heuristics. It is estimated that this evaluation was successful, as concise problems and directly related heuristics were identified.

Evaluating the usability of a chatbot is a complex challenge. A part of the solution lies, to a large extent, in improving the AI mechanisms of these types of systems, but it would be a simplistic approach to take just that into account. Another essential part is the context of the use of the chatbot; Most likely, a general evaluation mechanism is not enough to cover all the contexts in which a chatbot can be

used because, while in one scenario a characteristic is desirable, in another, it may be the opposite. However, some mechanisms can always be kept independent of the context, and that is where the proposed heuristics can be most helpful. If the interaction base elements of a chatbot have positive usability, this can lead to fewer UX problems in the future when the chatbot acquires characteristics of its environment.

The experts were satisfied with the heuristics as they allowed them to focus on their evaluation and identify problems more efficiently. However, they also suggested that heuristics should be refined, as they can be understood and applied more effectively in this way. This represents a fundamental challenge in the future since, as it was seen in Section 7.5, heuristics were misinterpreted in 15.78% of the cases (see Figure 7.1). Although a slight refinement was made to the H1 heuristic (see Section 7.6), it is necessary to improve them so that they are apparent to all evaluators at all times.

This work can be considered an exploratory nature investigation; we know that a heuristic evaluation does not constitute a complete usability evaluation but is simply one stage in an entire evaluation and design process. It is known that a single study is insufficient to obtain solid conclusions. However, the results obtained were satisfactory since they demonstrate that, although the heuristics need improvement, it is possible to use them in an evaluation.

A first step was established to create usability assessment methods in the little-explored terrain of chatbots. It is judged as a priority to identify the possible characteristics that have the most significant weight in the usability of chatbots, both the context-dependent and independent ones.

10.4 FUTURE WORK

As future work, we intend to replicate our AUX vs EUX tests, but this time with children. As Moser et al. (2014) work suggests, children can build prototypes with hedonic aspects in mind, i.e., we would expect to obtain results opposite to what we found. We also consider it essential to use other questionnaires besides Attrak-Diff, e.g., UEQ (Schrepp et al., 2017), SUS (Bangor et al., 2008) or Attrak-Work (Väätäjä et al., 2009), which would help validate our conclusions quantitatively. While in this work we focused on social networks, our assessment method can be used in multiple areas.

As for the chatbot development, all user stories of our personas suggest that the chatbot should be easily accessible from their smartphones. We consider the possibility that, in the future, we can provide the chatbot service through one (or some) of the most popular instant messaging applications on the market, e.g., WhatsApp, Messenger, Telegram. Of course, security and privacy issues should be taken into account. A possible alternative is to offer a limited service, i.e., that it does not include sensitive information, but merely academic activities in general.

To enrich the conversational heuristics, we intended to do tests to obtain expectations and ideals that younger populations have regarding this type of systems, e.g., generations "z" and "alpha". In this way, we will get insights from the groups likely to use chatbots the most and non-expert users.

For all these evaluations, we plan to improve the recruitment of participants, trying to do it through some probabilistic sample without forgetting the representation of target users. A significant challenge also arises in standardising or translating usability or UX evaluation tools for Mexican populations.

Finally, we have the opportunity to present an integrating case of the three approaches shown during this work: tasks, users and heuristics. We intend to explore accessibility in video games, particularly elements such as dyslexia and colour blindness, although some developments take measures for daltonism, these elements are not prevalent, and there is always room for improvement. First, our AUX vs EUX contrast method will allow us to know, through tasks, the participants' previous experience with the type of elements proposed and how they compare with video games on the market. Then, we will develop a user-centred study to find out the beliefs they have regarding these accessibility integrations. Finally, with all this knowledge collected, we will complement it with what is already known in state of the art to create accessibility heuristics for video games.

BIBLIOGRAPHY

Acemoglu, D. & Restrepo, P. (2019). *Automation and new tasks: How technology displaces and reinstates labor* (tech. rep.). National Bureau of Economic Research. https://doi.org/10.3386/w25684

Administration, U. G. S. (2013). Heuristic evaluations and expert reviews. https://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html

Agus Santoso, H., Anisa Sri Winarsih, N., Mulyanto, E., Wilujeng saraswati, G., Enggar Sukmana, S., Rustad, S., Syaifur Rohman, M., Nugraha, A. & Firdausillah, F. (2018). Dinus intelligent assistance (dina) chatbot for university admission services. *2018 International Seminar on Application for Technology of Information and Communication*, 417–423. https://doi.org/10.1109/ISEMANTIC.2018.8549797

Ajzen, I. (1991). The theory of planned behavior [Theories of Cognitive Self-Regulation]. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/https://doi.org/10.1016/0749-5978(91)90020-T

Aladwan, A., Kelly, R. M., Baker, S. & Velloso, E. (2019). A tale of two perspectives: A conceptual framework of user expectations and experiences of instructional fitness apps. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 394:1–394:15. https://doi.org/10.1145/3290605.3300624

AlHammadi, A., AlZaabi, A., AlMarzooqi, B., AlNeyadi, S., AlHashmi, Z. & Shatnawi, M. (2019). Survey of iot-based smart home approaches. *2019 Advances in Science and Engineering Technology International Conferences (ASET)*, 1–6. https://doi.org/10.1109/ICASET.2019.8714572

Alshamari, M. (2016). A review of gaps between usability and security/privacy. *International Journal of Communications, Network and System Sciences*, *9*(10), 413–429.

Alvarado, M. & Esquer, G. N. (1997). Contextual knowledge and belief: Representation and reasoning. *Computación y Sistemas*, *1*(1), 21–26.

Alvarado, M. & Sheremetov, L. (2001). Interaction modal logic for multiagent systems based on bdi architecture. *Memoria del 3er Encuentro Internacional de Ciencias de la Computacion ENC01*, 803–812.

Andrade, F. O., Nascimento, L. N., Wood, G. A. & Calil, S. J. (2015). Applying heuristic evaluation on medical devices user manuals. In D. A. Jaffray (Ed.), *World congress on medical physics and biomedical engineering, june 7-12, 2015, toronto, canada* (pp. 1515–1518). Springer International Publishing.

Andrejevic, M. & Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology*, *45*(2), 115–128. https://doi.org/10.1080/17439884.2020.1686014

Anić, I. (2018). The importance of visual consistency in ui design [[Online; accessed Oct-2018] https://www.uxpassion.com/blog/the-importance-of-visual-consistency-in-ui-design/].

Apostolou, B., Bélanger, F. & Schaupp, L. C. (2017). Online communities: Satisfaction and continued use intention. *Information Research*, *22*(4).

Argal, A., Gupta, S., Modi, A., Pandey, P., Shim, S. & Choo, C. (2018). Intelligent travel chatbot for predictive recommendation in echo platform. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 176–183. https://doi.org/10.1109/CCWC.2018.8301732

Aufderhaar, K., Schrepp, M. & Thomaschewski, J. (2019). Do women and men perceive user experience differently? *IJIMAI*, *5*(6), 63–67. https://doi.org/10.9781/ijimai.2019.03.005

Aula, A., Khan, R. M. & Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 35–44. https://doi.org/10.1145/1753326.1753333

Banfield, R., Lombardo, C. T. & Wax, T. (2015). *Design sprint: A practical guidebook for building great digital products*. " O'Reilly Media, Inc."

Bangor, A., Kortum, P. T. & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bannon, L. (2011). Reimagining hci: Toward a more human-centered perspective. *Interactions*, *18*(4), 50–57. https://doi.org/10.1145/1978822.1978833

Bardzell, J. & Bardzell, S. (2016). Humanistic hci. *Interactions*, *23*(2), 20–29. https://doi.org/10.1145/2888576

Bargas-Avila, J. A. & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698.

Barnum, C. M. (2011a). 1 - establishing the essentials. In C. M. Barnum (Ed.), *Usability testing essentials* (pp. 9–23). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-375092-1.00001-5

Barnum, C. M. (2011b). 3 - big u and little u usability. In C. M. Barnum (Ed.), *Usability testing essentials* (pp. 53–81). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-375092-1.00003-9

Basogain, X., Olabe, M. Á. & Olabe, J. C. (2017). Transition to a modern education system through e-learning. *Proceedings of the 2017 International Conference on Education and E-Learning*, 41–46. https://doi.org/10.1145/3160908.3160924

Baxter, K., Courage, C. & Caine, K. (2015). Chapter 5 - choosing a user experience research activity. In K. Baxter, C. Courage & K. Caine (Eds.), *Understanding your users (second edition)* (Second Edition, pp. 96–112). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-800232-2.00005-5

Benotti, L., Martínez, M. C. & Schapachnik, F. (2014). Engaging high school students using chatbots. *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education*, 63–68. https://doi.org/10.1145/2591708.2591728

Bevan, N., Liu, Z., Barnes, C., Hassenzahl, M. & Wei, W. (2016). Comparison of kansei engineering and attrakdiff to evaluate kitchen products. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2999–3005. https://doi.org/10.1145/2851581.2892407

Bødker, S. (2006). When second wave hci meets third wave challenges. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles*, 1–8. https://doi.org/10.1145/1182475.1182476

Bødker, S. (2015). Third-wave hci, 10 years later—participation and sharing. *Interactions*, *22*(5), 24–31. https://doi.org/10.1145/2804405

boyd danah m., d. m. & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Braun, P. (2020). Attrakdiff, i feel so i am ? measuring affects tested by digital sensors. *Digital Klee Esquisses Pédagogiques. Enquête sur le futur de la forme. Présent Composé (Rennes). Les Presses du Réel (Dijon), p.140-154*. https://hal.archives-ouvertes.fr/hal-02984960

Carey, K. & Helfert, M. (2015). An interactive assessment instrument to improve the process for mobile service application innovation. In F. Fui-Hoon Nah & C.-H. Tan (Eds.), *Hci in business* (pp. 244–255). Springer International Publishing.

Carta, S., Podda, A. S., Recupero, D. R., Saia, R. & Usai, G. (2020). Popularity prediction of instagram posts. *Information*, *11*(9), 453. https://doi.org/10.3390/info11090453

Castro, S. C., Hosseinpour, H., Quinan, P. S. & Padilla, L. (2021). Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. https://doi.org/10.1109/TVCG.2021.3114803

Chen, J.-S., Le, T.-T.-Y. & Florence, D. (2021). Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, *49*(11), 1512–1531. https://doi.org/10.1108/IJRDM-08-2020-0312

Chen, L.-S. & Chang, P.-C. (2010). Identifying crucial website quality factors of virtual communities. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*, 17–19.

Chen, V. H. H. & Duh, H. B. L. (2009). Investigating user experience of online communities: The influence of community type. *2009 International Conference on Computational Science and Engineering*, *4*, 509–514.

Cheng, T.-H., Chen, S.-C. & Hariguna, T. (2021). The empirical study of usability and credibility on intention usage of government-to-citizen services. *Journal of Applied Data Sciences*, *2*(2). https://doi.org/10.47738/jads.v2i2.30

Chin, J. & Fu, W.-T. (2010). Interactive effects of age and interface differences on search strategies and performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 403–412. https://doi.org/10.1145/1753326.1753387

Chuan, N. K., Sivaji, A. & Ahmad, W. F. W. (2015). Usability heuristics for heuristic evaluation of gestural interaction in hci. In A. Marcus (Ed.), *Design, user experience, and usability: Design discourse* (pp. 138–148). Springer International Publishing.

Ciechanowski, L., Przegalinska, A., Magnuski, M. & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, *92*, 539–548. https://doi.org/https://doi.org/10.1016/j.future.2018.01.055

Clarizia, F., Colace, F., Lombardi, M., Pascale, F. & Santaniello, D. (2018). Chatbot: An education support system for student. In A. Castiglione, F. Pop, M. Ficco & F. Palmieri (Eds.), *Cyberspace safety and security* (pp. 291–302). Springer International Publishing.

Cohn, M. (2004). *User stories applied: For agile software development*. Addison-Wesley Professional.

Coutaz, J. (2007). Meta-user interfaces for ambient spaces. In K. Coninx, K. Luyten & K. A. Schneider (Eds.), *Task models and diagrams for users interface design* (pp. 1–15). Springer Berlin Heidelberg.

Coutaz, J. & Calvary, G. (2008). Hci and software engineering: The case for user interface plasticity. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications-human factors and ergonomics series* (pp. 1107–1118). CRC Press.

Coyette, A., Kieffer, S. & Vanderdonckt, J. (2007). Multi-fidelity prototyping of user interfaces. In C. Baranauskas, P. Palanque, J. Abascal & S. D. J. Barbosa (Eds.), *Human-computer interaction – interact 2007* (pp. 150–164). Springer Berlin Heidelberg.

Cunningham-Nelson, S., Boles, W. W., Trouton, L. & Margerison, E. (2019). A review of chatbots in education: Practical steps forward. *Australasian Association for Engineering Education 2019*. https://eprints.qut.edu.au/134323/

Davis, R. C., Saponas, T. S., Shilman, M. & Landay, J. A. (2007). Sketchwizard: Wizard of oz prototyping of pen-based user interfaces. *Proceedings of the 20th Annual ACM Symposium on*

User Interface Software and Technology, 119–128. https://doi.org/10.1145/1294211.1294233

De', R., Pandey, N. & Pal, A. (2020). Impact of digital surge during covid-19 pandemic: A viewpoint on research and practice. *International Journal of Information Management*, 102171. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2020.102171

de Oliveira, R. & da Rocha, H. V. (2007). Consistency priorities for multi-device design. In C. Baranauskas, P. Palanque, J. Abascal & S. D. J. Barbosa (Eds.), *Human-computer interaction – interact 2007* (pp. 426–429). Springer Berlin Heidelberg.

de Souza, K. E. S., de Aviz, I. L., de Mello, H. D., Figueiredo, K., Vellasco, M. M. B. R., Costa, F. A. R. & da Rocha Seruffo, M. C. (2021). An evaluation framework for user experience using eye tracking, mouse tracking, keyboard input, and artificial intelligence: A case study. *International Journal of Human–Computer Interaction*, 1–15. https://doi.org/10.1080/10447318.2021.1960092

Dey, S. & Hossain, A. (2019). Session-key establishment and authentication in a smart home network using public key cryptography. *IEEE Sensors Letters*, *3*(4), 1–4. https://doi.org/10.1109/LSENS.2019.2905020

Dian Sano, A. V., Daud Imanuel, T., Intanadias Calista, M., Nindito, H. & Raharto Condrobimo, A. (2018). The application of agnes algorithm to optimize knowledge base for tourism chatbot. *2018 International Conference on Information Management and Technology (ICIMTech)*, 65–68. https://doi.org/10.1109/ICIMTech.2018.8528174

Díaz Ramiro, E., Rubio Valdehita, S., Martín García, J. & Luceño Moreno, L. (2010). Psychometric study of nasa-tlx mental workload index in a sample of spanish workers. *Revista de Psicología del Trabajo y de las Organizaciones*, *26*, 191–199. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1576-59622010000300003&nrm=iso

Díaz-Oreiro, I., López, G., Quesada, L. & Guerrero, L. A. (2021). Ux evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: A systematic literature review. *Advances in Human-Computer Interaction*, *2021*, 5518722. https://doi.org/10.1155/2021/5518722

Díaz-Oreiro, López, Quesada & Guerrero. (2019). Standardized questionnaires for user experience evaluation: A systematic liter-

ature review. *Proceedings*, *31*(1), 14. https://doi.org/10.3390/proceedings2019031014

Ding, H., Ranade, N. & Cata, A. (2019). Boundary of content ecology: Chatbots, user experience, heuristics, and pedagogy. *Proceedings of the 37th ACM International Conference on the Design of Communication*. https://doi.org/10.1145/3328020.3353931

D-LABS. (2019). Medium-fidelity-prototyping [Accessed: October 2019 https://www.d-labs.com/en/services-and-methods/medium-fidelity-prototyping.html].

Dong, T., Churchill, E. F. & Nichols, J. (2016). Understanding the challenges of designing and developing multi-device experiences. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 62–72.

Duarte, E. F. & Baranauskas, M. C. C. (2016). Revisiting the three hci waves: A preliminary discussion on philosophy of science and research paradigms. *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*. https://doi.org/10.1145/3033701.3033740

El Morr, C. & Eftychiou, L. (2017). Evaluation frameworks for health virtual communities. In L. Menvielle, A.-F. Audrain-Pontevia & W. Menvielle (Eds.), *The digitization of healthcare: New challenges and opportunities* (pp. 99–118). Palgrave Macmillan UK.

El-Moussa, F. (2018). Internet of things: A survey of technologies and security risks in smart home and city environments. *IET Conference Proceedings*, 30–37. https://digital-library.theiet.org/content/conferences/10.1049/cp.2018.0030

Ensina, L. A., Lee, H. D., Takaki, W. S. R., Maciejewski, N. A. R., Spolaôr, N. & Wu, F. C. (2019). Heuristics-based responsiveness evaluation of a telemedicine computational web system. *IEEE Latin America Transactions*, *17*(03), 444–452.

Febiyani, A., Febriani, A. & Ma'Sum, J. (2021). Calculation of mental load from e-learning student with nasa tlx and sofi method. *Jurnal Sistem dan Manajemen Industri*, *5*(1), 35–42. https://doi.org/10.30656/jsmi.v5i1.2789

Følstad, A. (2017). Users' design feedback in usability evaluation: A literature review. *Human-centric Computing and Information Sciences*, *7*(1), 19. https://doi.org/10.1186/s13673-017-0100-y

Fonte, F. A. M., Nistal, M. L., Rial, J. C. B. & Rodríguez, M. C. (2016). Nlast: A natural language assistant for students. *2016 IEEE*

Global Engineering Education Conference (EDUCON), 709–713. https://doi.org/10.1109/EDUCON.2016.7474628

Fragidis, G., Ignatiadis, I. & Wills, C. (2010). Value co-creation and customer-driven innovation in social networking systems. In J.-H. Morin, J. Ralyté & M. Snene (Eds.), *Exploring services science* (pp. 254–258). Springer Berlin Heidelberg.

Gaffney, G. (2018). Why consistency is critical [[Online; accessed Oct-2018] https://www.sitepoint.com/why-consistency-is-critical/].

Geffen, S. (2016). The ipod may be dead, but those iconic ads still shape the way we see music. https://www.mtv.com/news/2879585/ipod-ads-in-music-culture/

Geisen, E. & Romano Bergstrom, J. (2017a). Chapter 1 - usability and usability testing. In E. Geisen & J. Romano Bergstrom (Eds.), *Usability testing for survey research* (pp. 1–19). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-803656-3.00001-4

Geisen, E. & Romano Bergstrom, J. (2017b). Chapter 3 - adding usability testing to the survey process. In E. Geisen & J. Romano Bergstrom (Eds.), *Usability testing for survey research* (pp. 51–78). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-803656-3.00003-8

Geisen, E. & Romano Bergstrom, J. (2017c). Chapter 4 - planning for usability testing. In E. Geisen & J. Romano Bergstrom (Eds.), *Usability testing for survey research* (pp. 79–109). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-803656-3.00004-X

Gossman, R. R. F. (1998). *La gran corriente ornamental del siglo xx: Una revisión de la arquitectura neocolonial en la ciudad de méxico*. Universidad Iberoamericana.

Grice, R. A., Bennett, A. G., Fernheimer, J. W., Geisler, C., Krull, R., Lutzky, R. A., Rolph, M. G., Search, P. & Zappen, J. P. (2013). Heuristics for broader assessment of effectiveness and usability in technology-mediated technical communication. *Technical Communication*, *60*(1), 3–27.

Grosjean, J. C. (2018). Design d'interface et critère ergonomique 9: Cohérence [[Online; accessed Oct-2018] http://www.qualitystreet.fr/2011/01/23/design-dinterface-et-critere-ergonomique-9-coherence/].

Grudin, J. (1988). Why cscw applications fail: Problems in the design and evaluationof organizational interfaces. *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work*, 85–93. https://doi.org/10.1145/62266.62273

Grudin, J. (2018). From tool to partner: The evolution of human-computer interaction. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–3. https://doi.org/10.1145/3170427.3170663

Guerino, G. C. & Valentim, N. M. C. (2020). Usability and user experience evaluation of conversational systems: A systematic mapping study. *Proceedings of the 34th Brazilian Symposium on Software Engineering*, 427–436. https://doi.org/10.1145/3422392.3422421

Hao, K. (2020). The pandemic is emptying call centers. ai chatbots are swooping in. https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic/

Harper, R. H. R. (2019). The role of hci in the age of ai. *International Journal of Human–Computer Interaction*, *35*(15), 1331–1344. https://doi.org/10.1080/10447318.2019.1631527

Harrison, S., Tatar, D. & Sengers, P. (2007). The three paradigms of hci. *Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA*, 1–18.

Hart, S. G. & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology* (pp. 139–183). Elsevier.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction*, *19*(4), 319–349. https://doi.org/10.1207/s15327051hci1904_2

Hassenzahl, M. (2007). The hedonic/pragmatic model of user experience. *Towards a UX manifesto*, *10*.

Hassenzahl, M., Burmester, M. & Koller, F. (2003). Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch & computer 2003* (pp. 187–196). Springer.

Hassenzahl, M. & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, *25*(3), 235–260. https://doi.org/10.1080/07370024.2010.500139

Hassenzahl, M., Platz, A., Burmester, M. & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's

appeal. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 201–208.

Hernández-Sampieri, R. & Torres, C. P. M. (2018). *Metodología de la investigación* (Vol. 4). McGraw-Hill Interamericana Ciudad de México.

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V. & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *Proceedings of the 31st european conference on cognitive ergonomics* (pp. 207–214). Association for Computing Machinery. https://doi.org/10.1145/3335082.3335094

Hu, J., Le, D., Funk, M., Wang, F. & Rauterberg, M. (2013). Attractiveness of an interactive public art installation. In N. Streitz & C. Stephanidis (Eds.), *Distributed, ambient, and pervasive interactions* (pp. 430–438). Springer Berlin Heidelberg.

Hummel, J. & Lechner, U. (2002). Social profiles of virtual communities. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, 2245–2254.

Iniguez-Carrillo, A. L., Gaytan-Lugo, L. S., Garcia-Ruiz, M. A. & Maciel-Arellano, R. (2021). Usability questionnaires to evaluate voice user interfaces. *IEEE Latin America Transactions*, *19*(9), 1468–1477. https://latamt.ieeer9.org/index.php/transactions/article/view/4771

Intelligence, I. (2021). Chatbot market in 2021: Stats, trends, and companies in the growing ai chatbot industry. https://www.businessinsider.com/chatbot-market-stats-trends

Iriberri, A. & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Comput. Surv.*, *41*(2), 11:1–11:29.

Isleifsdottir, J. & Larusdottir, M. (2008). Measuring the user experience of a task oriented software. *Proceedings of the international workshop on meaningful measures: valid useful user experience measurement*, *8*, 97–101.

ISO. (2010). *Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems* (tech. rep.). International Organization for Standardization. Geneva, CH.

Isomursu, P., Virkkula, M., Niemelä, K., Juntunen, J. & Kumpuoja, J. (2020). Modified attrakdiff in ux evaluation of a mobile prototype. *Proceedings of the International Conference on Advanced Visual Interfaces*. https://doi.org/10.1145/3399715.3399930

Jacobsen, L. F., Tudoran, A. A. & Lähteenmäki, L. (2017). Consumers' motivation to interact in virtual food communities - the importance of self-presentation and learning. *Food Quality and Preference*, *62*, 8–16.

Jiménez, A. & Thamar, B. (2019). Determinación de carga mental en docentes universitarios del norte de méxico. *Instituto de Ingeniería y Tecnología*. http://cathi.uacj.mx/bitstream/handle/20.500.11961/8151/Memorias%20Academia%20Journals%20Oaxaca%202019%20-%20Carga%20mental.pdf

Johnson, J. (2008). Introduction. In J. Johnson (Ed.), *Gui bloopers 2.0* (pp. 1–6). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-012370643-0.50012-3

Johnson, J. (2014a). Chapter 1 - our perception is biased. In J. Johnson (Ed.), *Designing with the mind in mind (second edition)* (Second Edition, pp. 1–12). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-407914-4.00001-4

Johnson, J. (2014b). Chapter 10 - learning from experience and performing learned actions are easy; novel actions, problem solving, and calculation are hard. In J. Johnson (Ed.), *Designing with the mind in mind (second edition)* (Second Edition, pp. 131–148). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-407914-4.00010-5

Jordan, P. (1998). *An introduction to usability*. CRC Press. https://books.google.com.mx/books?id=aw%5C_2DwAAQBAJ

Joyce, A. (2019). Formative vs. summative evaluations. https://www.nngroup.com/articles/formative-vs-summative-evaluations/

Julien, J. (2012). The importance of knowing user intent. https://www.uxmatters.com/mt/archives/2012/10/the-importance-of-knowing-user-intent.php

Kairy, D., Mostafavi, M. A., Blanchette-Dallaire, C., Belanger, E., Corbeil, A., Kandiah, M., Wu, T. Q. & Mazer, B. (2021). A mobile app to optimize social participation for individuals with physical disabilities: Content validation and usability testing. *International Journal of Environmental Research and Public Health*, *18*(4). https://doi.org/10.3390/ijerph18041753

Karagianni, K. (2018). Optimizing the ux honeycomb. https://uxdesign.cc/optimizing-the-ux-honeycomb-1d10cfb38097

Karani, A., Thanki, H. & Achuthan, S. (2021). Impact of university website usability on satisfaction: A structural equation model-

ling approach. *Management and Labour Studies*, *46*(2), 119–138. https://doi.org/10.1177/0258042X21989924

Karapanos, E., Zimmerman, J., Forlizzi, J. & Martens, J.-B. (2009). User experience over time: An initial framework. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 729–738.

Karapanos, E., Zimmerman, J., Forlizzi, J. & Martens, J.-B. (2010). Measuring the dynamics of remembered experience over time [Modelling user experience - An agenda for research and practice]. *Interacting with Computers*, *22*(5), 328–335. https://doi.org/https://doi.org/10.1016/j.intcom.2010.04.003

Kaye, J. '. (2007). Evaluating experience-focused hci, 1661–1664. https://doi.org/10.1145/1240866.1240877

Kazi, H., Chowdhry, B. & Memon, Z. (2012). Medchatbot: An umls based chatbot for medical students. *International Journal of Computer Applications*, *55*(17).

Keijzer-Broers, W. J. W. & de Reuver, M. (2016). Applying agile design sprint methods in action design research: Prototyping a health and wellbeing platform. In J. Parsons, T. Tuunanen, J. Venable, B. Donnellan, M. Helfert & J. Kenneally (Eds.), *Tackling society's grand challenges with design science* (pp. 68–80). Springer International Publishing.

Kim, J., Kim, J. & Moon, J. Y. (2013). Does age matter in mobile user experience? impact of age on relative importance of antecedents of mobile user experience. *Proceedings of the Pacific Asia Conference on Information Systems*. https://aisel.aisnet.org/pacis2013/189

Király, O., Potenza, M. N., Stein, D. J., King, D. L., Hodgins, D. C., Saunders, J. B., Griffiths, M. D., Gjoneska, B., Billieux, J., Brand, M., Abbott, M. W., Chamberlain, S. R., Corazza, O., Burkauskas, J., Sales, C. M., Montag, C., Lochner, C., Grünblatt, E., Wegmann, E., … Demetrovics, Z. (2020). Preventing problematic internet use during the covid-19 pandemic: Consensus guidance. *Comprehensive Psychiatry*, *100*, 152–180. https://doi.org/https://doi.org/10.1016/j.comppsych.2020.152180

Klaassen, R., op den Akker, R., Lavrysen, T. & van Wissen, S. (2013). User preferences for multi-device context-aware feedback in a digital coaching system. *Journal on Multimodal User Interfaces*, *7*(3), 247–267. https://doi.org/10.1007/s12193-013-0125-0

Klobas, J. E., McGill, T. & Wang, X. (2019). How perceived security risk affects intention to use smart home devices: A reasoned action explanation. *Computers & Security*, *87*, 101571. https://doi.org/https://doi.org/10.1016/j.cose.2019.101571

Kocaballi, A. B., Laranjo, L. & Coiera, E. (2019). Understanding and Measuring User Experience in Conversational Interfaces. *Interacting with Computers*, *31*(2), 192–207. https://doi.org/10.1093/iwc/iwz015

Koh, J., Kim, Y.-G., Butler, B. & Bock, G.-W. (2007). Encouraging participation in virtual communities. *Commun. ACM*, *50*(2), 68–73.

Kothari, C. R. (2004). *Research methodology: Methods and techniques*. New Age International.

Kreitz, G. & Niemela, F. (2010). Spotify – large scale, low latency, p2p music-on-demand streaming. *2010 IEEE Tenth International Conference on Peer-to-Peer Computing (P2P)*, 1–10. https://doi.org/10.1109/P2P.2010.5569963

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E. & Sinnelä, A. (2011). Ux curve: A method for evaluating long-term user experience [Feminism and HCI: New Perspectives]. *Interacting with Computers*, *23*(5), 473–483. https://doi.org/https://doi.org/10.1016/j.intcom.2011.06.005

Kukka, H., Pakanen, M., Badri, M. & Ojala, T. (2017). Immersive street-level social media in the 3d virtual city: Anticipated user experience and conceptual development. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2422–2435.

Kuksenok, K. & Martyniv, A. (2019). Evaluation and improvement of chatbot text classification data quality using plausible negative examples.

Kumar, B. A. & Chand, S. (2018). Mobile app to support teaching in distance mode at fiji national university: Design and evaluation. *International Journal of Virtual and Personal Learning Environments (IJVPLE)*, *8*(1), 25–37.

Kumar, B. A., Goundar, M. S. & Chand, S. S. (2020). A framework for heuristic evaluation of mobile learning applications. *Education and Information Technologies*. https://doi.org/10.1007/s10639-020-10112-8

Lallemand, C., Gronier, G. & Koenig, V. (2015). User experience: A concept without consensus? exploring practitioners' perspect-

ives through an international survey. *Computers in Human Behavior*, *43*, 35–48. https://doi.org/https://doi.org/10.1016/j.chb.2014.10.048

Lallemand, C. & Koenig, V. (2020). Measuring the contextual dimension of user experience: Development of the user experience context scale (uxcs). *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. https://doi.org/10.1145/3419249.3420156

Lamprecht, J., Siemon, D. & Robra-Bissantz, S. (2016). Cooperation isn't just about doing the same thing – using personality for a cooperation-recommender-system in online social networks. In T. Yuizono, H. Ogata, U. Hoppe & J. Vassileva (Eds.), *Collaboration and technology* (pp. 131–138). Springer.

Law, E., Roto, V., Vermeeren, A. P., Kort, J. & Hassenzahl, M. (2008). Towards a shared definition of user experience. *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, 2395–2398. https://doi.org/10.1145/1358628.1358693

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–728. https://doi.org/10.1145/1518701.1518813

Lazar, J., Feng, J. H. & Hochheiser, H. (2017a). Chapter 3 - experimental design. In J. Lazar, J. H. Feng & H. Hochheiser (Eds.), *Research methods in human computer interaction (second edition)* (Second Edition, pp. 45–69). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-805390-4.00003-0

Lazar, J., Feng, J. H. & Hochheiser, H. (2017b). Chapter 4 - statistical analysis. In J. Lazar, J. H. Feng & H. Hochheiser (Eds.), *Research methods in human computer interaction (second edition)* (Second Edition, pp. 71–104). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-805390-4.00004-2

Lee, F. S., Vogel, D. & Limayem, M. (2003). Virtual community informatics: A review and research agenda. *JITTA: Journal of Information Technology Theory and Application*, *5*(1), 47.

Levin, M. (2014). *Designing multi-device experiences: An ecosystem approach to creating user experiences across devices*. O'Reilly.

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., Resnick, P. & Kraut, R. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, *10*(4), 00–00. https://doi.org/10.1111/j.1083-6101.2005.tb00273.x

Luctkar-Flude, M., Tyerman, J., Tregunno, D., Bell, C., Lalonde, M., McParland, T., Peachey, L., Verkuyl, M. & Mastrilli, P. (2021). Designing a virtual simulation game as presimulation preparation for a respiratory distress simulation for senior nursing students: Usability, feasibility, and perceived impact on learning. *Clinical Simulation in Nursing*, *52*, 35–42. https://doi.org/https://doi.org/10.1016/j.ecns.2020.11.009

Luo, C. J., Wong, V. Y. L. & Gonda, D. E. (2020). Code free chatbot development: An easy way to jumpstart your chatbot! *Proceedings of the Seventh ACM Conference on Learning @ Scale*, 233–235. https://doi.org/10.1145/3386527.3405932

Lynch, S. (2021). Peter norvig: Today's most pressing questions in ai are human-centered. https://hai.stanford.edu/news/peter-norvig-todays-most-pressing-questions-ai-are-human-centered

Magaña, C. (2019). *El art déco en ciudad de méxico: Retrospectiva de un movimiento arqutectónico*. Siglo XXI Editores México.

Magin, D. P., Maier, A. & Hess, S. (2015). Measuring negative user experience. In A. Marcus (Ed.), *Design, user experience, and usability: Users and interactions* (pp. 95–106). Springer.

Mai, H. T. X. & Olsen, S. O. (2015). Consumer participation in virtual communities: The role of personal values and personality. *Journal of Marketing Communications*, *21*(2), 144–164.

Mäkinen, L. A. (2016). Surveillance on/off: Examining home surveillance systems from the user's perspective. *Surveillance & Society*, *14*(1), 59–77.

Marcus, A. (1995). Principles of effective visual communication for graphical user interface design. In R. M. Baecker, J. Grudin, W. A. Buxton & S. Greenberg (Eds.), *Readings in human–computer interaction* (pp. 425–441). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-08-051574-8.50044-3

Margetis, G., Zabulis, X., Koutlemanis, P., Antona, M. & Stephanidis, C. (2013). Augmented interaction with physical books in an

ambient intelligence learning environment. *Multimedia Tools and Applications*, *67*(2), 473–495.

Marquis, K. H., Nichols, E. & Tedesco, H. (1998). *Human-computer interface usability in a survey organization: Getting started at the census bureau*. US Bureau of the Census.

Marshall, J. (2019). Ux: Do less, but with feeling. https://uxdesign.cc/ux-do-less-but-with-feeling-e58e2c8b3c90

Märtin, C., Bissinger, B. C. & Asta, P. (2021). Optimizing the digital customer journey—improving user experience by exploiting emotions, personas and situations for individualized user interface adaptations. *Journal of Consumer Behaviour*. https://doi.org/10.1002/cb.1964

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396. https://doi.org/10.1037/h0054346

Mathis, F., Vaniea, K. & Khamis, M. (2021). Prototyping usable privacy and security systems: Insights from experts. *International Journal of Human–Computer Interaction*, *0*(0), 1–23. https://doi.org/10.1080/10447318.2021.1949134

Maulsby, D., Greenberg, S. & Mander, R. (1993). Prototyping an intelligent agent through wizard of oz. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 277–284. https://doi.org/10.1145/169059.169215

McCormick, T. J. (2010). *A success-oriented framework to enable co-created e-services*. The George Washington University.

Melchior, J., Vanderdonckt, J. & Van Roy, P. (2011). A model-based approach for distributed user interfaces. *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 11–20. https://doi.org/10.1145/1996461.1996488

Mendoza, S., Hernández-León, M., Sánchez-Adame, L. M., Rodríguez, J., Decouchant, D. & Meneses-Viveros, A. (2020). Supporting student-teacher interaction through a chatbot. In P. Zaphiris & A. Ioannou (Eds.), *Learning and collaboration technologies. human and technology ecosystems* (pp. 93–107). Springer International Publishing.

Merz, B., Tuch, A. N. & Opwis, K. (2016). Perceived user experience of animated transitions in mobile user interfaces. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3152–3158. https://doi.org/10.1145/2851581.2892489

Meskens, J., Luyten, K. & Coninx, K. (2010). Jelly: A multi-device design environment for managing consistency across devices. *Proceedings of the International Conference on Advanced Visual Interfaces*, 289–296. https://doi.org/10.1145/1842993.1843044

Microsoft. (2018). Windows ribbon framework [[Online; accessed Oct-2018] https://docs.microsoft.com/en-us/windows/desktop/windowsribbon/-uiplat-windowsribbon-entry].

Miner, A. S., Laranjo, L. & Kocaballi, A. B. (2020). Chatbots in the fight against the covid-19 pandemic. *npj Digital Medicine*, *3*(1), 65. https://doi.org/10.1038/s41746-020-0280-0

Miyake, S. (2020). [mental workload assessment of health care staff by nasa-tlx]. *Journal of UOEH*, *42*(1), 63–75. https://doi.org/10.7888/juoeh.42.63

Molnár, G. & Szüts, Z. (2018). The role of chatbots in formal education. *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, 000197–000202. https://doi.org/10.1109/SISY.2018.8524609

Momenipour, A., Rojas-Murillo, S., Murphy, B., Pennathur, P. & Pennathur, A. (2021). Usability of state public health department websites for communication during a pandemic: A heuristic evaluation. *International Journal of Industrial Ergonomics*, *86*, 103216. https://doi.org/https://doi.org/10.1016/j.ergon.2021.103216

Morville, P. (2005). Experience design unplugged. *ACM SIGGRAPH 2005 Web Program*.

Morville, P. (2016). User experience design. https://semanticstudios.com/user_experience_design/

Moser, C., Chisik, Y. & Tscheligi, M. (2014). Around the world in 8 workshops: Investigating anticipated player experiences of children. *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, 207–216.

Müller, F., Hummers, E., Schulze, J. & Noack, E. M. (2021). [usability of an app to overcome language barriers in paramedic care]. *Notfall & Rettungsmedizin*, 1–7. https://doi.org/10.1007/s10049-021-00913-w

Nichols, J. (2006). Automatically generating high-quality user interfaces for appliances.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, 373–380. https://doi.org/10.1145/142750.142834

Nielsen, J. (2012). Usability 101: Introduction to usability. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Nielsen, J. (2020). 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 206–213. https://doi.org/10.1145/169059.169166

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249–256. https://doi.org/10.1145/97243.97281

Nik, I. (2021). Major behavioral theories, explained. https://uxdesign.cc/major-behavioral-theories-explained-7cad53369 4a2

Nikolov, A. (2017). Design principle: Consistency [[Online; accessed Oct-2018] https://uxdesign.cc/design-principle-consistency-6bocf7e7339f].

Norman, D. (2002). The design of everyday things.

Norman, D. (2004). Introduction to this special section on beauty, goodness, and usability. *Human–Computer Interaction*, *19*(4), 311–318. https://doi.org/10.1207/s15327051hci1904_1

Norman, D. A. (1983). Design rules based on analyses of human error. *Commun. ACM*, *26*(4), 254–258. https://doi.org/10.1145/2163.358092

Nov, O. & Ye, C. (2010). Why do people tag?: Motivations for photo tagging. *Commun. ACM*, *53*(7), 128–131.

O'Leary, K., Dong, T., Haines, J. K., Gilbert, M., Churchill, E. F. & Nichols, J. (2017). The moving context kit: Designing for context shifts in multi-device experiences. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 309–320. https://doi.org/10.1145/3064663.3064768

Oulasvirta, A. & Hornbæk, K. (2021). Counterfactual thinking: What theories do in design. *International Journal of Human–Computer Interaction*, *0*(0), 1–15. https://doi.org/10.1080/10447318.2021.1925436

Parlangeli, O., Marchigiani, E. & Bagnara, S. (1999). Multimedia systems in distance education: Effects of usability on learning.

Interacting with Computers, *12*(1), 37–49. https://doi.org/https://doi.org/10.1016/S0953-5438(98)00054-X

Peffers, K., Tuunanen, T., Rothenberger, M. & Chatterjee, S. (2007). A design science research methodology for information systems research. *J. Manage. Inf. Syst.*, *24*(3), 45–77.

Petrie, H. & Bevan, N. (2009). The evaluation of accessibility, usability, and user experience. *The universal access handbook*, *1*, 1–16.

Poole, L., Farber, D., Douglas, M., Bunnell, D., Young, J. S., Fluegelman, A., Willliam, A. T. & McCandless, J. (1984). Macworld: The macintosh magazine, premier issue.

Preece, J. (2000). *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.

Preece, J. (2001). Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology*, *20*(5), 347–356.

Preece, J., Abras, C. & Maloney-Krichmar, D. (2004). Designing and evaluating online communities: Research speaks to emerging practice. *Int. J. Web Based Communities*, *1*(1), 2–18.

Pyla, P. S., Tungare, M. & Pérez-Quinones, M. (2006). Multiple user interfaces: Why consistency is not everything, and seamless task migration is key. *Proceedings of the CHI 2006 workshop on the many faces of consistency in cross-platform design*.

Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J. & Chu, W. (2017). AliMe chat: A sequence to sequence and rerank based chatbot engine. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 498–503. https://doi.org/10.18653/v1/P17-2079

Quiñones, D., Rusu, C., Roncagliolo, S., Rusu, V. & Collazos, C. A. (2016). Developing usability heuristics: A formal or informal process? *IEEE Latin America Transactions*, *14*(7), 3400–3409.

Quiñones, D. & Rusu, C. (2017). How to develop usability heuristics: A systematic literature review. *Computer Standards & Interfaces*, *53*, 89–122. https://doi.org/https://doi.org/10.1016/j.csi.2017.03.009

Rafael, M. S., María, T. B. L., Antonio, F. U. & Hanns, D. L. F. M. (2019). Support to the learning of the chilean tax system using artificial intelligence through a chatbot. *2019 38th International Conference of the Chilean Computer Science Society (SCCC)*, 1–8.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151. https://doi.org/10.1145/3375627.3375820

Ranoliya, B. R., Raghuwanshi, N. & Singh, S. (2017). Chatbot for university related faqs. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1525–1530. https://doi.org/10.1109/ICACCI.2017.8126057

Ravi, R. (2018). Intelligent chatbot for easy web-analytics insights. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2193–2195. https://doi.org/10.1109/ICACCI.2018.8554577

Reeves, L. M., Lai, J., Larson, J. A., Oviatt, S., Balaji, T. S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.-C., McTear, M., Raman, T., Stanney, K. M., Su, H. & Wang, Q. Y. (2004). Guidelines for multimodal user interface design. *Commun. ACM*, *47*(1), 57–59. https://doi.org/10.1145/962081.962106

Ren, R., Castro, J. W., Acuña, S. T. & de Lara, J. (2019). Evaluation techniques for chatbot usability: A systematic mapping study. *International Journal of Software Engineering and Knowledge Engineering*, *29*(11n12), 1673–1702. https://doi.org/10.1142/S0218194019400163

Ribeiro, I. M. & Providência, B. (2021). Quality perception with attrakdiff method: A study in higher education during the covid-19 period. In N. Martins, D. Brandão & F. Moreira da Silva (Eds.), *Perspectives on design and digital communication ii: Research, innovations and best practices* (pp. 217–231). Springer International Publishing. https://doi.org/10.1007/978-3-030-75867-7_14

Rogers, Y. (2012). Hci theory: Classical, modern, and contemporary. *Synthesis lectures on human-centered informatics*, *5*(2), 1–129.

Roto, V., Law, E. L.-C., Vermeeren, A. & Hoonhout, J. (2011). 10373 Abstracts Collection – Demarcating User eXperience. In J. Hoonhout, E. L.-C. Law, V. Roto & A. Vermeeren (Eds.), *Proceedings of dagstuhl seminar on demarcating user experience*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany. http://drops.dagstuhl.de/opus/volltexte/2011/2949

Rowland, C., Goodman, E., Charlier, M., Light, A. & Lui, A. (2015). *Designing connected products: Ux for the consumer internet of things*. O'Reilly.

Rubin, J. & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design and conduct effective tests*. John Wiley & Sons.

Sánchez-Adame, L. M. (2016). *Consistency heuristics for designing meta-uis* (Master's thesis). CINVESTAV-IPN.

Sandars, J. (2010). The importance of usability testing to allow e-learning to reach its potential for medical education. *Education for Primary Care*, *21*(1), 6–8. https://doi.org/10.1080/14739879.2010.11493869

Sari, E. & Tedjasaputra, A. (2017). Designing valuable products with design sprint. In R. Bernhaupt, G. Dalvi, A. Joshi, D. K. Balkrishan, J. O'Neill & M. Winckler (Eds.), *Human-computer interaction – interact 2017* (pp. 391–394). Springer International Publishing.

Sato, G. Y., de Azevedo, H. J. S. & Barthès, J.-P. A. (2012). Agent and multi-agent applications to support distributed communities of practice: A short review. *Autonomous Agents and Multi-Agent Systems*, *25*(1), 87–129.

Sauro, J. (2010a). Are the terms formative and summative helpful or harmful? https://measuringu.com/formative-summative/

Sauro, J. (2010b). Do you need a random sample for your usability test? https://measuringu.com/random-sample/

Sauro, J. (2016). Measuring the quality of the website user experience.

Sauro, J. & Lewis, J. R. (2012a). Chapter 6 - what sample sizes do we need?: Part 1: Summative studies. In J. Sauro & J. R. Lewis (Eds.), *Quantifying the user experience* (pp. 105–142). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-384968-7.00006-0

Sauro, J. & Lewis, J. R. (2012b). Chapter 7 - what sample sizes do we need?: Part 2: Formative studies. In J. Sauro & J. R. Lewis (Eds.), *Quantifying the user experience* (pp. 143–184). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-384968-7.00007-2

Schmettow, M., Schnittker, R. & Schraagen, J. M. (2017). An extended protocol for usability validation of medical devices. *J. of Biomedical Informatics*, *69*(100), 99–114. https://doi.org/10.1016/j.jbi.2017.03.010

Schrepp, M., Hinderks, A. & Thomaschewski, J. (2017). Construction of a benchmark for the user experience questionnaire (ueq). *Int. J. Interact. Multim. Artif. Intell.*, *4*(4), 40–44.

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L. & Callison-Burch, C. (2019). ChatEval: A tool for chatbot evaluation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 60–65. https://doi.org/10.18653/v1/N19-4011

Segura, C., Palau, À., Luque, J., Costa-Jussà, M. R. & Banchs, R. E. (2019). Chatbol, a chatbot for the spanish "la liga". In L. F. D'Haro, R. E. Banchs & H. Li (Eds.), *9th international workshop on spoken dialogue system technology* (pp. 319–330). Springer Singapore.

Shaw, A. (2012). Using chatbots to teach socially intelligent computing principles in introductory computer science courses. *2012 Ninth International Conference on Information Technology - New Generations*, 850–851. https://doi.org/10.1109/ITNG.2012.70

Shawar, B. A. & Atwell, E. (2007). Chatbots: Are they really useful? *Ldv forum*, *22*(1), 29–49.

Shehan, E. & Edwards, W. K. (2007). Home networking and hci: What hath god wrought? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 547–556. https://doi.org/10.1145/1240624.1240712

Shneiderman, B. & Plaisant, C. (2010). *Designing the user interface: Strategies for effective human-computer interaction*. Pearson Education.

Smutny, P. & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, *151*, 103862. https://doi.org/https://doi.org/10.1016/j.compedu.2020.103862

Soure, E. J., Kuang, E., Fan, M. & Zhao, J. (2021). Coux: Collaborative visual analysis of think-aloud usability test videos for digital interfaces. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2021.3114822

Southall, H., Marmion, M. & Davies, A. (2019). Adapting jake knapp's design sprint approach for ar/vr applications in digital heritage. In M. C. tom Dieck & T. Jung (Eds.), *Augmented reality and virtual reality: The power of ar and vr for business* (pp. 59–70). Springer International Publishing. https://doi.org/10.1007/978-3-030-06246-0_5

Spotify. (2018). About [[Online; accessed Oct-2018] https://www.spotify.com/about-us/].

Stone, D., Jarrett, C., Woodroffe, M. & Minocha, S. (2005). *User interface design and evaluation*. Morgan Kaufman.

Strohmann, T., Höper, L. & Robra-Bissantz, S. (2019). Design guidelines for creating a convincing user experience with virtual in-vehicle assistants. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Stroll, A. & Martinich, A. (2021). Epistemology. https://www.britannica.com/topic/epistemology

Svaigen, A. R. & Martimiano, L. A. F. (2018). Netanimations mobile app: Improvement of accessibility and usability to computer network learning animations. *IEEE Latin America Transactions*, *16*(1), 272–278.

Takahashi, L. & Nebe, K. (2019). Observed differences between lab and online tests using the attrakdiff semantic differential scale. *Journal of Usability Studies*, *14*(2), 65–75. http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=135193687&site=ehost-live

Talin. (2019). Why google+ failed [Accessed: October 2019 https://onezero.medium.com/why-google-failed-4b9db05b973b].

Tella, A. & Babatunde, B. J. (2017). Determinants of continuance intention of facebook usage among library and information science female undergraduates in selected nigerian universities. *International Journal of E-Adoption (IJEA)*, *9*(2), 59–76.

Tractinsky, N. (2004). A few notes on the study of beauty in hci. *Human–Computer Interaction*, *19*(4), 351–357. https://doi.org/10.1207/s15327051hci1904_3

Tulaskar, R. & Turunen, M. (2021). What students want? experiences, challenges, and engagement during emergency remote learning amidst covid-19 crisis. *Education and Information Technologies*. https://doi.org/10.1007/s10639-021-10747-1

Urquhart, L. & Rodden, T. (2017). New directions in information technology law: Learning from human–computer interaction. *International Review of Law, Computers & Technology*, *31*(2), 150–169. https://doi.org/10.1080/13600869.2017.1298501

usability.gov. (2014). User experience basics. https://www.usability.gov/what-and-why/user-experience.html

Väätäjä, H., Koponen, T. & Roto, V. (2009). Developing practical tools for user experience evaluation: A case from mobile

news journalism. *European Conference on Cognitive Ergonomics: Designing beyond the Product—Understanding Activity and User Experience in Ubiquitous Environments*, 1–8.

Valtolina, S., Barricelli, B. R. & Gaetano, S. D. (2020). Communicability of traditional interfaces vs chatbots in healthcare and smart home domains. *Behaviour & Information Technology*, *39*(1), 108–132. https://doi.org/10.1080/0144929X.2019.1637025

Vanderdonckt, J. (2010). Distributed user interfaces: How to distribute user interface elements across users, platforms, and environments, 3–14.

Vergadia, P. (2020). How can chatbots help during global pandemic (covid-19)? https://medium.com/google-cloud/how-can-chatbots-help-during-global-pandemic-covid-19-9c1a4428d8c2

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J. & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–530.

Virzi, R. A. (1989). What can you learn from a low-fidelity prototype? *Proceedings of the Human Factors Society Annual Meeting*, *33*(4), 224–228. https://doi.org/10.1177/154193128903300405

Walker, M., Takayama, L. & Landay, J. A. (2002). High-fidelity or low-fidelity, paper or computer? choosing attributes when testing web prototypes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(5), 661–665. https://doi.org/10.1177/154193120204600513

Walsh, T., Varsaluoma, J., Kujala, S., Nurkka, P., Petrie, H. & Power, C. (2014). Axe ux: Exploring long-term user experience with iscale and attrakdiff. *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*, 32–39.

Wang, X., Li, Y., Cai, Z. & Liu, H. (2021). Beauty matters: Reducing bounce rate by aesthetics of experience product portal page. *Industrial Management & Data Systems*, *121*(8), 1848–1870. https://doi.org/10.1108/IMDS-08-2020-0484

Wang, Y. & Li, Y. (2016). Proactive engagement of opinion leaders and organization advocates on social networking sites. *International Journal of Strategic Communication*, *10*(2), 115–132.

Weinschenk, S. (2010). The psychologist's view of ux design. https://uxmag.com/articles/the-psychologists-view-of-ux-design

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wenning, C. J. (2009). Scientific epistemology: How scientists know what they know. *Journal of Physics Teacher Education Online*, *5*(2), 3–15. http://samuelkrueger.com/PTE/publications/scientific_epistemology.pdf

Wesolko, D. (2016). Peter morville's user experience honeycomb. https://medium.com/%5C@danewesolko/peter-morvilles-user-experience-honeycomb-904c383b6886

Wiederhold, B. K. (2020). Social media use during social distancing [PMID: 32255658]. *Cyberpsychology, Behavior, and Social Networking*, *23*(5), 275–276. https://doi.org/10.1089/cyber.2020.29181.bkw

Wiener, E. L. (1989). Human factors of advanced technology (glass cockpit) transport aircraft.

Wikipedia. (2021). Peter morville. https://en.wikipedia.org/wiki/Peter_Morville

Winckler, M., Bernhaupt, R. & Bach, C. (2016). Identification of ux dimensions for incident reporting systems with mobile applications in urban contexts: A longitudinal study. *Cognition, Technology & Work*, *18*(4), 673–694. https://doi.org/10.1007/s10111-016-0383-1

Wong, E. (2018). Principle of consistency and standards in user interface design [[Online; accessed Oct-2018] https://www.interaction-design.org/literature/article/principle-of-consistency-and-standards-in-user-interface-design].

Wong, R. Y. (2021). Using design fiction memos to analyze ux professionals' values work practices: A case study bridging ethnographic and design futuring methods. *Proceedings of the 2021 chi conference on human factors in computing systems*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445709

Woodrow, W. W. (2016). Designing for interusability: Methodological recommendations for the systems engineer gleaned through

an exploration of the connected fitness technologies space. *INSIGHT*, *19*(3), 75–77. https://doi.org/10.1002/inst.12115

Wurhofer, D., Krischkowsky, A., Obrist, M., Karapanos, E., Niforatos, E. & Tscheligi, M. (2015). Everyday commuting: Prediction, actual experience and recall of anger and frustration in the car. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 233–240.

Yin, J., Goh, T.-T., Yang, B. & Xiaobin, Y. (2020). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research*. https://doi.org/10.1177/0735633120952067

Yocco, V. (2016). *Design for the mind: Seven psychological principles of persuasive design*. Manning Publications.

Yogasara, T., Popovic, V., Kraal, B. J. & Chamorro-Koc, M. (2011). General characteristics of anticipated user experience (aux) with interactive products. *Proceedings of IASDR2011: the 4th World Conference on Design Research: Diversity and Unity*, 1–11.

Zeng, E., Mare, S. & Roesner, F. (2017). End user security and privacy concerns with smart homes. *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, 65–80. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/zeng

Zhang, E., Culbertson, G., Shen, S. & Jung, M. (2018). Utilizing narrative grounding to design storytelling gamesfor creative foreign language production. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 197:1–197:11.

Zhang, J., Johnson, T. R., Patel, V. L., Paige, D. L. & Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics*, *36*(1), 23–30.

Zhao, Z., Chen, J., Zhou, S., He, X., Cao, X., Zhang, F. & Wu, W. (2021). Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation.

Zheng, S., Apthorpe, N., Chetty, M. & Feamster, N. (2018). User perceptions of smart home iot privacy. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 200:1–200:20. https://doi.org/10.1145/3274469

Zhou, T. (2011). Understanding online community user participation: A social influence perspective. *Internet Research*, *21*(1), 67–81.