TÉCNICO LISBOA

TUx: Testing UX between Web Frameworks

Luís Carlos Alves Henriques

Thesis to obtain the Master of Science Degree in

Engenharia Informática e de Computadores

Supervisor: Prof. Manuel João Caneira Monteiro da Fonseca

Examination Committee

Chairperson: Prof. José Luís Brinquete Borbinha
Supervisor: Prof. Manuel João Caneira Monteiro da Fonseca
Members of the Committee: Prof. Daniel Jorge Viegas Gonçalves

May 2015

Acknowledgement

In first place I would like to thank you to my supervisor Professor Manuel da Fonseca for all de help and support on my work for this dissertation. Also I would like to thank you to my external advisor Sabrina Mach, James Page, Webnographer and all the staff that helped me and allow me developing this work with them. Thank you to all my friends for the motivation and support to take this work until the end.
Finally thank you to my family for all the support and help, because without them I would never be able to do all the way until here the way I did.

Thank you

Resumo

Recentemente têm vindo a ser discutidos os potenciais problemas de usabilidade que o Flat Design pode causar nas interfaces de utilizador que não acontecem caso usemos Skeuomorphism. Neste trabalho foi desenvolvido um estudo comparativo, de modo perceber se o Flat Design influencia ou não o desempenho dos utilizadores durante o uso de uma aplicação em comparação com o design utilizado até agora o Skeuomorphism. Para compreender melhor este tema realizámos uma investigação dos trabalhos desenvolvidos acerca da influência da estética e design na usabilidade de uma interface. Dado irmos realizar este trabalho com a Webnographer e irmos utilizar a sua ferramenta de avaliação remota, adicionalmente realizámos uma investigação para demonstrar que este era um bom método de avaliação de usabilidade a aplicar. No nosso estudo começámos por aplicar os dois estilos a uma interface: Flat Design e "Skeuomorphism". De seguida realizámos um teste de usabilidade a cada um deles para perceber se o Flat Design afecta ou não o desempenho do utilizador ao utilizar uma interface. Adicionalmente realizámos um segundo teste com uma variação de estrutura da interface para validar a hipótese desenvolvida na primeira avaliação. Finalmente foi-nos possível validar a hipótese de que o Flat Design é tendencialmente menos usável que o "Skeuomorphism". Adicionalmente também nos foi possível verificar que a diferença de usabilidade é mais relevante em interfaces complexas do que em interfaces mais simples.

Keywords: Flat Design, Skeuomorphism, Experiência de Utilizador, Usabilidade, Avaliação

Abstract

Recently has been discussed the problems that Flat Design can potentiate in the usability of the user interfaces compared with Skeuomorphism. In this work we performed a comparative study between the two designs to understand if Flat Design influences the user performance while using an application. To better understand this issue we investigated work done in this area to find any evidence that aesthetics and design could influence the usability of an interface. Since the work was planned to be done at Webnographer using their remote usability tool, we also studied the existing usability test methods to check that their method was a good solution to apply in our study. In our work we applied the two styles: flat design and skeuomorphic design, to an application. After we tested the usability of these "different" applications and analyzed the results to check if the flat design affects or not the user performance while using the interface. Additionally we did a second test with a structural variation of the interface to validate the hypothesis developed after the first evaluation. In the end we were able to validate the hypothesis that Flat Design tends to be less usable than Skeuomorphism. Additionally, we also found that the difference can be more or less relevant depending on the complexity of the interface. In other words we can see more improvements in complex interfaces than in simple interfaces.

Keywords: Flat Design, Skeuomorphism, User Experience, Usability, Evaluation

Table of Contents

Table of Figures

Table of Tables

1 INTRODUCTION

In the last years a minimalistic design genre called Flat Design is becoming popular and used in different interfaces such as websites, applications, etc. One of the first usages of flat design was the windows phone 7 in the end of 2010. However the first big application changing to flat was the Microsoft operating system, Windows 8, in 2012.

Figure 1 – Windows 8 Start Screen

Since then Flat Design has become widely used in the newest websites, tools and applications, like Mac OS and IOS, from Apple, or Android and Google web applications. We also have front-end frameworks like Bootstrap [1]and Zurb Foundation[2] that use as default the flat style.

But what is Flat Design? Which are the characteristics of this style? This style can be described as a simplification of an interface by removing aesthetic elements such as shadows, bevels, textures or gradients. In other words flat design removes any aesthetic element that give the three dimension illusion and depth sense focusing on the minimalism using simple elements, flat colours and typography's (by aesthetic elements we mean any decorative element of the interface that can be used to transmit to the user the feeling of interactivity).

In contrast we have Skeuomorphism, which is more than a design genre, it's a design technique. The word skeuomorph was defined in 1890 as "*An ornament or ornamental design due to structure*" [32], or a physical ornament or design on an object made to resemble another material or technique. In human computer interaction

[1] http://getbootstrap.com/
[2] http://foundation.zurb.com/

they have the same meaning on the interfaces. Skeuomorphs are the metaphors that help the user understand the functionality of the interface. In other words they use the aesthetic elements that were removed in the flat design to create this metaphors. Thus, a skeuomorphic graphical user interface emulates the aesthetics of physical objects as explained by Mullay [33]. For example if we look to the flat button and flat label on Figure 2 and Figure 3, it is not clear which one is the button and which one is the label.

Figure 2 – Bootstrap Default Button

Figure 3 – Bootstrap Default Label

However if we look to Figure 4 the button as bevels and gradients that make the button looks clickable in comparison to the other two. This cues are known as perceived affordances.

Figure 4 – Skeuomorphic Button (Bootstrap 2.3.2)

The concept that objects have a couple of characteristics that tell to the person or animal what they should do with them is old and they are called affordances. The first person calling this characteristics as affordances was James Gibson [25]. In his work Gibson said that an affordance is something that transmit to the user the meaning for what that object can be used. For instance water could afford drinking. After Gibson's work Donald Norman did a work where he applied this concept to the human computer interaction. That work was called *The Psychology of Everyday Things* [28]. In this work Norman related the affordances not only with the physical object but also with the user goals, plans, past experiences, etc. Some years later he revisited his work [29] to explain that he was not talking about affordances. He claimed that he was talking about "perceived affordances". The difference is that the affordances are always there, are something that are with the object. Perceived affordances, on the other hand, only exist when the user as the need to accomplish a goal. Some other works were developed combining the concept of affordance with the technology as ***Technology Affordances*** from Gaver [30] or Affordances in HCI from Kaptelinin et al. [31], which we will explain in detail later.

So if we go back to the skeuomorphism what this concept applies is replicate the affordances of the real world on the interfaces. And the way that they do it is through the effects to create the metaphors with the real

world. For example we can use bevels and gradients to do a button look like a button. However all of this clues and “affordances” were removed on Flat Design for the sake of minimalism.

Based on the works that we described before and that we will describe with more detail on the literature review our hypothesis is that Flat Design is less usable than a design that apply skeuomorphism. In other words by removing the style effects on flat we could cause a lack of affordances making the interface less usable.

In the next sections we will describe the objectives of our study. We will also do an overview of our solution, with a quick description of our contributions and results. Finally we give an overview of the dissertation structure.

1.1 OBJECTIVES

The main purpose of this study is to understand if flat design and/or Skeuomorphism influence the usability of applications. In particular, we want to know if by changing the interfaces from Skeuomorphic to Flat we are affecting the usability. The book published by Donald Norman in 1988 *The Psychology of Everyday Things* [28] is one of the most known works developed describing how important the affordances present in objects, that we use every day are on giving clues to the user on how to use them. William Gaver work Technology Affordances [32] is another work that describes how important are the affordances for the usability of an interface. In addiction we also have other works researching the relation between affordance and usability like Affordances in HCI [33], Human Affordance [34] and Affordance as Context [36].

To prove that flat design can influence the usability, due to the lack of affordances, we developed the hypothesis that the Flat Design is less usable than Skeuomorphism. In other words we believe that by removing the stylish (aesthetic elements like gradients and bevels for example) from the interactive elements (buttons, title bars, etc.), we are also removing the affordances for that element. Consequently the usability of the interface will be affected.

To do and validate this comparison we will perform usability tests with a real application using flat design and change their interfaces adding affordances by changing buttons, links, etc. Then we will perform to compare between each variable Flat and Skeuomorphism. Finally, we will analyse the results and feedback from the users to understand if flat design influences usability or not.

1.2 SOLUTION

To prove the hypothesis explained in the previous section, we tested a real application related to tax return submission. We used this application since it was a Webnographer project developed and planned by Sabrina Mach and the Webnographer team (Additionally Webnographer was allowed by the client to use the test data as

a Case Study of the company and allowed them to publish as a Sample project, for that reason we were able to use it on our study since Webnographer gently provided their data). Also two interfaces with different structure were tested, this two interfaces were also developed and planned by Sabrina Mach and Webnographer with Simpletax Company.

To build the test conditions to compare flat with Skeuomorphism design we applied some aesthetic changes between the two interfaces (derived from the same interface). In other words we changed the appearance of buttons, widgets, etc. without changing the structure or organization of the application. In each interactive element we changed it from flat to skeuomorphism by adding effects like gradients, bevels or underlines on words for example. In summary the test was current design Flat vs current design Skeuomorphic and new design Flat vs new design Skeuomorphic. In other words a current interface design with a level of usability and a new interface design with a different level of usability. The main concern was not to affect the two variables at the same time.

In order to compare between the two different conditions the same usability test was performed to each one of the "different" applications with the user performing the test in only one of the conditions. After that we analysed and compared the results to understand if the different styles (Flat and Skeuomorphic) influenced the usability of the applications.

To perform these tests we used an asynchronous remote usability test method. This has the main advantage of allowing to perform the tests with a considerable sample size, doing the tasks on their own environment minimizing the test influence on the user.

Finally, a survey, developed by Sabrina Mach from Webnographer, was performed on both interfaces tested with the remote method, to get demographic data and also satisfaction rates from the users. All the steps of the test questionnaire and usability test were performed on a single survey that is possible due to the Webnographer tool that we will explain better on a section dedicated to it.

1.3 CONTRIBUTIONS AND RESULTS

In this research our main goal was to validate if Flat design is less usable than Skeuomorphism or not. After evaluate the usability test results we were able to understand that Flat has influence on usability. In other words after the first test we found that by doing the change from flat to skeuomorphism the users were able to slightly improve their performance. However we could also understand that this improvement was not for the overall application but only in some particular steps. With this observation we developed the hypothesis that the improvement of the usability level would not be relevant in simple interfaces. After performing the second test we were able to verify that hypothesis.

In summary with our research we were able to validate that Flat tends to be less usable than skeuomorphic, but this difference is only relevant when we are using complex interfaces. For example interfaces like forms are so simple that the user will be able to understand what he need to do even on a flat interface.

1.4 WEBNOGRAPHER COLLABORATION

This Study was designed and developed at Webnographer[3] with the collaboration of Sabrina Mach, as external advisor, and James Page (which also had an active role during all the process). Without them as without all the other members of the company this work would not be possible since it was based in all the knowledge methods developed inside the company. Since this work was developed at Webnographer and with Webnographer all the processes and methodologies used by the company and applied during this work were designed and developed by Sabrina and James for Webnographer. Additionally, some of these processes and the way they are used by the company are confidential and for that reason in some parts of the dissertation we could not provide all or any details on how we applied their methods to perform our work.

1.5 DISSERTATION STRUCTURE

This dissertation has five chapters. The first and current chapter where we introduce and explain the context of our work, what is our solution, the role of Webnographer and the main results and findings of our research.

Then on the second chapter we present research that shows the context of our work and the related work previously done on this subject. This chapter is divided in two main topics **Design and Usability** and **Usability Test Methods.** In the section design and usability we have two different research areas one that is related to the concept of affordance and the importance of it to help the user how to use an interface or object that he can use. The other research is about the relation between aesthetics and usability. Basically in this section is shown the influence that the style/aesthetics of an interface can have in the user performance while he is performing his task. On the section about the usability test methods we present the current options available to apply usability tests. The main objective is to compare the available methods and understand their advantages and disadvantages in order to explain that the asynchronous remote testing (Webnographer tool and methods) is a good solution for our work.

On chapter 3 we explain our proposed solution. First in section 3.1 we explain our approach to validate our hypothesis that Flat Design is less usable than Skeuomorphic, like how we will test and the tool we will use, then on section 3.2 we will explain the basic research methods applied to perform the usability test that were adapted and applied by Webnographer.

After the theoretical context on chapter four we will show the preparation that we have done to perform the usability test, this will be developed on section 4.1. Then on section 4.2 we will present and analyse the results that we got from our usability tests and check if we can validate our hypothesis or not. Finally on section 4.3 we will discuss the results and present our conclusions based on what we observed on the previous section.

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

In the end on chapter 5 we will do an overview of all the dissertation content and present the main findings and conclusions about our work, and we will propose some ideas on how our work can be continued and improved with future research.

2 CONTEXT AND RELATED WORK

In this section we describe two main topics: **Design and Usability** and **Usability Test Methods**. In the first topic we analyse some papers describing the relation between affordance and usability and also other works describing the relation between aesthetics and usability and how this aesthetics can influence the user performance.

Then in the second topic we analyse the different methods that we can use to evaluate user interfaces and we explain why the remote asynchronous usability testing that we will use is a good solution.

Finally we present a discussion were we relate the works described in section 2.1 and 2.2 with the work that we want to develop. Basically on section 2.1 we analyse and relate the papers about affordances, aesthetics and usability with the Flat Design problem, and then on section 2.2 we summarize the advantages and disadvantages of each usability test method to explain why our solution is a good solution.

2.1 DESIGN AND USABILITY

In this section we will present (2.1.1) some of the works that we found describing the relation between affordances and usability with special focus on the works related with Human Computer Interaction. After this we will also present other works (2.1.2) about the influence of aesthetics on the interface usability and user performance.

2.1.1 **Affordances and Visual Perception**

The idea that objects have certain characteristics that help us understand how to use them is a concept that started a long time ago. This attributes that are contained on the objects were named by James Gibson as affordances [25]. Affordances are perceived by animals as possibilities for action in the environment. Also the affordance is always there even if it is not perceived. Either because it is not needed or because is not visible. As Gibson explain in his work ***The Ecological Approach to Visual Perception*** [26]:

> "The concept of affordance is derived from these concepts of valence, invitation, and demand, but with a crucial difference. The affordance of something does not change as the need of observer changes. The observer may or may not perceive or attend to the affordance, according to his needs, but the affordance, being invariant, is always there to be perceived."

In summary, to Gibson affordances are not dependent of interpretation they are perceived directly. Also they are relational properties that emerge in the interaction between animal and object. In other words is something that is contained in the objects and always present. However, it will only be perceived if the user (human or animal) as the need of using it.

In 1988 Donald Norman introduced the concept of affordances to human computer interaction [28]. In his work Norman described affordances as perceived or real properties of the object that determine how to use them. In other words the properties are cues on how to use or operate the object. And also according to him, we can use the affordances as an advantage to allow the user to know what to do, even without labels or other kind of instructions [28]. Later in 1999 Donald Norman felt the need to clarify his work on ***The Psychology of Everyday things*** [28]. This happened because people misunderstood affordances from Gibson [26], the real affordances, with Norman "affordances" that are actually perceived affordances (as he clarify on ***Affordance, Conventions, and Design***) [29]. In other words Norman was talking about the reaction caused on the user by the affordance that do not need to be a real affordance.

Between this two works from Norman, in 1991, we had also an interesting work that tried to clarify and apply the concept of affordance on the human computer interaction field. This work is called Technology Affordances and was developed by Gaver [30]. In this work the author lays out a framework for developing ways to apply the notion of design on interfaces. More precisely Gaver shows how we can improve the usability of interfaces by applying the affordances concept to the computer interfaces with the objective of giving the clues to the user of how to work with the interface. However the way that Gaver approached the concept of affordance was based on Gibson's concept of affordance. In a recent work from 2012 Kaptelinin and Nardi [31], argue that Gibson's concept is correct but can't directly he applied on the world of human computer interaction. For them HCI needs a broader concept of affordances. So for them the theory for the affordances in HCI needs to be different from the Gibson's theory. As they argue the most fundamental insight of socio-cultural approach is that human action and mind are inherently mediated. Our action capabilities to a large extent depend on socially developed mediating means, first and foremost tools, including technological tools. Based on that they propose understanding technology affordances as possibilities to mediated human action. On their work they present an initial outline of the mediated action perspective on affordances that focuses on individual human action. As future work they say that a necessary next step is to extend the analysis to collective actions.

2.1.2 Aesthetics and Usability

The influence of aesthetics in user experience has been studied in several works using different approaches. Bargas-Avila, J.A. and Hornbæk, K. [2] made critical analysis of empirical studies on user experience. In this study they identified that the most frequent researches were about aesthetics. One of the first studies on this subject was from Kurosu, M. & Kashimura, K. [9], where they concluded that the apparent usability is correlated with the apparent beauty. Two years later Tractinsky, N. [16] revisited this study and concluded this relation too. However these two papers are theoretical works.

The first experimental study that we found about this subject was done by Tractinsky, N., Katz, A.. & Ikar, D. [17]. In this article authors intended to relate the perceived aesthetic and usability of pre-use and post-use. For that they defined two main goals. The first was to test if the initial correlation of perceived aesthetics and usability reflected a general tendency to associate aesthetics with other system attributes. And the second was to explore what happens to the user's perceptions of aesthetics and usability after they use the system.

After defining the objectives they developed the method to perform their study. Relatively to the participants they selected 132 students from Industrial Engineering. This students were all from the third year, 67% were males and the average age was 25 years old. Then they used two different factors aesthetics and usability. For the aesthetics factor they gave to the participants 26 ATM layouts to rate relative to aesthetics. After that they choose nine of this 26 layouts. Three were the most rated, the other three were the lowest rated and finally the last three ATM layouts were rated in the middle. Then to select which one was the layout that each participant would work they gave them the layout that the user rated as the better in relation to a factor of aesthetic evaluation. Relatively to the usability factor they presented to the participants a set of 11 tasks to be performed on the ATM. The usability factor was manipulated by introducing interaction problems to the machines like delays and malfunction buttons.

In the test procedure the authors gave three layouts (a first one with low aesthetics a second one with high aesthetics and a third one between) to each participant to test. After they tried the three different layouts they were asked to perform the 11 tasks in each of the layouts. This 11 tasks were comprised of the following four types: inquiring about their account balance; withdrawing cash; checking out the account balance and withdrawing cash simultaneously; and depositing money. This tasks were presented in a secondary panel aside the main panel.

This study corroborates the results of earlier studies (Kurosu, M. & Kashimura, K., 1995 and Tractinsky, N., 1997) that found a strong correlation between user's perception of an interface aesthetics and their perception of the usability of the entire system as we can see in Figure 5. They also found that users tended to rate the aesthetics better after using the system. According to the authors this can be explained by natural adaptation of the human being to something that is required to use.

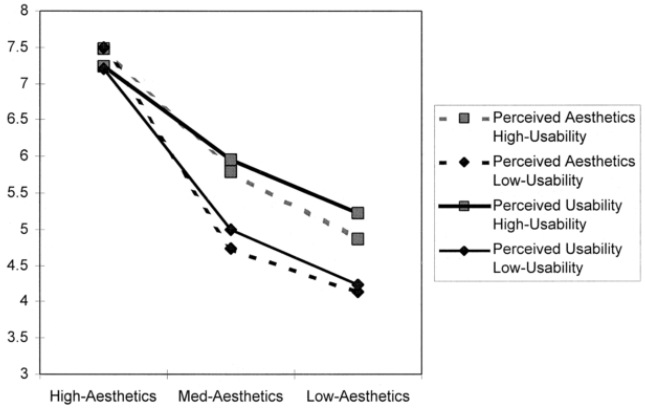

Figure 5 Post-experimental perceptions of usability an aesthetics (on a 1-10 scale) under three levels of ATM aesthetics and two levels of ATM usability

A very interesting finding in this study is the fact that post-experimental perceptions of the system usability were affected by the interface's aesthetics and not by the actual usability of the system (like we want to verify).

Two limitations of this study are the users who performed the tests and the interface. Because all the users have the same background and the test has a lack of variety of personalities. Then the authors generalize these findings from a single interface. In conclusion the author admit that is important to continue studying these relationships during a longer time frame.

Other studies were developed after this, and most of them found the same correlation, like Van Schaik, P. & Ling, J. [20] or Lavie, T. & Tractinsky, N. [10]. However some studies like Hassenzahl, M. [7] or Van Schaik, P. & Ling, J. [21] did not find any correlation between perceived aesthetics and perceived usability.

Lee, S. et al. [11] developed a work answering to the methodological limitations of the previous studies by using a new methodology to examine perceived usability/aesthetics and user preference in an experimental setting. To execute this work they developed nine hypotheses based on usability and aesthetics divided in three parts: interaction before actual use (hypotheses 1-1, 1-2 and 2); interaction after actual use (hypotheses 3-1, 3-2 and 4); and comparison of interactions before and after actual use (hypotheses 5-1, 5-2 and 5-3). To test the nine hypotheses the authors implemented an experiment that used four simulated systems with different usability and aesthetics levels. To do this they selected seventy three students majoring in engineering. From these users 59 were males, with an average age of 23.68 and with 3 different nationalities.

To apply the tests the authors developed four different systems that vary between low/high usability and bad/good aesthetics, all of them with the same information content (as illustrated in Figure 6 and Figure 7). Then the participants were required to perform three major experimental tasks: evaluate perceived aesthetics, perceived usability and user preference before actual use; complete four scenarios tasks on system; assess perceived usability, perceived aesthetics and user preference after actual use.

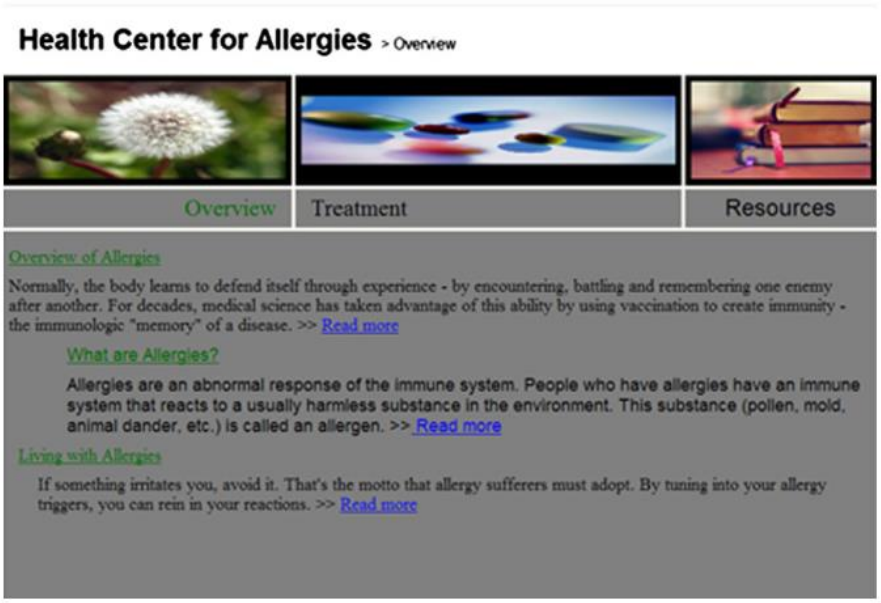

Figure 6 System with low aesthetics

Figure 7 System with high aesthetics

The first task was to rate an assigned system with regard to usability, aesthetics and user preference before actual use with 8 statements for perceived usability, 11 statements to perceived aesthetics and 1 statement to user preference. In the second task participants were required to complete four scenario tasks on the assigned system. These tasks allow the participants to use the system and perform the last major task. In the third task after the participants use the assigned system they were asked to rate the system using the same method that was used on the pre-use evaluation form (only the statement tense changed).

In the results they began by checking the manipulation between high/low aesthetics and high/low usability. Relatively to aesthetics they obtained a result of 4.74 points in the high aesthetics website and a 3.13 points in the low aesthetics website (the scale is between 1 and 7). This results indicated that the manipulation aesthetics was useful. Relatively to the usability manipulation check they compare the average high completion time. They obtained a result of 153s in high usable interface against 299s in low usable interface. They concluded that the manipulation of the usability factor was successful. Comparing the high aesthetics usability with the low aesthetics usability they also concluded that the usability was free from any aesthetics side effect. However in our opinion and based on the Figure 6 and Figure 7 this statement may contain some doubts given little difference between the two websites.

Regarding their hypotheses based on the results obtained in the analysis they can say that all of them were supported with the exception of the hypothesis 1-2 (before actual use user preference was marginally affected by differences in usability) that it was only partially supported.

In the analysis of hypothesis 2 the authors did a very interesting finding. Basically they identified that before actual use, the rating of perceived aesthetics was higher in the high usability condition than in the low usability condition. This supported that the aesthetics and usability were interrelated and affected by each other.

Another interesting finding in this study was the relation between aesthetics and usability. They concluded after analysing the results of users on tasks and his satisfaction in the high usability systems, that although users did not have a significant worse performance on the tasks, users rated the interface with the worse aesthetic as less usable. This supports our hypothesis that flat design can affect the usability of a user interface. Mainly because the difference between a flat interface and a skeuomorphic interface is much more significant than the aesthetic difference in this study.

The authors also detected that systems with a low usability were low rated by the users in aesthetic. In other words on hypothesis 5 making a comparison between the rates before and after actual use the users rated better systems with high aesthetics and low usability before actual use than after actual use. This indicates that the aesthetics can be influenced by the usability too.

In conclusion the authors found a high correlation between perceived aesthetics/usability and user preference. Also this study confirmed and clarified the findings made by others previous studies. This study introduce a new methodology where usability, aesthetics and occurrence of actual use were simultaneously considered in a more complete setting. However the authors identified four limitations that need to be solved in future works. First, we need to test different applications in different areas so that we can verify that the same results are obtained. Second the system was not considered as an influence factor. The author consider that in future studies is necessary that the users are interested in the system and with little or no experience so that the results are less influenced by external factors. Third the population used in the tests was principally male engineering students and the study only can be generalized by this homogenous nature of participants. In future works is necessary that participants are more scattered with different environments and it is necessary to take into account the cultural factor. Finally, like we had identified in future works the differences between aesthetic and usability levels need to be more deeply study.

Tuch, A.N. et al. [12], in 2012 performed another study about the correlation between interface aesthetics and perceived usability. Based on the study of Hassenzahl, M. & Monk, A. [13], they identified the lack of experimental studies on this subject. To perform this study they identified the principal problems in previous studies and tried to present solutions in order to solve them. For this study they formulated 3 hypotheses: Interface aesthetics affects perceived usability before usage, interface aesthetics affects perceived usability after usage and interface usability affects perceived aesthetics after usage. To perform their work they build four different websites of an online shop with two variables interface aesthetics (low vs high) and interface usability (low vs high). Then they choose 80 participants (42 females) with an average age of 25.7 years old and the mean experience in using web was 10.8 years and all of them had previously shopped online. We can consider this propose a solution to the same problems found by Lee, S. et al. [11] in their study.

To manipulate the usability they maintain the same structure and menus but change the labels of the menus and submenus. Basically they change the categories like can be seen in Figure 8. Then in order to choose the ugly and the beautiful design they pick 30 professionally designed website templates. After that 4 experts choose the 10 most ugly and the 10 most beautiful. Finally 178 users choose from this two sets the ugliest/beautiful pair.

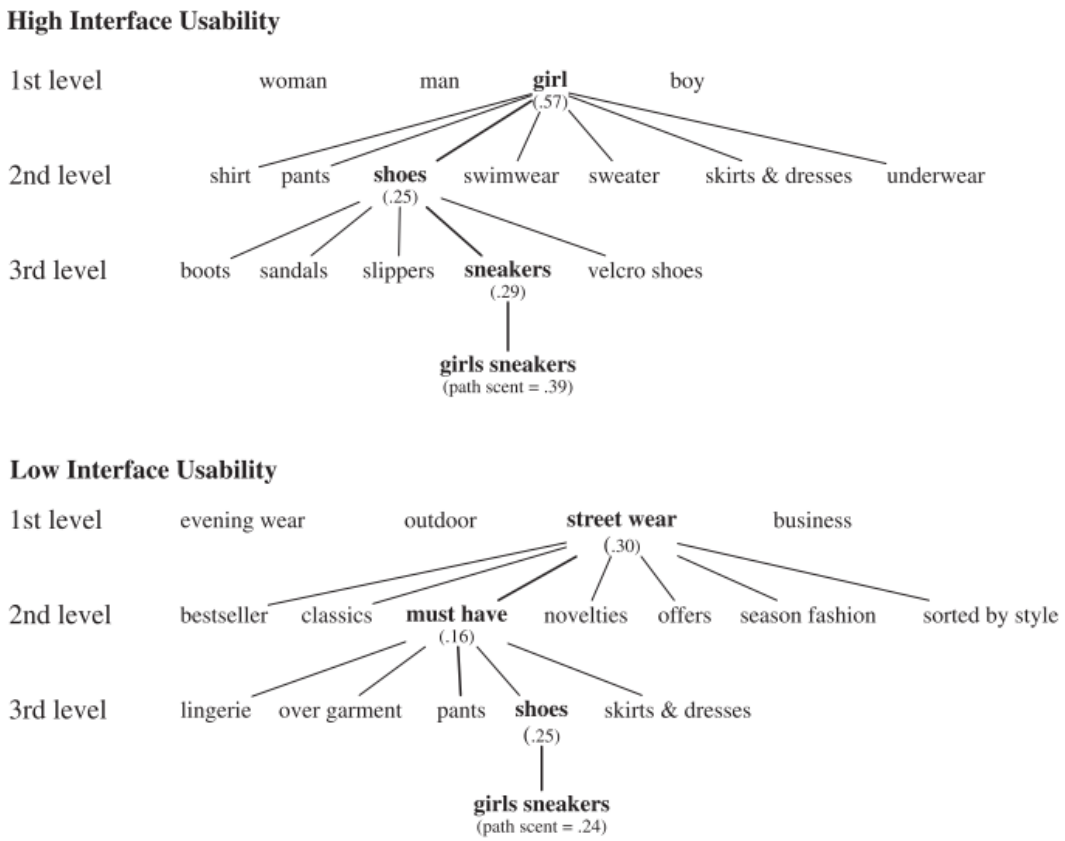

Figure 8 Example of navigation path on the online shop with high and low usability

The online shop was a clothes shop and was fully implemented in order to perform this test. Then they defined four similar tasks that consisted in finding a product and add this product to the cart. The users had 5 minutes in maximum to perform each task.

In the test procedure the users perform three steps. First were presented for 10s a preview of the online shop and the user rated this one according to perceived aesthetics and perceived usability. After that the user performed the four tasks and after each task they rated their user experience answering some questions. In the end the user was asked to evaluate the entire interaction principally in terms of aesthetics and usability.

Before the main analysis the authors could check that the factors interface aesthetics and interface usability were usefully manipulated performing a two way ANOVA with perceived aesthetics and performance as dependent variables.

Against the authors expectations the first hypothesis was refuted. In their experiment the users did not use the interface's aesthetics as a proxy for pre-use perceived usability. In the post use phase they also did not find relation between aesthetics and usability, refuting the second hypothesis and contradicting previous studies in this subject (like Tractinsky, N., Katz, A.. & Ikar, D. [17]). With regard to the third hypothesis the authors could observe that after use, the perceived aesthetics was influenced by the usability of the website. In other words if the usability of an interface was bad then the user will reduce the rate of the aesthetics. This can be explained by the affective experience of the user, i.e. if the user can't use their interface easily or with success He will have tendency to dislike the application thereby reducing his review on various aspects.

In conclusion the authors not only contradicted the influence of aesthetics in perceived usability supported by some previous studies as also support that the usability can influence the perceived aesthetics. However they assume that their usability manipulation was stronger than the aesthetics manipulation and this might have influenced the results. A limitation in this study was again the use of a single product to support their findings. Another limitation identified by the authors is the performance oriented tasks defined that may have led the user to focus too much on usability issues distracting the problems of aesthetics. Finally the authors concluded that in further studies the manipulation level of usability an aesthetics need to be more worked in order to understand the boundary conditions of the aesthetics usability correlation.

Another interesting study is from Sonderegger, A. & Sauer, J. [14]. The difference between this work and the others already described is that this one is more focused in the influence of aesthetics in usability testing. To perform this study the authors choose two mobile phones because it has a stronger affective component than most other interactive consumer products. Based on the literature reviewed by the authors they form three hypothesis: User performance will be better for the more aesthetically pleasing product than for the less pleasing one; Perceived usability will be higher for the aesthetically more pleasing product than for the less pleasing one; and The difference in perceived usability between the two conditions will be less pronounced after the usability test than prior to it. To perform this work the authors selected 60 participants from a secondary school aged between 13 and 16 years old. The average use was of 8.7 times per day and they have rated their experience on 65 in 100. Besides that they have no difference between males and females. Then the users was randomly attributed to the appealing and non-appealing mobile phone. To measure the study the authors defined three categories: Perceived product attractiveness, Perceived usability and User performance. In the Perceive product attractiveness the users need to rate the product in some items with a scale of seven points from strongly agree to strongly disagree. For the Perceived usability the users was departed against with some items to rate with the same scale. Besides that the users need to answer a questionnaire to better understand user opinion. Related to

the User performance the authors measure three indexes: Task completion, Interaction efficiency and Number of error messages (when the user choose a wrong navigation option).

To assure the usability of the two mobile phone interfaces the prototypes were based in an already existent mobile phone (SonyEricsson SE W800i). However in the two computer prototypes only the functionality needed for the study was implemented and not all the functionality. To define the aesthetics of these two interfaces the authors performed a pilot study selecting 10 participants to choose the appealing and non-appealing mobile phones aesthetics.

For the test execution the authors defined two tasks. The first task consisted in sending a text message to someone. The second task was a little bit more complex and involved changing the mobile settings in such a way that one's own phone number is supressed when making a call.

After analysing the results the authors found that the appealing prototype was better rated than the unappealing. They verified too that after usage the appealing increased the rating against the unappealing that decrease very significantly. Relatively to the perceived usability the authors observed that the usability rating was the same before and after usage in the two prototypes and that this rating was not influenced by the aesthetics. In User performance was detected that the users needed less time to complete the tasks on the appealing prototype. Like the Task completion the Interaction efficiency was superior on the appealing prototype. Finally, the users performed fewer errors on the more attractiveness prototype.

In conclusion one more time was demonstrated the influence of the aesthetics in the perceived usability like in previous studies. However the authors proved in contrast with other studies that the user performance is affected by the aesthetics obtaining better results in good aesthetics than in bad aesthetics. The limitations that could exist in this study are the population tested that could be more embracing including other ages. The other limitation is the lack of tasks (only two) that we believe are few to actually prove the results obtained.

In section **Error! Reference source not found.** we will draw some conclusions about the related work analysed relating it to our work and how this research will help us in not doing the same mistakes made in the previous studies.

2.2 USABILITY TEST METHODS

To prove our hypothesis we needed to use an usability test method. The method planned to be used in our work was the remote asynchronous usability testing, supported by Webnographer. In order to demonstrate that this method was a good solution to perform our tests, we performed a research of the main available methods to compare their advantages and disadvantages. Relative to the remote asynchronous usability method that we describe in particular on this study, even being different from the Webnographer method we consider being a

good paper to understand the concept of a remote asynchronous method and the general advantages of this method. In the end of the section 2.3.2, we do a comparison between the method used by Tullis, T. et al. [19] and Webnographer to show the main differences.

2.2.1 Heuristic Evaluation

One of the most known usability test techniques is heuristic evaluation that was developed by Nielsen et al. [12]. This technique consists in an evaluation carried out by experts. In other words to assess a user interface we give the interface to some experts so then they do an evaluation identifying possible issues based on heuristics. To identify the issues they provide a description about it, which heuristic is being violated (one or more) and the severity of this problem and also a possible solution if asked.

However this evaluation has some problems. One of this problems was identified by Jiménez, C. et al. [8] in the paper ***Formal specification of usability heuristics***. In this paper the authors address the problem of the difficult interpretation or various interpretations that the defined usability heuristics have. Additionally the authors try to prove that standardize the way heuristics are applied need to be well specified. This is necessary to guarantee that the heuristics are interpreted and applied in the same way by all experts. To prove this point they did tests with 20 evaluators without experience to prove that different people have different interpretation when they see the heuristic for the first time. After analysing the results of the tests, the authors were able to prove their idea. Because of the lack of specification of the heuristics the evaluators had some difficulty in relating a heuristic with a usability problem. For this reason they conclude that if the definition of the heuristics were more specific, by being better described and including examples, evaluators probably could apply better this heuristics.

In conclusion the authors say that this study can be more explored and that they will do another type of analysis and other techniques to confirm the result obtained in this work.

Although we can agree with the authors in this interpretation that the heuristics can be confusing or difficult to understand their meaning, the truth is that this study needs further testing and try to test more users with different mind sets in order that one can say with certainty that actually this kind of evaluation can bring these problems.

2.2.2 Laboratory Testing vs Remote Testing

Another technique for usability evaluation is user testing. This type of tests can be divided in two different approaches: lab testing and remote testing. Lab testing is done in a controlled environment where we tell the users what we want them to do and how, while we observe and measure their performance during the test. Remote testing like lab testing also uses a script with tasks where we describe what we want the user do. However the main difference is that we can test a user wherever he is, because the test is not limited by the distance or the local taking into account that it runs remotely. On the paper ***An Empirical Comparison of Lab***

and Remote Usability Testing of Web Sites Tullis, T. et al. [19], do a comparison between this two types of test methods and describe the advantages and disadvantages of both.

There are two types of remote tests, the synchronous and the asynchronous (the type of method used by Webnographer, that we compare with Tullis, T. et al. [19] in section 2.3.2 and we explain in some detail in section 3.1.3). The first only have the advantage of the local, since all the other protocols are very similar to the lab-based tests. Basically instead of the user being observed by the moderator in the lab the moderator observes the user using a webcam and a microphone or with something that can substitute these tools. However, for the author the synchronous remote tests are not particularly interesting because they require that the moderator spends time observing the user and we can't reach more users than in the lab tests. For this reason the author chose to test the asynchronous remote tests against the lab tests because this way he can reach much more users than with the synchronous method.

When the study was carried the only way to capture certain types of interactions was using instrumented browsers installed on the user's computers. Because of that the authors opted for another approach. Basically when the user initialized the test two windows were showed on the screen, the normal browser with the website and other window with the task. When the user finished the task he confirmed on the window and answered a little survey about the task and rated it. Then it shows the next task, until the last one. For each task they saved the time that the user spend in that task.

The authors conducted two different experiments. Firstly they collected some data to compare the two types of tests, then they did a second experiment to validate and improve the first experiment. They wanted to identify the advantages and disadvantages of both tests too.

In the first experiment the lab and remote testers did the same tasks and answered the same surveys and both were alone in the moment of the test. The difference between them was that the lab testers were been observed by the moderator and everything that the user did was recorded. However regarding the remote user the moderators only knew what the user reported by the surveys. Four main types of data was considered to evaluate the tests: task completion, Task Time, Subjective Ratings and Usability Issues. It was possible to conclude through the first two metrics that the difference between lab and remote users was not significant so the environment didn't have influence in this subjects. In relation to usability issues although the number in both tests was different the most important problems were the same. However relatively to the subjective ratings the remote users gave more negative ratings. This can be explained by the difference between the sample sizes (8 lab users against 29 remote users). One surprise in the remote tests was the very reach comments provided by the users that almost substitute the direct observation.

In the second experiment was possible to verify the results of the first experiment. The results of the two first metrics were very similar to the first experiment. Relatively to the usability issues it was possible to observe again that the principal problems were the same. However they concluded that the remote users detected some relevant problems that the lab users couldn't detect, probably derived to the sample size and diversity of users. Relative to the subjective ratings unlike the first experiment this time remote users gave better rating than the lab users which support that only 8 users are not reliable.

In conclusion, by analysing the experiments results was possible to prove that the different environments don't influence the behaviour of the users and that they find the same big problems. The authors realize that the comments provided by the remote users are very rich and that can in some cases substitute the data collected via direct observation. This information complemented with software that capture the users interactions can be very complete. A particularly advantage of the remote tests is the diversity of users that we can reach and the several environments that we can test. This type of tests also provide more reliable subjective assessments, because of the sample size. However, the remote tests implies always the loss of the information provided by the user observation and this is a clear disadvantage.

Finally, according to the authors if we want a complete usability evaluation we have to use the two types of tests. However if we only want to solve the biggest usability problems they believe that the remote evaluation is better because it allows us to identify more usability problems than the lab tests.

2.2.3 Moderated Remote Usability Tests

After analysing the study made by Tullis, T. et al. [19], and considering the conclusions, we could support that remote usability testing in general is a good approach to evaluate our hypothesis (furthermore if we consider that Webnographer has a method more developed and complete than the method explained before, that we describe). We consider remote methods in general a good approach since the differences of usability between Flat and Skeuomorphism can maybe not be visible with a low sample of users, and also our main concern is not about the severity of usability issues but more about if the usability issues exists or not, which is more likely to get more usability issues with remote since we are testing more users. However as Tullis proved on is study [19], the main problems are the same. Another advantage is that users do not feel so pressured when they give their opinion on asynchronous remote method, this happens because the researcher is not present on the room/session (like lab testing) which can make the user not be honest if he wants to give a bad feedback to the evaluator [19].

Since we were able to support that remote usability testing is a good solution, we concentrate our search on current approaches using remote methods. As is mentioned by Tullis et al. [19], there are two different types of remote usability tests synchronous and asynchronous. Synchronous remote usability test is what was tested by Anon [1], in the article ***Here, there, anywhere***. The propose of this article was to perform an evaluation of remote usability tests to determine if this type of tests could be as effective as the lab tests and if this is true then we can do the same thing with less money and where we want.

To do this test they selected a well-known website. Then to proceed with the remote tests they selected a web software able to share screen, create log files, ease to use, and cheap (like licenses for example, etc.). In the end they choose Microsoft NetMeeting because of the ready availability and low cost. They used the web site marketing department to obtain the user profiles, then they choose ten participants (five for each test type) all with the same characteristics. They only chose five based on the study developed by Nielsen, J. [12]. After choosing the participants to the lab test, the administrator did the tests at a formal usability testing lab using the think aloud protocol. All the tasks were exactly the same that the remote users did.

For the remote usability tests the first thing they made was to send through U.S. mail all the things that the user will need to do the test. Then to ensure that the test would run well the administrator did a test drive of the software with the user before the test. The remote users did exactly the same tasks as the lab users and they have the same conditions like think aloud protocol and recorded sessions. In the end they were asked to answer a survey with some questions about the test and the web site. Finally they had to return all material provided for review by the administrator. In the results authors divided the analysis in four different topics: Time on Task, Number of Errors, Usability Problems identified and Post Test Survey. They concluded that the remote users took more time to complete the tasks than the lab users. They realized also that the remote users made more mistakes than the lab users, which clearly influenced the time of each task. They also concluded that the remote and lab users discovered almost the same problems and that they found the same number of usability problems. Finally with regard to the surveys they realize that both user groups gave similar answers.

In the end they concluded that the remote usability tests can be as good as the lab tests and can give good data as well. However, this type of tests have bad things too. For example we are unable to see the reaction of an user to something that he see, however this is compensated by the advantage of not having to create an environment in the laboratory and of not requiring the users to move to the evaluation place.

2.2.4 Automatic Remote Usability Tests

Another way to perform remote usability tests is using automatic evaluation softwares like is mentioned in one study made by De Vasconcelos, L.G. & Baldochi, L.A. [4]. In this article the authors talk about a recent problem related to the easiness with which anyone can develop a website just needing to know basic programming concepts. This is due mainly to the amount of frameworks that exist to help on the development of websites. However several of these websites do not respect some essential rules of design and usability heuristics. To identify these problems already exists some remote automatic and semi-automatic evaluation software's to facilitate the usability evaluation, which provide a more convenient way of evaluation and cheaper to the developers. However, the authors realized that these tools have problems in large and very dynamic websites (like commercial websites).

To solve this problem the authors created a tool called USABILICS. The main functionality of this tool is provided when a developer defines a task, then USABILICS use the COP model to identify all the alternative paths that the user can make to complete this task and add these alternative paths to the evaluation. The COP model is based on the identification of objects (buttons or textboxes), containers (that contain multiple objects) and pages (that contain multiple containers). Using this model the algorithm compares the similarities between the various instances in the website to verify if these instances have the same functionality of the task that the developer previously defined. This way it is easy to identify several tasks in the website without being necessary to describe them and the usability tests are much more detailed and comprehensive.

In next stage the tool does an auto evaluation of the data, then identify the problems and the errors and report them to the user. Later users suggested that the tool could also suggest corrections to the identified

problems. The authors accepted this suggestion and they decided to implement this functionality. After the tests the authors verified that this function has a good grade of confidence.

In conclusion this solution is good because it is totally automated and does not burdening developers or end users giving results with a good grade of confidence. However this system has a problem on the usability evaluation about tasks that are not linear. For example if we have a commercial website and we add an item to the cart and then continues the navigation without finishing the purchase, the tool assume this as an error, when the user in reality only want to add more items to the cart.

2.2.5 The different asynchronous remote usability methods

Since we will be using asynchronous remote usability testing, we think that it is important to explain some of the existent alternatives that can be used and how they can be used.

One of the existent methods is setting a task that the user will perform on the interface, and that we want to evaluate with freedom to navigate in the whole website/toll. In other words a task description is given to the user and he has to perform the instructions to reach the goal. To perform this kind of method remotely we have different alternatives. An alternative is by video and/or audio recording that will allow record all the user interaction with the interface and then the researcher can analyse the videos to get the data for the usability test analysis. Another way can be by recording the user interactions with the interface. This will not only allow to record the main interactions of the user but it will also allow some automatic data analysis that will help on the usability test analysis (like is done by Webnographer). This automatic analysis of data is not possible with simple video recording. Still regarding to the second way of testing this can be done like is done by Webnographer, that uses a web tool that is used by the user to perform all the test. Or it can be done through software that as the disadvantage of requiring the installation on the participant side, like is done by Userzoom[4] for example.

Another method is first click testing. This kind of test, like is explained in the website usability.gov, is good to see where a participant would click in order to complete a task that he wants to perform. With this kind of test we can get two interesting measures: If the user performed the correct action? And how much time they took? This is also a kind of test that can be performed by Webnographer.

A\B testing is also another kind of remote usability testing. As is explained by Jeff Sauro in his blog[5], A\B testing is basically a split test where we test for example a website with two different designs, and then we test a version A with half of the participants and another version B with the other half. In the end of the test we compare the usability test results of each test to understand which of the two versions is the best. However this kind of test has two limitations. First it only allows testing one variable at a time (like which kind of headers is

[4] http://www.userzoom.co.uk/
[5] https://www.measuringu.com/blog/ab-testing.php

better? Or which is the best color for the buttons?). The other one is that it requires a big sample size to make the test reliable.

This are just examples of the most known methods to perform unmoderated remote usability tests, since there are other interesting methods, like Multivariate analysis, which are an alternative to A\B testing, that allows analyse multiple variables at a time. However we will not describe all of them in this dissertation.

2.3 DISCUSSION

In this section we describe the main learnings of the two researches performed. In the first subsection, **Design and Usability Discussion**, we discuss the conclusions of the studies and the limitations that they identified and what we will do to solve this limitations. Then in the **Usability Test Methods Discussion**, we analyse the different usability test methods researched and compared the advantages and disadvantages of each one showing that the remote asynchronous solution that we are going to use (from Webnographer) is a good solution to the kind of study that we want to perform.

2.3.1 Design and Usability Discussion

To better understand the subject that we want to study (influence of Flat Design in usability) we searched works already performed regarding to design and affordances and the influence of aesthetics on usability of user interfaces.

About the concepts of affordance we could learn with the works from Gibson, The theory of affordances [25] and The Ecological Approach to Visual Perception [26], we could learn the importance of this concept to user. Basically for him the affordances are the clues to the user (human or animal) that transmit how to manipulate the objects. Another interesting finding is from Norman [28] that explained on his work the importance of design and apply this affordances in the objects produced by the human (physical or nonphysical like computer interfaces). We also learned with him on his work from 1999 [29] the difference between affordance and perceived affordance. Affordance is the characteristic of the object that is always present, although perceived affordance is the user perception that only exists if the user wants to use the object. Additionally with a more recent work from Kaptelinin and Nardi we could understand that the way how the concept of affordance was merged from Gibson theory to the technological field is not the correct. Not because Gibson theory is wrong but because the way that affordances work in computer interfaces is not the same as he developed for physical objects.

Related to the influence of aesthetics we have a few studies like Hartmann, J., Sutcliffe, A. & De Angeli, A. [5], which demonstrate no correlation between aesthetics and usability however we have much more studies that prove this correlation. The four papers that we summarize in the section 3.1 are examples of these studies.

In the first study that we analysed (Tractinsky, N., Katz, A.. & Ikar, D. [17]) the main lesson from it was the introduction to the subject we will address in our work. This paper that is one of the first studies in this subject was good to understand the basis and was very important to understand better the following works. We could with this work learn more about the correlation between aesthetics and usability and understood how they affected each other.

The other work that we analysed was Lee, S. et al. [11]. In this study besides the influence that perceived aesthetics has in perceived usability (already supported in the first study) we can learn that perceived usability also has influence on the perceived aesthetic. Another learning from this work is that if we really want to get good results we need to do a thorough manipulation of aesthetics and usability of the interface.

Then we analysed the study performed by Tuch, A.N. et al. [18], in 2012. This study unlike Tractinsky, N., Katz, A.. & Ikar, D. [17] demonstrates that perceived aesthetics does not affect perceived usability but in reality is perceived usability that affects perceived aesthetics.

In the last study (Sonderegger, A. & Sauer, J. [14]) we identified more similarities with that we want to perform in our work. The main focus of this study was not the influence of perceived aesthetics on the perceived usability but the aesthetic influences on user performance. The main conclusion from this study was that the aesthetics can really influence the user performance taking into account the results obtained by the authors. Since the similarity between our study and this one we will have it as a good reference for the development of our work.

In conclusion, we could identify two limitations that we consider be recurring in all works analysed. The first one is the number of interfaces tested. In all of the studies the authors only test one interface (that was changed in different aspects). However, we consider that in order to generalize our findings we should test different applications and compare the results of both so then we can validate our results. Also this interfaces should have different contexts in order to check if the results are the same for different contexts. Then the other problem that we identified was the diversity of the population. In all the studies the authors choose the users in a closed circle which meant that all users have the same mind-set. However these limitations are not relevant in our study since we are going to use Webnographer tool, a remote evaluation method that makes it possible to perform the usability tests with a more diverse population.

2.3.2 Usability Test Methods Discussion

Like we already explained on section 2.2.3, since our goal is to compare the usability of two different user interface styles Flat and Skeuomorphic the difference in usability could not be easy to see with a low number of test participants. And based on the works analysed before remote usability test method, that we will use, is a

good option because we can reach more users and with different experiences or cultures (Tullis, T. et al., July [19]).

We also performed a comparison between the asynchronous remote method and the other two different types of remote tests founded during the research performed. Relatively to the automatic evaluation De Vasconcelos, L.G. & Baldochi, L.A. [4], we consider that this approach is not good for our study, since from what we conclude from the study this method found issues based on the website structure and our hypothesis is not depending from structure, but from the style applied Flat or Skeuomorphism.

Thus only remains for us to compare the synchronous method with the asynchronous. The main advantage of synchronous method (Anon [1]) is that we can replicate almost all of the techniques used in lab testing (like think aloud for example). However with synchronous we can test more diverse people because we have not a distance restriction and the usability tests becomes cheaper. But if we compare with asynchronous method the principal disadvantage is that we can't achieve more users than in the lab tests because we still have time restrictions. Another advantage of the asynchronous method is more reliability on results because we can get more participants. However this method also has disadvantages. One of them is that we can not see the user reactions by direct observation which is only possible being present on the room or by video. But like Tullis, T. et al. [19] identified we can obtain good conclusions from the questions given to users to assess each task performed and this information can sometimes replace the direct observation. Other advantage that we can obtain of asynchronous method is that the users can perform the usability tests without the stress of being observed by someone and they don't fell being evaluated, which give us more realistic data because they perform tasks in a "real environment" [19].

Based in all of these findings even considering that the method described by Tullis [19] is very different from Webnographer method (that we will explain on section 3.1.3) the concepts that we learned from there and conclusions that we got, let us understand that the asynchronous remote usability test method is a good solution to perform our study and to evaluate if the usability is affected by the Flat Design or not. Moreover, if we compare the two methods (the method used by Tullis [19] and the Webnographer method) we can say that we have even more advantages, mainly because Webnographer solve some limitations that we can easily identify on the method described. First is the control that the researcher has during the test. In other words, conducting the test with Tullis method [19] the user need to perform the test in two different windows, one where he has the interface to use and the second one where he performs the questionnaire and insert the test data (like usability problems, time on task, etc). However in Webnographer the survey is completely done in a single window and is performed in a single flow, i.e. the user will perform all the survey steps (questionnaire, usability tasks, etc) without need to care with setups because Webnographer tool will guide him through the different steps. In second another big advantage from Webnographer to the method described is the dependency from the user to collect data. While in the method described the user will be in charge of start and finish the timer for the task and communicate all the problems during the task, in Webnographer all of this measures (like time, interactions, etc) are recorded directly by the tool without interference from the participant. This allows us to get much more detail about the user interaction with the tool and total control of what we record. We consider this very important since it allows us to base our analysis on users behaviour instead of only users feedback. Additionally

Webnographer also allows to collect additional information given by the user through questionnaires. In conclusion these are the main differences (and advantages) of Webnographer tool compared to the approach used by Tullis et all [19]. Nevertheless there are other very useful and important functionalities in Webnographer tool, like the automatization of data analysis for example, that makes this tool a very good solution to use on our study. However as we already mentioned in section 3.1.3 we will do a further explanation of the tool functionalities and also additional conclusions about the usefulness of some functionalities in our study.

2.4 SUMMARY

On the current section we discuss two topics **Design and Usability** and **Usability Test Methods**. In the first topic we resumed some works that were done describing the importance of the affordances for the usability and also for user understanding on what he can do with the object or interface that is being presented to him. Related to the influence of aesthetics on the usability. From this works we could understand that usability is influenced by aesthetics and also that perceived aesthetics is influenced by usability. An important conclusion was the user's background. This could be a problem due to the influence on test results.

On the topic about **Usability Test Methods** we described the most used methods in usability testing and we analysed the advantages and disadvantages of each one. We also compared them in a discussion section where we concluded that our solution, the asynchronous remote usability testing with Webnographer tool, is a good solution to our study.

3 PROPOSED SOLUTION

In this chapter we describe the approach used during the study and some of the methods and tools used at Webnographer and implemented by James Page and Sabrina Mach that were used to test our hypothesis and to develop and evaluate the Usability Test.

On the Used Approach section we explain three main concepts. The first is how we tested the difference between flat and skeuomorphism. Then we explain that we will use different interfaces (a current one and a new interface trying to improve usability) and why. Finally we explain how Webnographer method works and what we can do with it.

After, we have another section called research methods where we describe a statistical method adapted and implemented by Sabrina Mach and James Page in Webnographer that we used on our research. Here we explain and justify why the alternative method that we used is a good way to do the statistical evaluation. For that we compare the Bayesian inference (applied in Webnographer by James Page and Sabrina) against frequentist inference and we explain the advantages and disadvantages of each one.

3.1 USED APPROACH

In this section we will describe the approach used to prove our hypothesis. First we describe how we compare the difference of usability between the two designs Flat and Skeuomorphism. Then we describe how we validated the findings of the experiment. Finally we explain Webnographer tool and what we can do with this tool.

3.1.1 Flat Design vs Skeuomorphism

To perform the comparison between the two different designs we decided to use an application were we would apply the two styles flat and skeuomorphism. In other words for the same application we developed two "different" interfaces where only the style applied changes (without changes on the structure). So, for example, if we develop a website with flat (or getting one already done) then we will only change the style to skeuomorphism. These changes are done, for example, by adding gradient and bevels to a button to make them look like buttons.

After having the two variations of the application our method is compare the usability test results that we collected. Then to validate if our hypothesis is correct or not we performed two separated usability tests. Additionally, the participants should also perform task only in one of the variations to avoid affecting the results.

The reason is that participants could remember the interactions from the other interface and the test will not be done in the same conditions. The goal was to check if the users have a better performance doing the tasks with skeuomorphism or on the flat version.

3.1.2 Testing Different Interfaces

For this test we used a real application from Simpletax[6] (a real client and project from Webnographer). This company has a tool to help users on tax submissions. And as Webnographer project one of the goals was to compare the usability of the tool with two interfaces with different structure. To that end the current structure of Simpletax was changed. The reason why we evaluated this was to verify if the results would be the same if the interface had differences in usability.

To do this comparison the current version of the application was changed with the goal of improve the usability. In other words in the first two test iteration the current interface was tested with the original flat design and with Skeuomorphism. Then in the second test iteration the interface structure was changed, mainly by changing the workflow but also doing some visual improvements, like changing links to buttons for example. In the end we performed the second test iteration with this second interface also applying the two designs (Flat and Skeuomorphic).

Finally comparing the results with two different usability conditions we can understand if the conclusions are similar when we compare between Flat and Skeuomorphism. If they are similar then we can conclude that the style has influence on the usability. If not we will need to analyse the results to check how and why the issues are not similar.

3.1.3 Webnographer Method

On section 2.3.2, we supported why the non-moderated remote test that we use on our study to perform the usability test is a good solution compared with the other methods identified. Now in this section we will briefly explain how Webnographer works and the main steps performed in a Usability test. The Webnographer[7] tool is a proprietary tool designed and developed by James Page and Sabrina Mach from Webnographer. The tool allows us to perform the usability test and the questionnaire in one single survey. In other words we can in one single survey perform the questionnaire and perform usability tasks. Also the order is not fixed we can do questions when we want and usability tasks when we want and even as multiple tasks and questions mixed in the survey. Additionally a very important feature of this tool is that is a browser tool, in other words contrary to other tools, there is no need to install additional software on the participant side which allows an easier access to them. This is also true relatively to the client, in other words the client doesn't have to install anything to allow the evaluation by Webnographer tool. However Webnographer is not just a tool, they follow a method that was developed by Sabrina Mach and James Page on their Usability tests. To apply that method they will perform some steps. First they do a preliminary analysis, the objective of this analysis is to understand which is the

[6] www.gosimpletax.com
[7] www.webnographer.com

current status of the interface being tested and what are the potential problems of this one, due to confidentiality we can't give details in how this analysis is performed. Then based on the analysis is designed the Usability test. After preparing the Usability test, it is launched to be performed by the participants that can be sent by the client or can be used a recruitment agency to send participants to the test for example. Finally the results are evaluated and the conclusions can be reached based on the data collected.

Now to better understand how the Usability test works we will explain how Webnographer tool works. The tool has two different perspectives, the participant and the researcher. From the participant perspective what is seen is just the Survey. As we already explained the survey can be composed by questions and usability tasks. Relatively to the questionnaire it can be presented to the participant different types of questions depending on what we want to ask. These questions can be open question, multiple choice (like rating time task), multiple selection, etc. The way that the questionnaire is presented is also variable, in other words we can have multiple questions in one single page or we can do it in multiple pages. For example we can want to do pre and post task questionnaires.

Then we have the Usability task view, for the usability task, where a first page will be presented to the user with the task that s/he will have to perform, an interactive help is also available for the user to understand how the interface tool works. In the usability task itself the user has to perform the task asked and can always review the task description on the top of the page if needed. He can always quit the task if he can't finish it successfully. Additionally Webnographer tool supports different kind of Usability tests like one click test that consists in asking to the user to try to perform a single action (like click a button for example). Other different type of test can be a fully interactive test that can consist in a complex task were the user can navigate on the website and finish the task when he thinks that is done. In the end of the task we can just finish, when the user click to finish, or if needed we can ask the user to give a small answer like a price or a number for example. Relatively to the data recorded during the test, a part from the questionnaire answers, it also saves data during the usability tasks. This data can be clicks on buttons or input boxes, text inserted in text boxes, scroll on page, webpages visited, time spent on a task or on each page, Ajax calls, among others.

Finally from the researcher perspective it is also an interesting tool. In first place it allows an easy setup of the survey, since the tool gives to the researcher templates for all the possible questions or functionalities that can be setup on the survey. Also it allows to delete and add new questions or usability tasks to the survey without having to redo the test. Also all the setup, questions or usability tests, are defined on the same place without having to do separated setups. Relatively to the usability tests it also gives us tools that make the analysis of the test data much easier. Finally another functionality that is not only good but also very useful (considering that we intend to change the visual appearance of the page) is that it allows us to easily show to the participant a modified page of the client. In other words Webnographer tool allows the researcher to change what the participant will see on the test without having to change the real client webpage (only through Webnographer, i.e. since this is not changing the client code the normal users will not be affected and we still can test changes on the client page without a big effort, this functionality was designed and developed by James Page.). In the end we will need to get the pre-processed data that Webnographer tool gives to us to do the analysis and develop the conclusions of the test.

In conclusion we consider this tool a very good solution to perform our study based on the functionalities offered and our needs. First, it is easy to use from the participant perspective, since he has no need to setup anything to perform the usability test. Second, the setup from the research perspective is also easy, due to the available tools. Additionally the data analysis is also made easier by the Webnographer tool with the pre-processed data. Finally, the functionality that allows the changes on the interface shown to the participants, allows us to perform the changes that we need to apply much easier to execute.

3.2 RESEARCH METHODS

With the method to test our hypothesis prepared we needed to study some of the basics from Webnographer methods to understand and apply them in order to prepare and evaluate the usability tests and also understand all the process (that as we told on section 1.4 we cannot detail due to confidentiality). In this section we describe Bayesian Statistics. Here we explain the two statistical methods that can be used, Bayesian and Frequentist (the most commonly applied) and we do a comparison between their advantages and disadvantages. These method was adapted from the original versions and implemented in Webnographer by James Page and Sabrina Mach.

3.2.1 Statistics and Usability Results Analysis

As known we have different methods to analyse the data collected during the usability tests. The two main methods are Descriptive Statistics and Inference Statistics. Descriptive statistics is used to describe the data collected and to get some preliminary conclusions.

However to do a proper analysis and be able to generalize the results with a good degree of certainty we need to use inference statistics. This analysis could be done with frequentist statistics that is the most used statistical method to do inference statistics, like t-student test or chi-square for example. Another alternative is Bayesian analysis, like Wagenmakers [23] and Masson [24] explained, to do this kind of analysis.

In this section we explain the advantages and disadvantages of each method, Bayesian and Frequentist and we explain why Bayesian Test (method used in Webnographer and implemented by James Page and Sabrina Mach) is better to evaluate our results from usability testing.

Starting with the frequentist statistics, like we said before is the most known method. The major advantages of this method are that it provides a systematic approach to wide range of statistical methods and do not required additional specification beyond that of the probabilistic representation of the data-generating process [24],[35]. A key problem in principle in frequentist formulations is that of ensuring that the long-run used in calibration is relevant to the analysis of the specific data being analyzed [23]. Another issue in applying the ideas is that technically exact solutions are available only for a limited class of situations. Usually, approximations have to be used based on asymptotic analysis [23].

Relatively to Bayesian we identified three main advantages. First is the sample sizes. The size that Bayesian require to have reliable results is small, comparing with frequentist [35]. The other advantage of this approach is that the hypothesis is only based on collected data. In other words the probability of our hypothesis being correct is calculated only with the data collected [23],[24],[35]. For last Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions. However with this method if we want to test large amounts of data the calculations are computationally heavy [24].

So, after comparing the advantages and disadvantages of each method we can say that Bayesian method is a good method comparing with frequentist. The main reasons were that it requires smaller samples to have reliable results and that inference includes uncertainty in the probability model. Relatively to uncertainty we consider that it is easier to understand, because with this method we are quantifying a difference instead of validating a difference, like in frequentist. For example, if we consider the values in Table 1 and apply a N-1 Chi-Square Test, we will get a result of 5.2. Now, if we calculate a p-value for this result with a chi-square distribution we will get a p-value of 0.02, which validates the hypothesis that Flat is less usable than skeuomorphism.

	Clicked	Didn't Clicked	Success Rate
22. Flat	12	62	11%
22. Non Flat	18	55	24%

Table 1 – Number of clicks on the two conditions

However If we apply the Bayesian method to the same table what we will get is for a static value that represents the percentage of improvement how much confident we are, in other words based on our results how much is the probability of this improvement being true. So if we apply the calculations for 1% of improvement we can say that we are 97% sure that this improvement is true as Masson explains [24].

So by comparing the two methods, frequentist and Bayesian, we consider that Bayesian is a good method. And one of the main reasons is that it allows equivalent levels of reliability as frequentist using less users. Additionally, since we will use Bayesian on our study to perform the results analysis (like it is already done by Webnographer) we will briefly explain more details about how the evaluation is done.

Basically we have two separated steps for the evaluation: one where we check the probability of our results being correct, in other words how reliable they are; and the second step is test our hypothesis, i.e. pick the results for the two separated conditions and compare then to check the probability of our hypothesis (Flat less usable than skeuomorphic) being correct.

To perform this analysis we applied an implementation done by Matthew Leitch[8] and that is used in Webnographer. For a better understanding of how we applied the Bayesian statistics we will use an example of a button being clicked in two different interfaces (flat and skeuomorphic).

So, for the step one as we said we check how much is the probability of our results being right. For that we calculate the probability density for each of the two cases base on the following formula:

[8] The implementation can be checked in this website http://www.workinginuncertainty.co.uk/conj_beta.shtml

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Where x is the success rate that we are checking, α is the successful interactions and β the unsuccessful interactions and B is:

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$$

So then what we do is generate a curve by applying the formula iteratively from a success rate of 0 to 100 percent. The number of iterations depends in how detailed we want the results, for our solution we applied in an interval of 5%.

So for the values present on Table 1 we got the following graph:

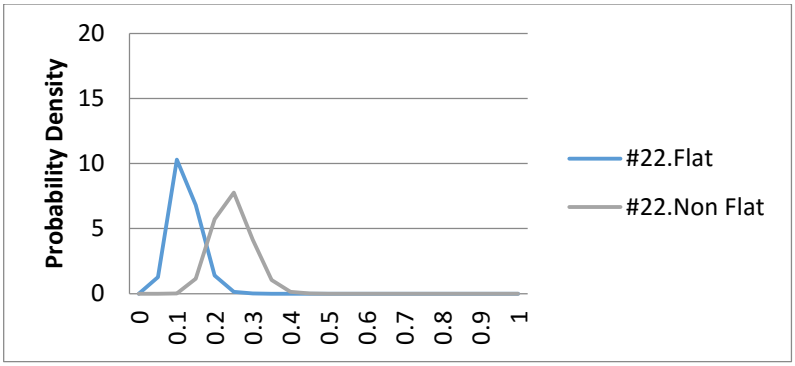

Figure 9 – Graphic for the probability of success rate results being correct

The results are more reliable as higher and narrow is the curve. Also the overlap between the two curves is the probability of our hypothesis being wrong. However here starts the step two. To exactly know this probability we will apply another calculation that will tell us how much is the probability of our hypothesis being true. For that we apply the implementation from Matthew Leitch to test how much is the probability of our hypothesis being right for percentage of difference defined for us. For our example we defined a difference of 1% and the result is: We are 97% confident that Interaction 2 is 1% better than interaction 1.

Additionally we have on Table 2, from Raftery [27], mentioned in Wagenmakers [23] works, a relation between the level of evidence and the probability that we calculate. So for this situation we have a Strong evidence of improvement.

Evidence	Weak	Positive	Strong	Very Strong
P(Hypothesis|Data)	50% - 75%	75% - 95%	95% - 99%	> 99%

Table 2 – Interpretation of Bayes Probability in terms of evidence

3.3 SUMMARY

In this chapter we explained the methods that we used to implement our solution. In the first section we described how we proceed to verify our hypothesis, like what we are comparing, Flat and Skeuomorphic design, how we are testing it and finally the tool we are using to implement our solution with remote usability testing Webnographer.

On the second section we described the basic concepts used to apply the solution that we are proposing. We did an explanation about the method implemented by Sabrina Mach and James Page used in Webnographer to analyse the results, the Bayesian Test, which is not the conventional method to do the analysis but that we checked being better to demonstrate our results.

4 CASE STUDY - SIMPLETAX

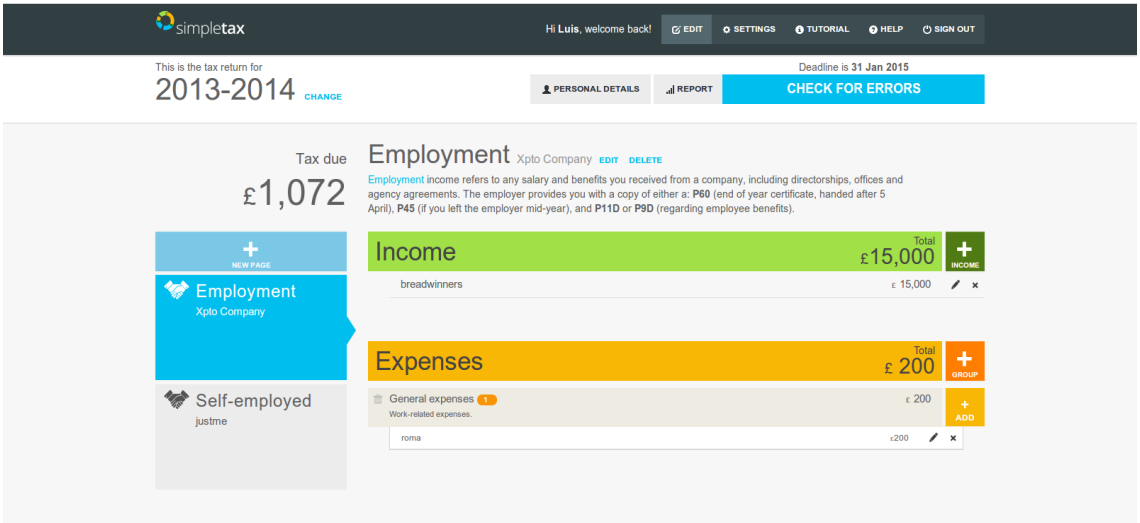

Figure 10 – Main Page of Simpletax Tool

The case study was done using Simpletax, as we mentioned before this company was a client and project planned and performed by Webnographer, by that reason we would be able to use the application and the usability results on our case study. Their tool is a webapp where the main functionality is the tax return submission to HMRC[9]. The goal of Simpletax is give to the user an easy way of filling their tax return without the complexity that they have on the official tool. Additionally we think that this application is a very good case study, because it should be accessible for all kind of users.

For this study in addition to the two style variations (Flat Design and Skeuomorphism) we also added another variable. The second variable was between the design used at the time on the website and an alternative design suggested by the owners of the tool.

On this chapter we will explain the preparation done to create the usability test. We will also show the results and draw conclusions from the results that we got from the users performances. Additionally we will explain the test preparation and results analysis where necessary.

[9] HMRC is Her Majesty´s Revenue Collection. A non-ministerial department of the UK Government that is responsible for taxes collection

4.1 TEST PREPARATION

Survey

On the survey development apart from the normal demographic questions as age or gender, it is identified the online experience of the user. This is done by asking which tasks they usually perform online and if they already submitted tax returns online. This survey (developed and performed by Sabrina Mach and Webnographer) was done for evaluating the tool for the company Simpletax and the results were made available to us to prove our hypothesis.

Usability Test

Another component of the research was the usability test, designed by Sabrina based on the client's needs. This one was composed by the task and a quick post task survey for the user to classify his/her satisfaction with the Simpletax tool.

As a first step we have the task we performed. This task has two subtasks: First task is to **signup** and the second task is to **fill a tax return**. To perform this task we give to the user a task description (before and during the task on the top of the screen) with all the needed details that can be checked bellow:

Use Simpletax to fill a 2013-14 tax return.

Please use the following details:

Email: sam.smith@gmail.com
Birth date: 01/10/1980
Unique taxpayer reference: 4325648151
Trading name: Creative designs
Business type: Sculptor
Address: 49 Featherstone Street LONDON EC1Y 8SY
Income: £15,500 of income in 2013-14 fiscal year.
This income results from Invoices for £8000, £3500, £4000;
Receipts for Purchases: Adobe Photoshop license for £329.
Additional information: The business is not registered for VAT;
Exempt of class 4 National insurance contributions;
Has not been approved for class 4 national insurance contribution deferment;
Does not have a balance sheet; Losses from previous years £0

Then we did an initial evaluation of Simpletax Tool (using Webnographer methods) where we could identify some issues that were present on the tool. The main findings for possible issues were:

1 **Step 9 – Click Personal Details Button**

The main problem here is the color of the button. First the color is gray which can cause two different issues. One of them is that the gray is the standard color to disabled buttons, due to that reason the user can maybe ignore the button thinking that is not an available action. The other reason is the gray being so light that is hard to see on the interface due to the contrast with the background color.

Also the position can be a problem too. The reason is that almost all of the editable details are on the bottom layer. So would be expected that this functionality was also there, or maybe asked in a next stage.

2 **Step 13 – Click Edit Link For Tax Payer Details**

The issue with this interaction is mainly the lack of visibility. First it's a link without any affordance apart from the blue color (that is the convention color for links). Also it uses a small font size if compared to the other elements on the page. The choice for the position was not the better also, considering all the interface. For example if this was placed on the layer that is showing to the user which tax category are we submitting, maybe it would be more natural and easy to see to the user.

3 **Step 18 – Click Add Income**

Relatively to this interaction the possible issue that we identified is the lack of contrast. In other words the color of the button is very similar to the color used on the title bar. This makes hard to the user identify the button in the end of the bar.

Additionally this button doesn't have visual feedback when we put the mouse over it. This is something that makes it even harder to realize.

4 **Step 22 – Click Add Expense Group**

In this step we found the same problem as is step 18. However on this button we have visual feedback when the mouse goes over the button. It is not enough still, but it helps when the user is inspecting the page trying to find this functionality and eventually get the visual feedback after placing the mouse over the button.

5 **Step 25 – Click Add Expense**

On this step the contrast is not an issue comparing to the ones already mentioned. However the labeling of this button is not clear. The function is add a new expense, but the labeling is only "ADD" which can be confusing to the user.

After the usability test was performed a post task questionnaire (ASQ – After Scenario Questionnaire) from James Lewis [34], that is normally asked in the end of a task on Webnographer tool and it was applied by

Sabrina Mach and James Page. The goal of this questionnaire was to understand how satisfied the user was after performing the task.

The post task questionnaire was composed for the following questions:

- Do you feel you were successful in completing the task?
- How would you describe how difficult or easy it was to complete this task? (rated between 1 – very easy and 5 – very hard)
- How satisfied are you with using this application to complete this task? (rated between 1 – very unsatisfied and 5 – very satisfied)
- How would you rate the amount of time it took to complete this task? (rated between 1 – very slow and 5 – very fast)
- What worked well on this task? (open question)
- What didn't work well and needs improving? (open question)

With this questions we could understand how much satisfied the users were with the tool and also we were able to get interesting comments on the last two questions that helped us to understand the reasons for some problems.

Design Variations Developed

To setup the different conditions to perform the tests it was required to do some changes on the platform both structural and stylish.

First we had to change the flat style application to an application with skeuomorphism. To do this changes we performed CSS changes on the web application. The changes were mainly effects on buttons as gradients, bevels, shadows, etc. and also create the animations for the button clicked or fix the button pressed when they are toggle buttons. Another change was adding the underline on the links that on flat design are just blue without underlines.

Additionally we did not have to change the structure of the website between current flat and current skeuomorphism. The reason was to try to change the perception of the user on the website without changing the "structural usability of the website". In other words the only variable changing between the two versions (Flat and Skeuomorphism) was the style as can be seen on Figure 11 and Figure 12.

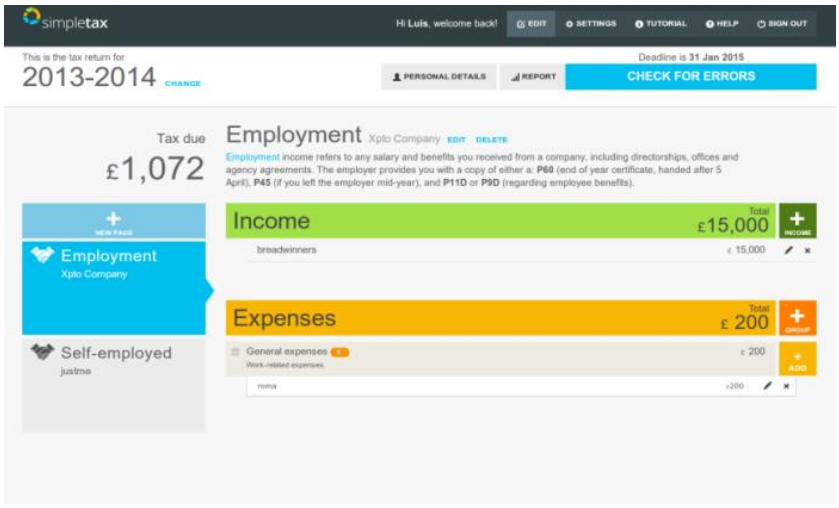

Figure 11 – Simpletax Dashboard with Flat Style

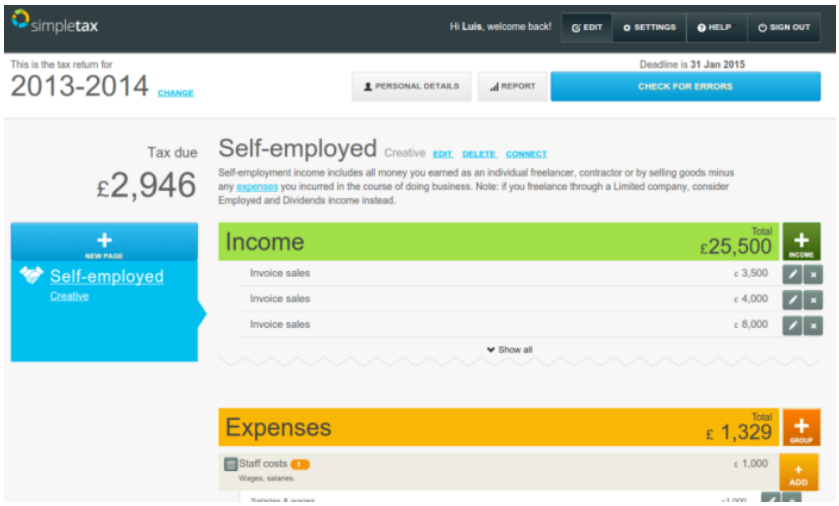

Figure 12 – Simpletax Dashboard with Skeuomorphism

The other variable changing on this case study was a variation of structure suggested by the owners of the tool. This change had as main goal create a flow on the task that the user need to perform to fill the tax return. For example on the current design the user perform all the task on the same window with multiple modals and also need to find the right element to get the respective modal. With the changes applied for the new design not only the number of popups shown to the user was reduced, as a sequence of steps that need to be performed to achieve the final goal were defined.

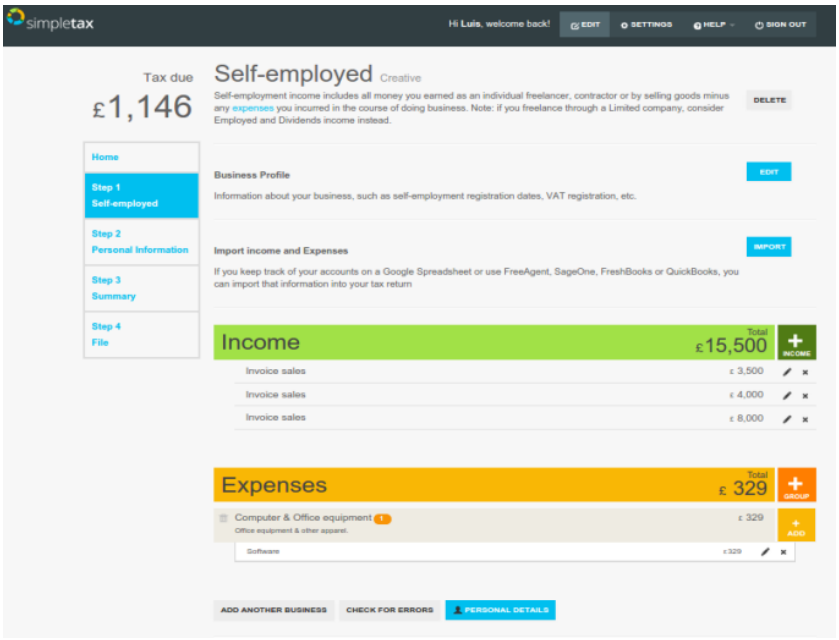

Figure 13 – Simpletax Dashboard with Flat Style and New Structure

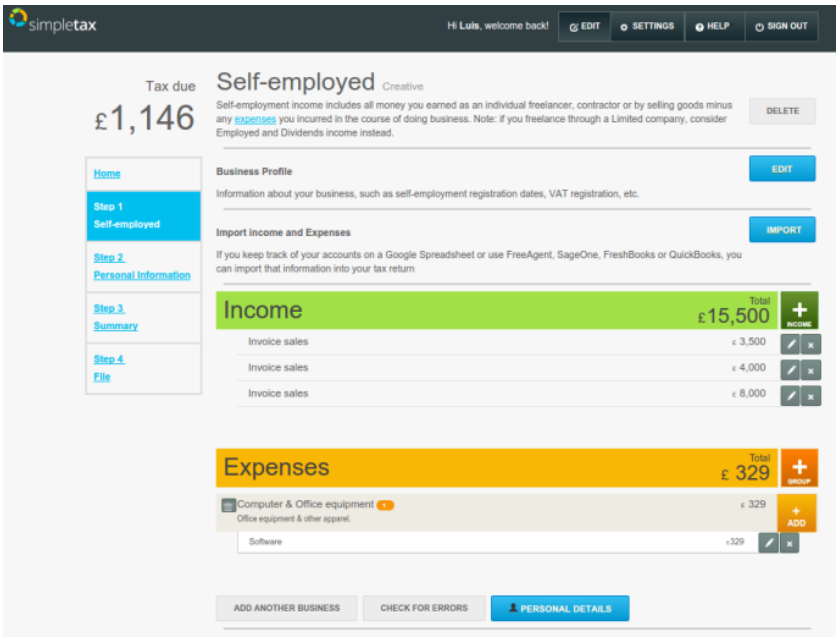

Figure 14 – Simpletax Dashboard with Skeuomorphism and New Structure

To apply this changes a script in JavaScript and using jQuery library was developed to change, delete or add elements on Simpletax page (this changes were visible only through Webnographer tool and not actually made on the client) to setup the new design prototype as can be seen on Figure 13 and Figure 14.

As we explained before all this changes are done on the fly (functionality available on Webnographer tool) and only can be visible through Webnographer tool and in consequence only for who as access to the test, since

the client page and code is not being effectively changed. In other words all the CSS and JavaScript code needed to apply the style and structural changes mentioned before are done through Webnographer using the method explained before on section 3.1.3.

User Recruitment

For this test was defined by Webnographer a goal between 70 and 80 users for each iteration (four variations of the interface). Also due to the application domain (tax submission for UK), we would need that the participants were UK residents. In order to achieve this two goals (number of users and demographic restrictions) a panel agency was used to do the participant recruitment. However, due to the high rate of users giving up or not performing the test truthfully, on the last iteration we only reach 24 participants. As a consequence the test is not reliable enough for some of the results. However for some particular interactions the difference is so big that even with this small sample we can see the improvement.

4.2 RESULTS ANALYSIS

On the analysis we show some general results and identify some conclusions that are general to all the different variations of the application.

Then we split the analysis in two different topics. First we will analyse the Current Design with the two styles variations. After that we will also compare the differences between the New Design with the two styles. Finally we discuss all the results that we got from the four variations and compare them to come up with our conclusions about the influence of the styles when we have different structures.

4.2.1 General Results

On the Table 3 we can see the number of participants in each test and also the number of interactions. The number of interactions is the same as could be expected since on new design the only change is the flow, which means that the user still need to perform the same steps on the task.

	Current Design Flat	Current Design Non Flat	New Design Flat	New Design Non Flat
Number of Participants	**73**	**78**	**73**	**24**
Number of Interactions	**30**	**30**	**30**	**30**

Table 3 – Number of Users per Iteration

Interactions List	
#1 Insert First Name	#16 Insert Address
#2 Insert Last Name	#17 Save Source Income Details
#3 Insert Email	#18 Add Income
#4 Insert Password	#19 Select Income category - SALES
#5 Click "Get Started" to submit the registration form	#20 Insert Amount
#6 Select Total Income	#21 Save Income
#7 Select source Self-Income	#22 Add Expense Group
#8 Click Continue	#23 Select Expense Group - Computer & Office equipment
#9 Click Personal Details	#24 Save Expense Group
#10 Insert Birthdate	#25 Add Expense
#11 Insert UTR Number	#26 Select Expense type - Software
#12 Save Personal Details	#27 Insert Amount
#13 Click Edit for "Self-employed"	#28 Save Expense
#14 Insert Trading Name	#29 Click "Check for Errors" ("Submit to HMRC" on new design)
#15 Insert Business Type	#30 Click Yes to Submit

Table 4 – List of Interactions for the task in both Current and New Designs

On Table 4 we have the list of all the steps identified to perform the task successfully for this test. The order present on the table is for the current version of the tool and all the interactions are performed in the same page. For the new design we have two small changes. First the steps for the personal details (between 9 and 12) are performed only after step 28. Second the name for the buttons on step 12 (step 28 on new design) and step 29 are different. On step 12 the button is now called “Continue to Summary” and on step 29 the button is called “Submit to HMRC”. Also the big change between this two versions is the creation of multiple pages for the form. Now the user has a page to fill the tax details (income, expenses, etc.), a second page for the personal details and a last page with the summary of the tax return where the user should do the submission.

Table 5 indicates the percentage of users that were able to achieve the end of the task. However this number does not mean the number of users that completed the task completely. This happens because the users can finish the task without perform all the steps since some of them are not required. For example in our task we have multiple income sources, however the user can only have one income source. The same think can be applied to the expenses, they don’t need to have all the expenses that we listed on the task. So they can finish the task but maybe succeeded.

Success Rate			
Path	**N Total**	**N Succeeded**	**% Succeeded**
Current Design Flat	73	12	16%
Current Design Non Flat	78	13	17%
New Design Flat	73	6	8%
New Design Non Flat	24	5	21%

Table 5 – Success rate of the task (including users that didn't complete not required steps)

4.2.2 Current Design Flat vs Current Design Non Flat

In this section we will focus only in compare between the two variations of the current design of the Simpletax tool, flat and skeuomorphism. We will show the detailed results for the two designs and compare them to show the improvement that we got in the user performance from flat to the version with skeuomorphism.

	Current Design Flat	Current Design Non Flat				
#Interactions	Total of Successful Interactions	Total of Successful interactions	A[2] Route Pass Rate	B[3] Route Pass Rate	Evidence Of Improvement (5%)[1]	Evidence Of Improvement (1%)[1]
#1	74	79	99%	99%	1%	24%
#2	74	79	99%	99%	1%	24%
#3	74	79	99%	99%	2%	25%
#4	74	78	92%	94%	21%	58%
#5	74	79	99%	99%	1%	25%
#6	73	78	90%	91%	18%	58%
#7	73	78	82%	87%	50%	82%
#8	74	78	99%	92%	0%	1%
#9	73	72	44%	53%	69%	87%
#10	32	38	94%	95%	23%	70%
#11	32	38	84%	84%	27%	58%
#12	32	38	78%	84%	52%	80%
#13	73	72	29%	29%	25%	54%
#14	21	21	81%	86%	50%	78%
#15	22	21	95%	95%	15%	41%
#16	21	21	95%	95%	18%	70%
#17	21	21	81%	86%	50%	77%
#18	73	72	16%	26%	76%	94%
#19	12	19	42%	26%	13%	29%
#20	7	11	86%	91%	45%	85%

#21	7	11	86%	82%	26%	65%
#22	73	72	12%	24%	83%	97%
#23	10	17	90%	76%	10%	15%
#24	10	17	90%	76%	9%	15%
#25	10	13	90%	92%	34%	53%
#26	10	12	90%	92%	33%	51%
#27	10	13	90%	92%	38%	54%
#28	10	13	90%	92%	35%	54%
#29	73	72	32%	43%	80%	93%
#30	23	31	52%	42%	14%	29%

Table 6 – Summary of the results for the Current Design in both Styles

[1] This value represents in probability how much confident we are that the non-flat is at least 5% (or 1%) better than the flat version.

[2] The A route is the test with the current interface of Simpletax with Flat design.

[3] The B route is the test with the current interface of Simpletax with Skeuomorphism.

On Table 6 we have a digest of the results that we got for Simpletax with current design. In this table we resume the number of successful interactions per design variation, the success rate of each one and also the evidence of improvement from flat to skeuomorphic in each step of the task. Evidence, as we explained before on section 3.2, is the probability that our hypothesis (Skeuomorphism is more usable than Flat) is correct. Then this probability can be translate in weak, positive, strong and very strong as can be seen on Table 2. For our analysis we focus on the improvement of 1%.

After analyze and compare our results we found 6 steps where the users got an evidence of improvement of at least 80%. The steps are the following:

- Step 9 – Click Personal Details (evidence of 1% improvement 87% probable)
- Step 12 – Save Personal Details (evidence of 1% improvement 80% probable)
- Step 18 – Add Income (evidence of 1% improvement 94% probable)
- Step 20 – Insert Amount (evidence of 1% improvement 85% probable)
- Step 22 – Add Expense Group (evidence of 1% improvement 97% probable)
- Step 29 – Check for Errors (evidence of 1% improvement 93% probable)

Conclusions

Step 9 – Click Personal Details

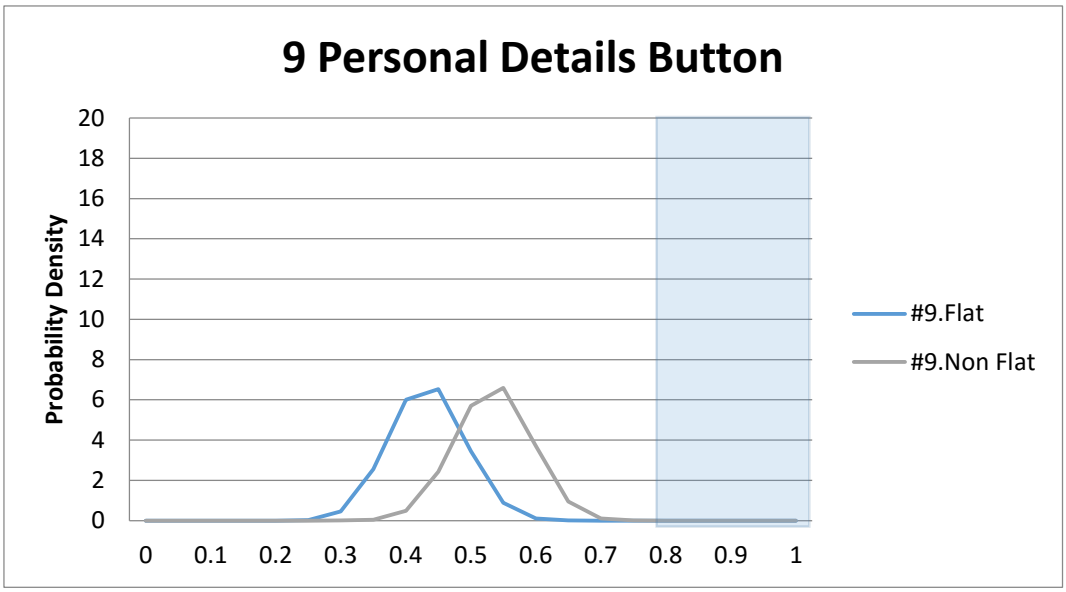

Figure 15 - Bayesian Test results for step 9

On this step the results show us a positive evidence of improvement for a difference of 1% (probability of 87%). So as we can conclude based on the test data the change from Flat to Skeuomorphism allowed a tendency of improvement on the user performance. We think that this happened due to the gradient and bevels added to the button that led to the user click the button more naturally. Because even if the color is still not a proper choice for this button, just the 3D appearance is the enough to create a contrast with the background and make the element more visible as we can check on Figure 16 and Figure 17

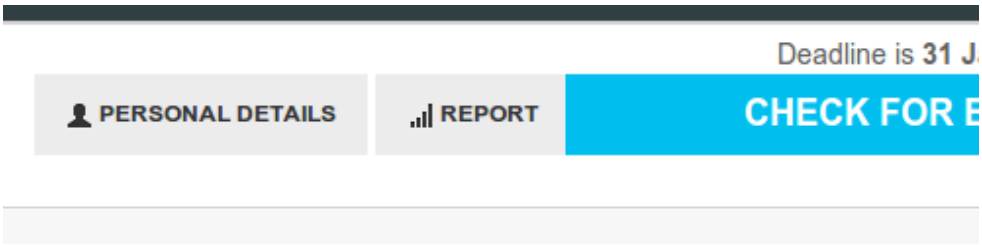

Figure 16 – Personal Details Button with Flat Design

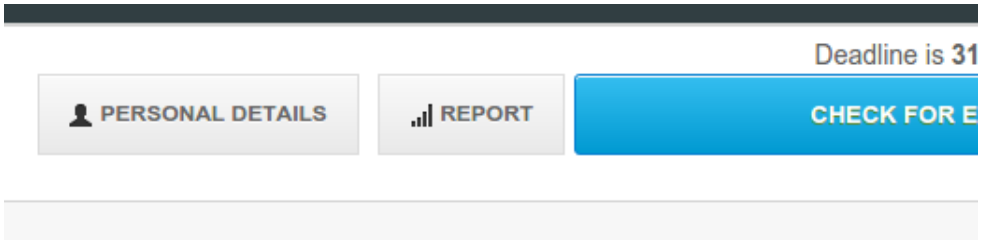

Figure 17 – Personal Details Button with Skeuomorphism

Step 12 – Save Personal Details

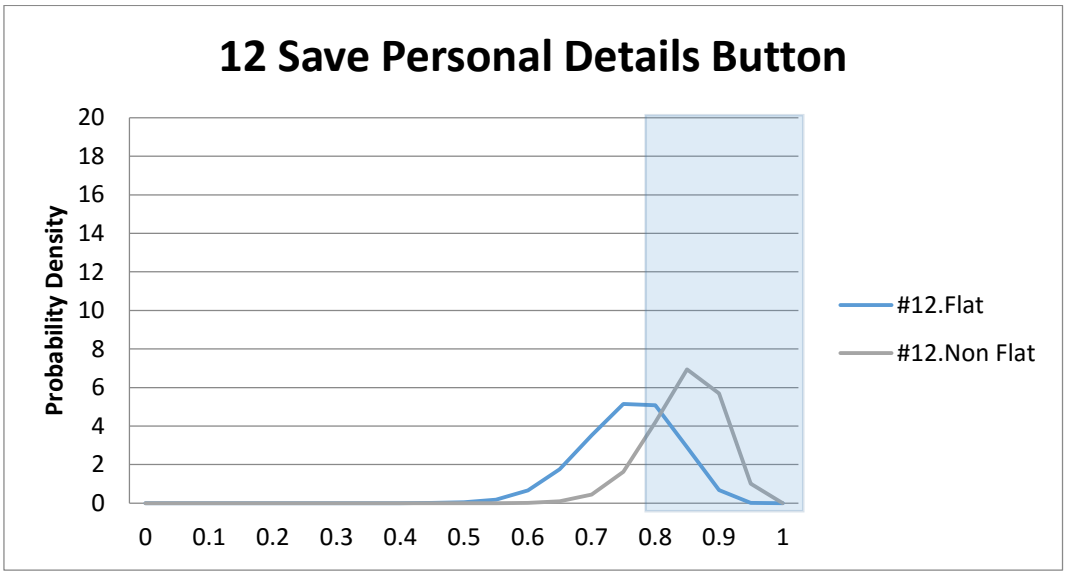

Figure 18 - Bayesian Test results for step 12

This step also seems to be better with Skeuomorphism compared to the flat version, however this is weaker than the last step. What we think being the main reason for the increased performance in this step was the button location. As we can see on Figure 19 the button is blue which we think that creates a good affordance to the user to click the button. However the form that the user needs to fill is quite long which can distract the user from the button. So we think that just for adding the gradients and bevels to the button was again a factor to catch the user attention and help him completing the step with success by clicking the button.

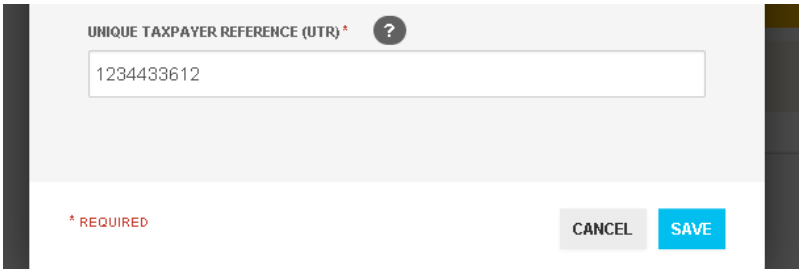

Figure 19 – Personal Details Popup with Flat Design

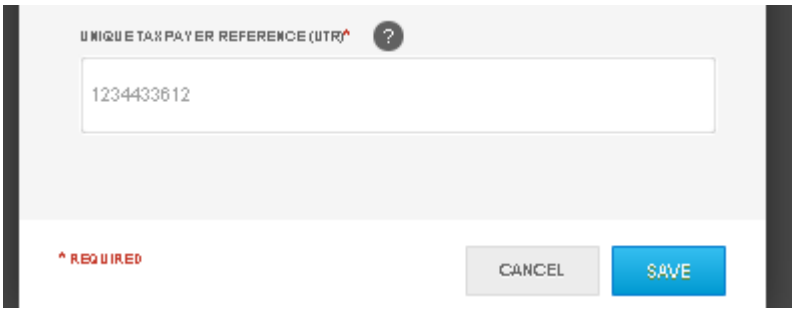

Figure 20 – Personal Details Popup with Skeuomorphism

Step 18 – Add Income

This step consisted in click the "Add Income" button to create a new entry on the table for an income source and the respective value.

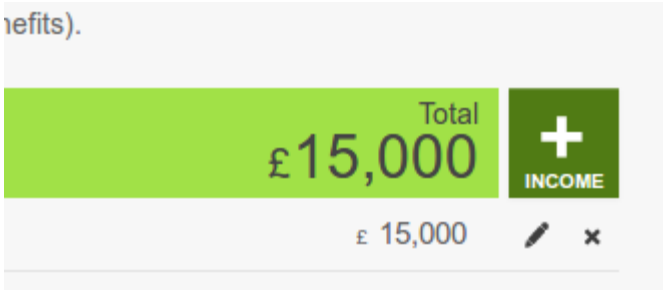

Figure 21 – Add income button with Flat Design

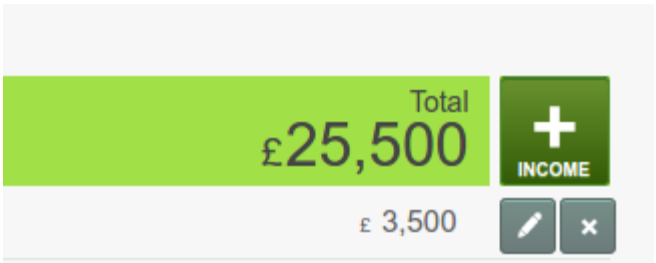

Figure 22 – Add income button with Skeuomorphism

Looking to the Figure 21 we can see that the contrast of the button could cause problems to identify it due to the color used compared with the title bar color. After applying the change we could verify this hypothesis. After analyze the results we got an evidence of improvement with a probability of 94%. Probably after adjust the colors used in both, button and title bar, we could get an improvement on performance even better.

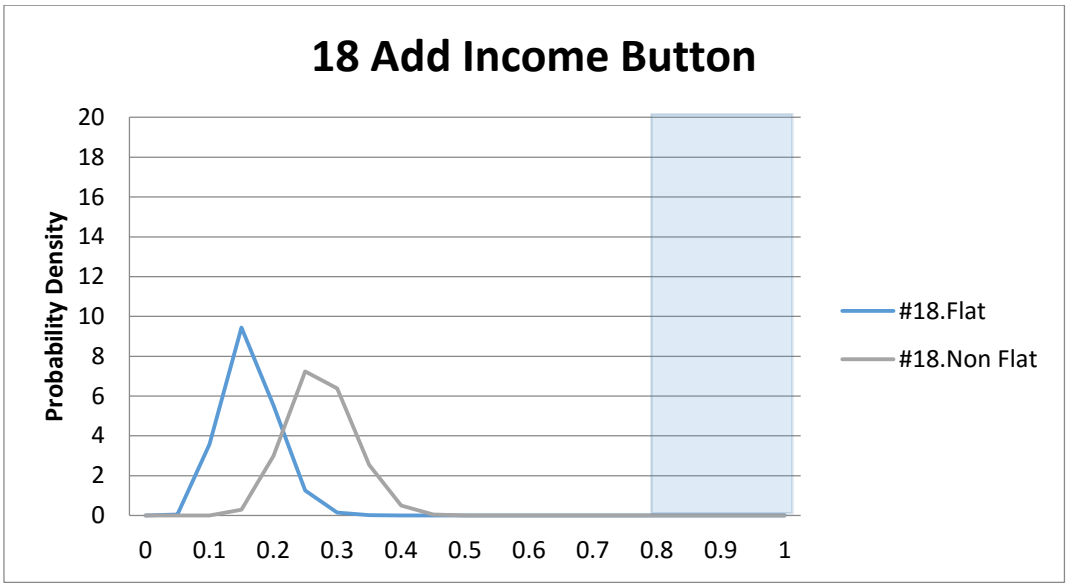

Figure 23 - Bayesian Test results for step 18

Step 20 – Insert Amount

In this step the goal of the user was select a text box and insert a value of an income for a work done to another entity. In other words an invoice value.

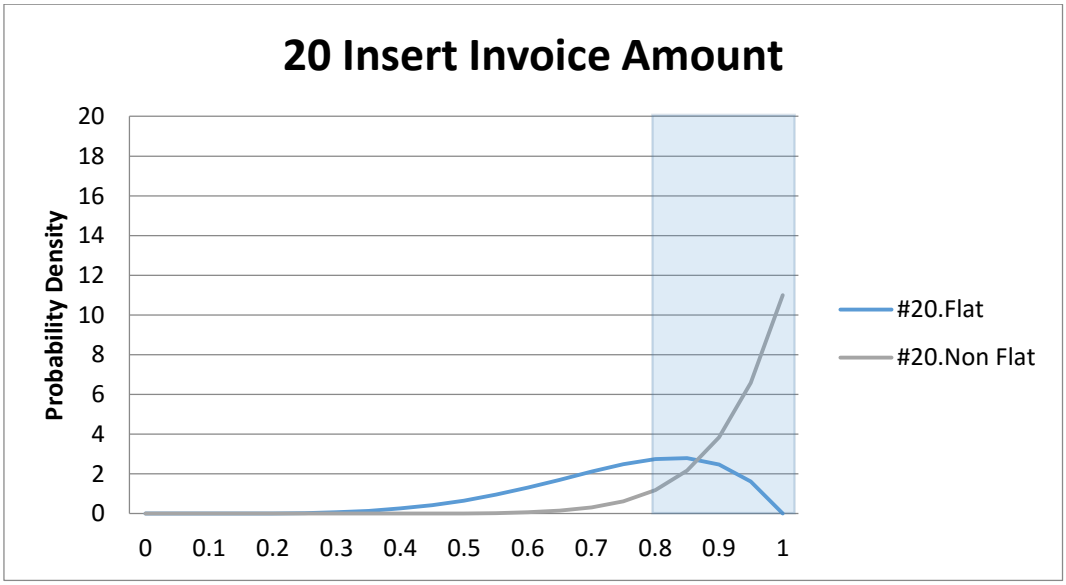

Figure 24 - Bayesian Test results for step 9

When we compare the flat version of the tool with the skeuomorphic version we could found a positive evidence (a probability of 85%) of improvement. A reason for this could be the shadow and the bevels added to

the textbox. However since we have any issue with a significant evidence in the other forms needed for the task we are not sure if this is really an improvement caused by Skeuomorphism.

Step 22 – Add Expense Group

Relatively to this step the explanation is the same that we detailed on the step 18. Basically the button to add the expense group is hard to be seen due to the title bar having a very similar color compared with the button. For this reason after adding the skeuomorphism tends to have an improvement compared to the flat version.

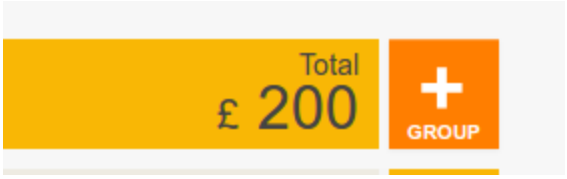

Figure 25 – Add expense group button with Flat Design

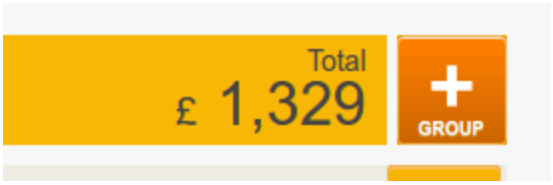

Figure 26 – Add expense group button with skeuomorphism

If we look to the level of evidence we have 97% chance of improvement which is a strong evidence according to the Bayesian method.

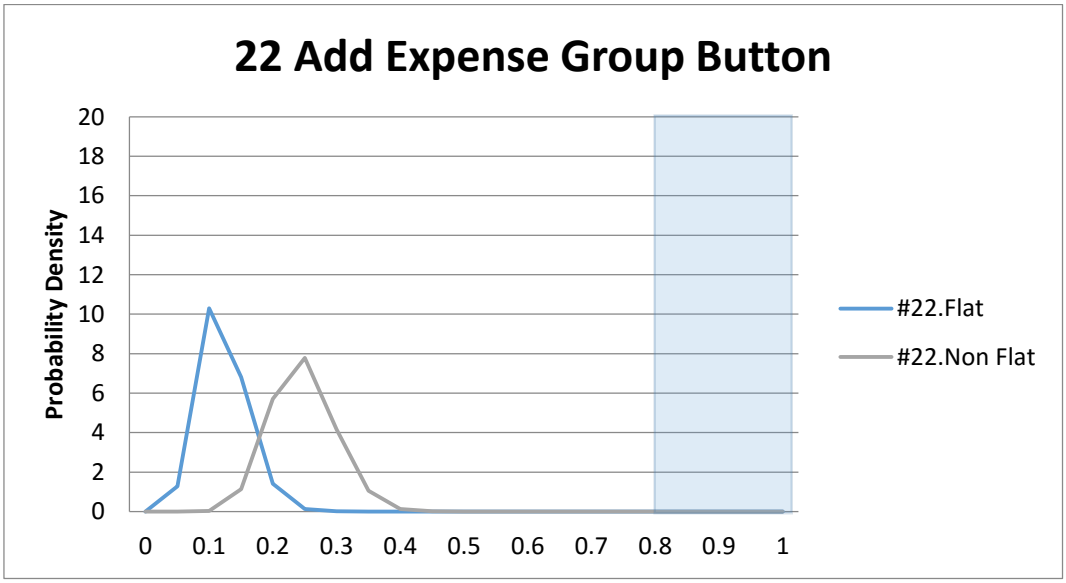

Figure 27 - Bayesian Test results for step 22

Step 29 – Check For Errors

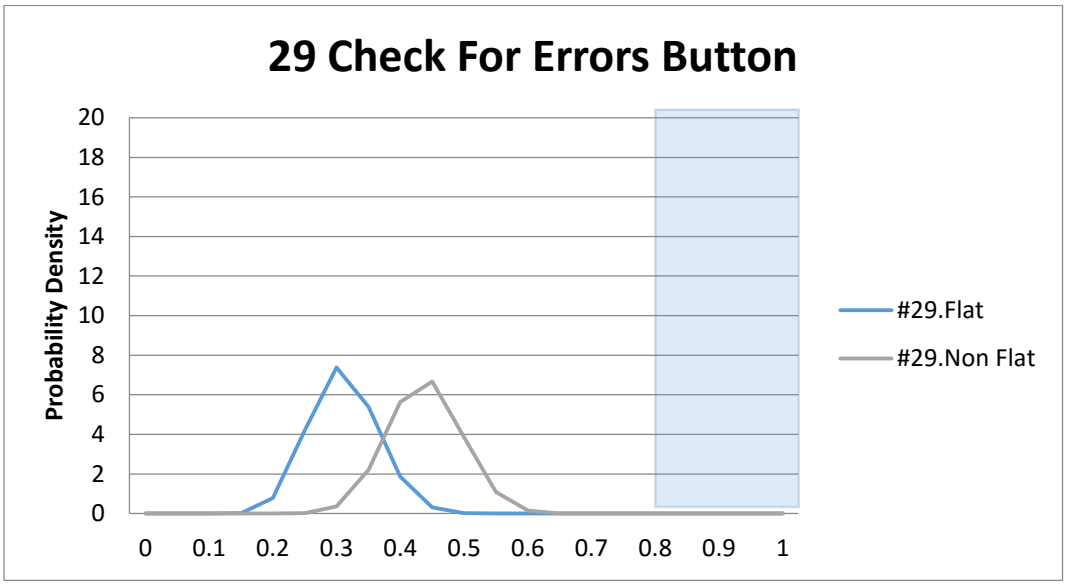

Figure 28 - Bayesian Test results for step 29

For the "Check for Errors" button the issue is also the labeling that is not clear about his functionality. However looking to the results that we got with the Bayesian analysis we got a chance of improvement of 93% that is a positive evidence (and very close to the 95% that is the value to be considered a strong evidence).

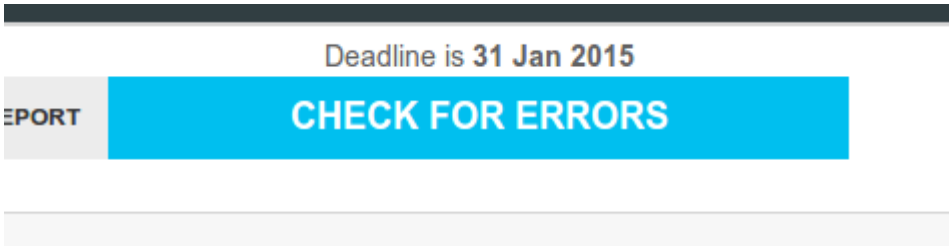

Figure 29 – Check for Errors button with Flat Design

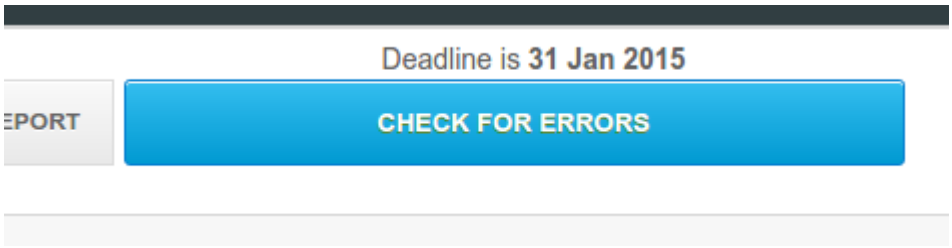

Figure 30 – Check for Errors button with Skeuomorphism

We believe that the reason for this improvement was the look and feel that the button got with our changes between the flat design and the skeuomorphism. In other words, however the labeling is still a problem since the

button looks really a button will catch the user attention. For this reason the user will feel compelled to click the button.

Post-Questionnaire results

After the task a satisfaction survey was done (the ASQ from James Lewis [34]) to check how the users felt after using the tool. This results are already calculated in Webnographer tool and this function it was design and developed by James Page and Sabrina Mach. Surprisingly even having a slightly improvement on the success rate finishing the task (but not successfully, in other words they reach the end of the task but they didn't complete all the steps that we defined), the users were a little less satisfied with the Skeuomorphic design than the flat design. One reason for this could be the strange appearance of the tool since the design was developed to match the flat design and not skeuomorphic. For this the final appearance of the tool with skeuomorphism looked strange. (The satisfaction score can be checked on Table 8)

ASQ Questionnaire				
Design Variation		**How would you describe how difficult or easy it was to complete this task?**	**How satisfied are you with using this application to complete this task?**	**How would you rate the amount of time it took to complete this task?**
Current Design Flat	1	15%	25%	16%
	2	33%	19%	23%
	3	21%	21%	25%
	4	18%	23%	21%
	5	14%	12%	15%
Current Design Non Flat	1	14%	21%	23%
	2	32%	28%	21%
	3	35%	24%	32%
	4	14%	23%	15%
	5	5%	4%	9%

Table 7 – Participants Answers After Scenario Questionnaire (the rate goes from 1 to 5 being 1 the worst rate and 5 the better)

Satisfaction score			
Design Variation	**Average satisfaction rating**		
Current Design Flat	1	20	27%
	2	20	27%
	3	15	20%
	4	14	19%
	5	6	8%
Current Design Non Flat	1	21	27%
	2	23	29%
	3	23	29%
	4	9	12%
	5	2	3%
	Score	Number of Participants	Percentage of Participants

Table 8 – Satisfaction Rates for the task in the current design being 1-low and 5-high

4.2.3 New Design Flat vs New Design Non Flat

As we did on last section we will focus our analysis in comparing the two variations between them, New Design Flat and New Design Non Flat. For that we will show the results that we got with our test and see the improvements that we were able to get between the two design variations. In this second test as we already mentioned, due to the high drop rate with the test participants we got only 24 participants for the second variation. This is the cause for a lack of results reliability. However the difference in some steps is so big that even with this low participant rate on the second test we are able to see the improvement.

	New Design Flat	New Design Non Flat				
#Interactions	Total of Successful Interactions	Total of Successful interactions	A Route Pass Rate[2]	B Route Pass Rate[3]	Evidence Of Improvement (5%)[1]	Evidence Of Improvement (1%)[1]
#1	74	25	99%	96%	0%	12%
#2	74	25	99%	96%	1%	12%
#3	74	25	99%	96%	1%	12%
#4	73	25	92%	96%	45%	79%
#5	74	25	99%	96%	1%	12%
#6	73	24	95%	83%	1%	4%
#7	73	24	86%	88%	33%	56%
#8	73	24	99%	96%	1%	12%
#9	72	23	39%	65%	96%	99%
#10	28	15	89%	87%	23%	39%
#11	28	15	79%	80%	42%	54%
#12	28	15	79%	80%	42%	55%

#13	72	23	47%	48%	35%	54%
#14	34	11	85%	73%	10%	20%
#15	34	11	94%	82%	3%	12%
#16	34	11	85%	82%	29%	42%
#17	34	11	82%	73%	16%	27%
#18	72	23	24%	9%	1%	7%
#19	17	3	47%	67%	71%	86%
#20	13	3	92%	67%	10%	17%
#21	14	3	93%	67%	7%	15%
#22	72	23	13%	17%	46%	74%
#23	9	4	78%	50%	10%	19%
#24	9	4	89%	75%	16%	28%
#25	7	3	86%	67%	18%	28%
#26	7	3	86%	67%	15%	28%
#27	7	3	86%	67%	16%	27%
#28	7	3	86%	67%	16%	27%
#29	22	12	64%	83%	88%	91%
#30	14	10	43%	50%	53%	64%

Table 9 – Summary of the results for the Current Design in both Styles

[1] This value represents in probability how much confident we are that the non-flat is at least 5% (or 1%) better than the flat version.

[2] The A route is the test with the current interface of Simpletax with Flat design.

[3] The B route is the test with the current interface of Simpletax with Skeuomorphism.

On Table 9, presented before, we have a digest of the results with current design but for the New Design. In this table we resume the number of successful interactions per design variation, the success rate of each one and also the evidence of improvement from flat to skeuomorphic in each step of the task. Evidence, as we explained before on section 3.2, is the probability that our hypothesis (Skeuomorphism is more usable than Flat) is correct. Then this probability can be translate in weak, positive, strong and very strong as can be seen on Table 2. Also like on the first experiment we focus the analysis for the improvement of 1%.

After analyzing and comparing the results we found 3 steps where the users got an evidence of improvement of at least 80%. The steps are the following:

- Step 9 – Click Personal Details (evidence of improvement 99% probable)
- Step 19 – Add Income (evidence of improvement 86% probable)
- Step 29 – Check for Errors (evidence of improvement 91% probable)

Conclusions

Step 9 – Click Personal Details

In this step the user has to click the button “Personal Details” to continue the tax return submission. In this new design the button had two main changes: the color is now blue; and the position is now in the end of the page as can be seen on Figure 31.

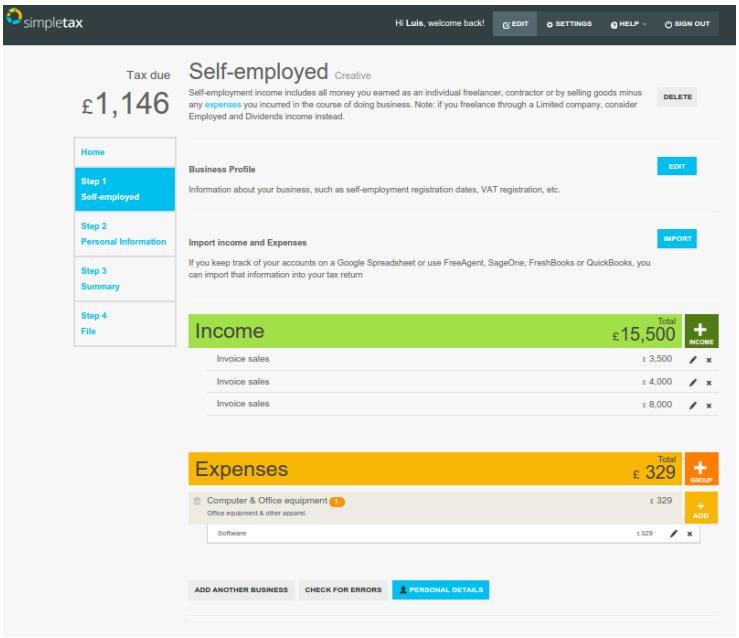

Figure 31 – Self-Employed Page for new flat design Simpletax

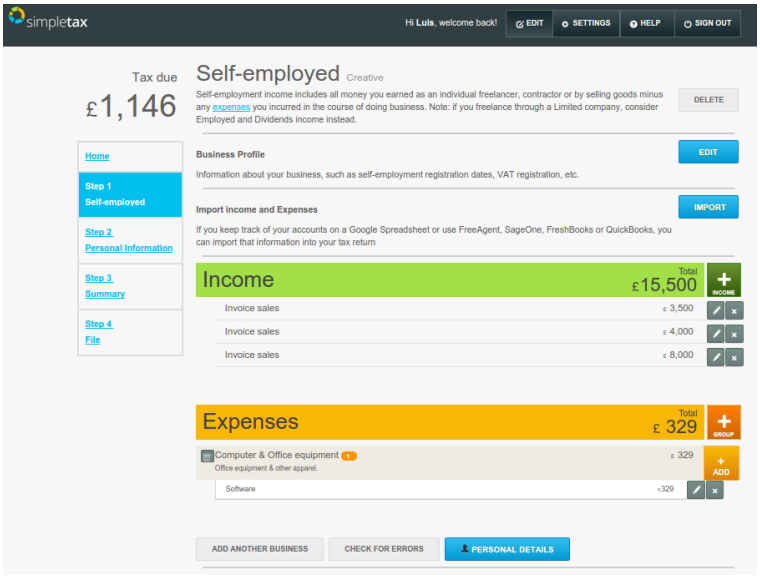

Figure 32 - Self-Employed Page for new non Flat design Simpletax

In our opinion and based on the 99% evidence of improvement (very strong evidence), we believe that the change from flat design to skeuomorphism was responsible for this result. In other words due to the button being positioned in the bottom of the page is not easily visible to the user. But since we change the button to

skeuomorphism we are adding the affordances that will catch the user attention and make him notice and click the button.

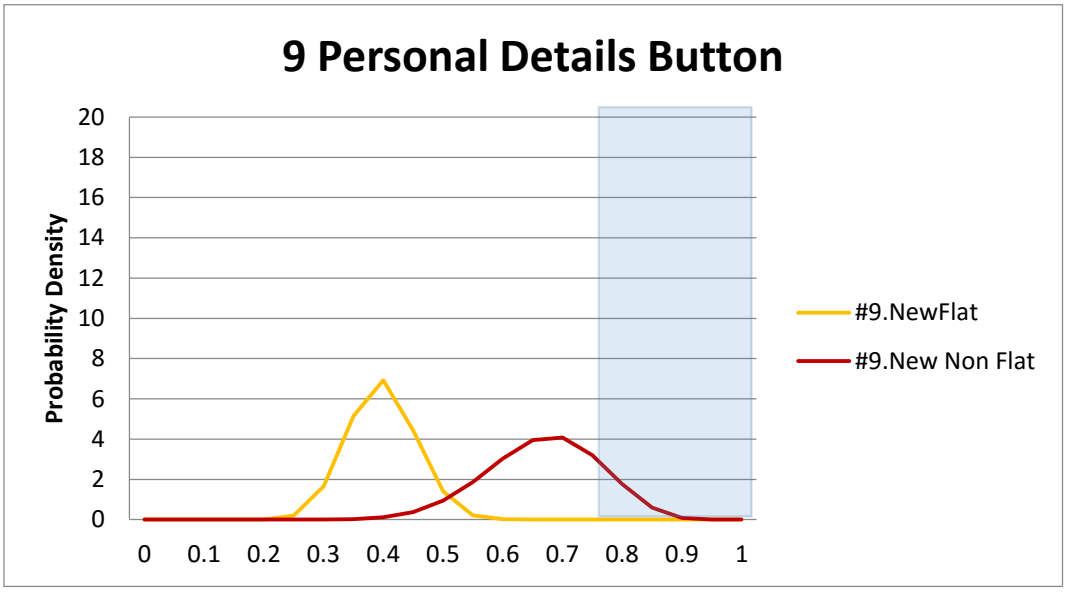

Figure 33 - Bayesian Test results for step 9

Step 19 – Select Income category – SALES

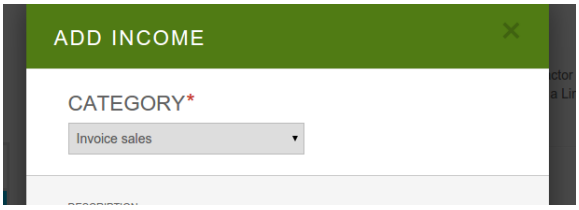

Figure 34 – Select Category Dropdown for new flat design Simpletax

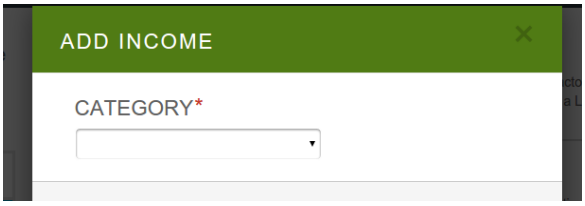

Figure 35 – Select Category Dropdown for new non flat design Simpletax

Relatively to this step the user needs to select the income category in a dropdown selection box. The only reason that can be causing this chance of improvement (86%) is the shadows and depth feeling that was added on the skeuomorphic variation as you can see on Figure 35. However since we have so few participants on non flat test and we didn't had the same issue in other similar steps we can't really tell that this is a real improvement.

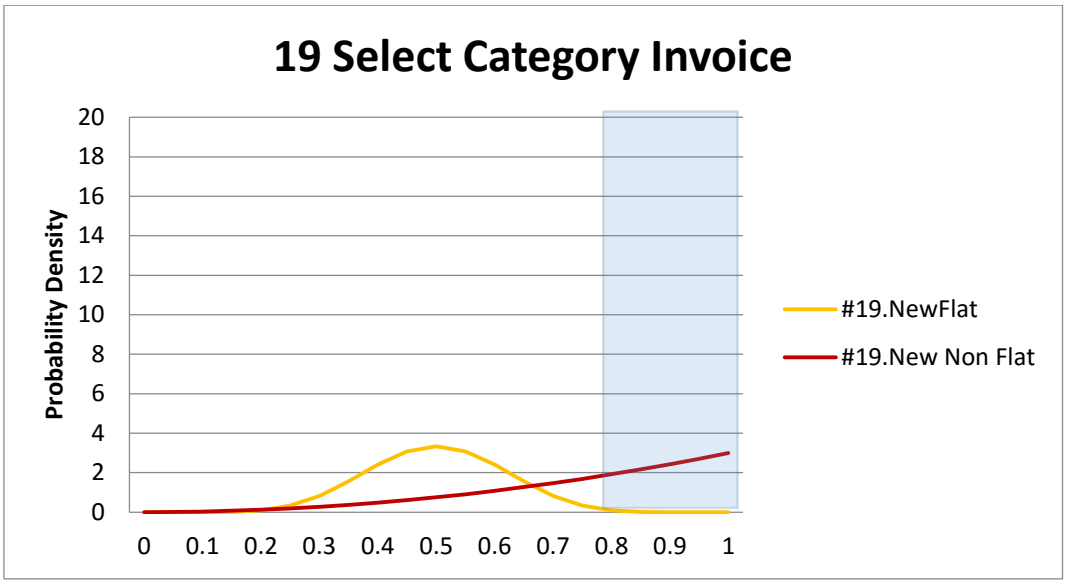

Figure 36 - Bayesian Test results for step 19

Step 29 – Click "Submit to HMRC"

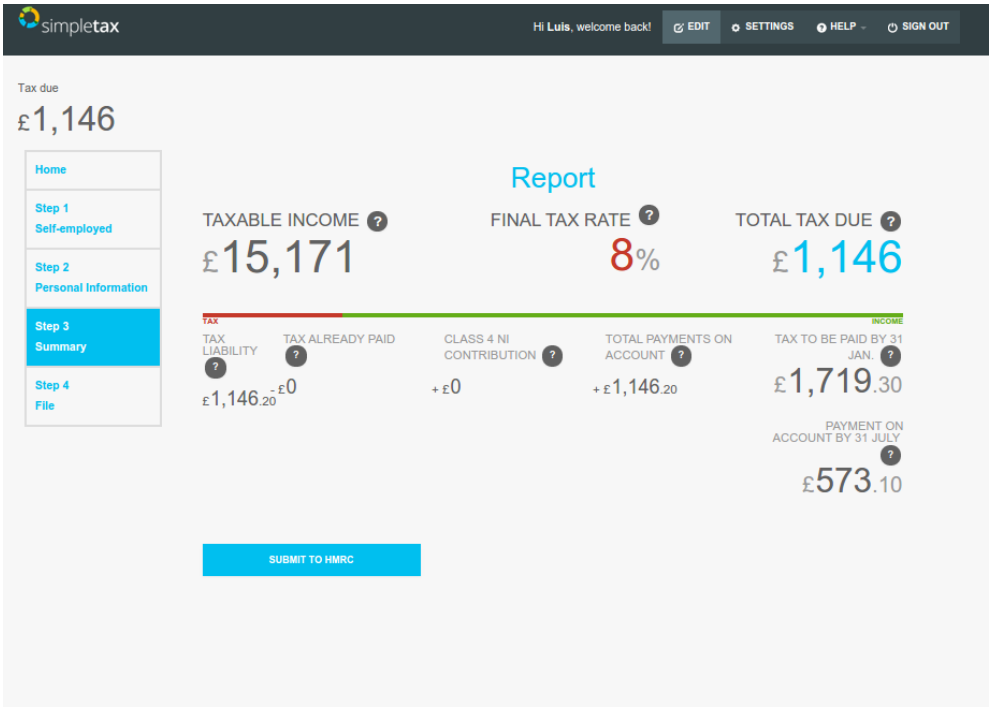

Figure 37 – Summary Report Page for new flat design Simpletax

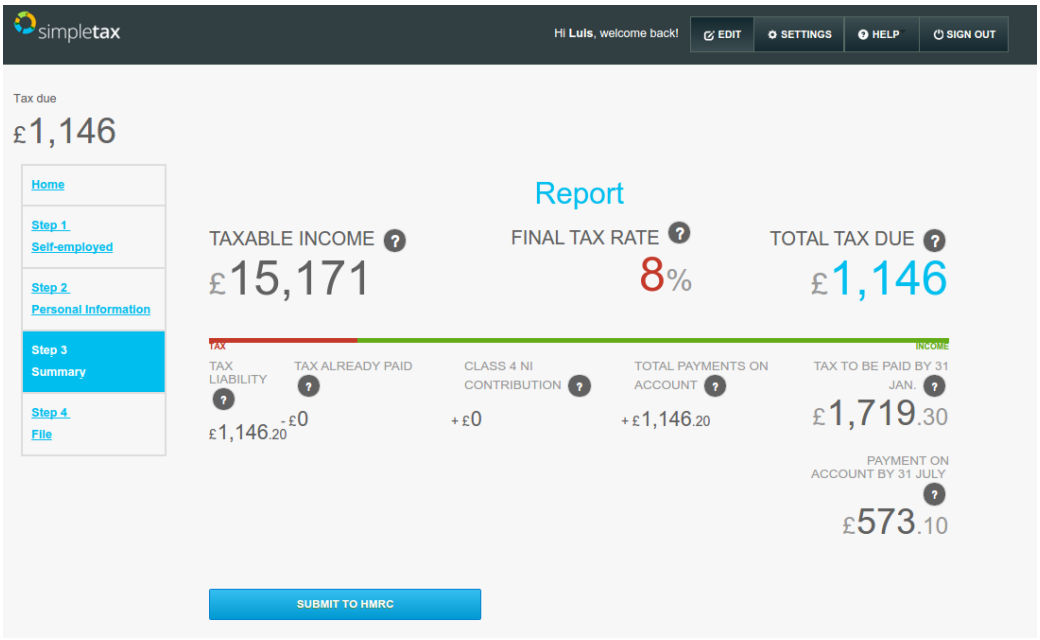

Figure 38 – Summary Report Page for new non flat design Simpletax

In this step the user has to click the button to finish the complete the submission to HMRC and complete the task. Similarly to the step 9 this button is also located in the bottom of the page which we think that is the reason for the evidence of improvement when we compare the Flat Design with Skeuomorphism. However this improvement is not as strong as step 9 (91% on step 29 compared with the 99% on step 9). In our opinion the reason for this difference, between the two steps, is the complexity of the page on step 9 that due to the length of the form that need to be filled, when the page for step 29 only as the summary for the tax return and the option to go back or submit. In other words the last page is so simple that is easy to the user understand even the flat design.

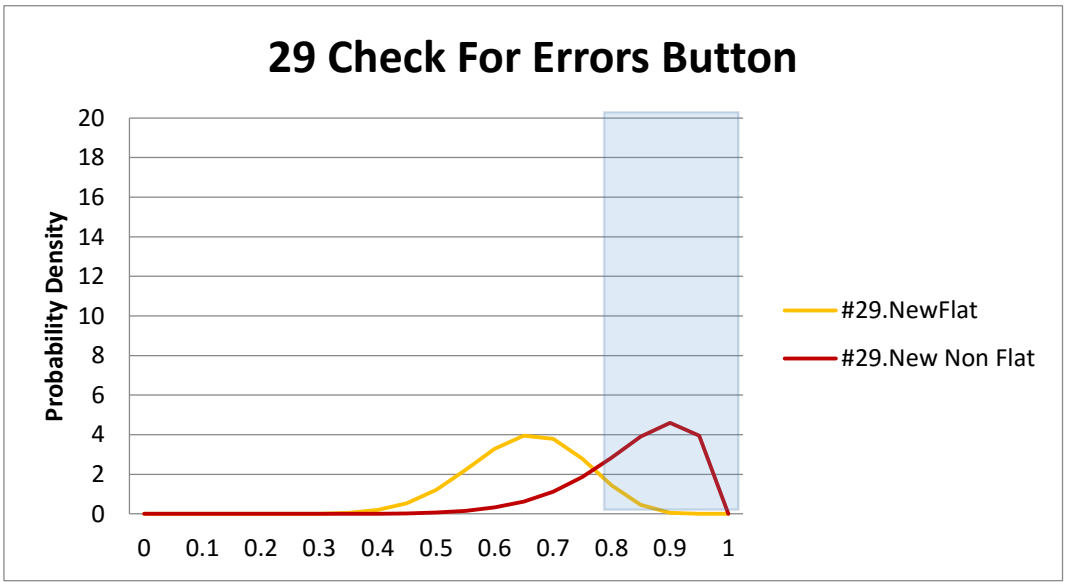

Figure 39 - Bayesian Test results for step 12

Post-Questionnaire results

As it is done on the first test a satisfaction survey was performed after the task (ASQ [34]) to check how the users felt after using the tool. By looking to the satisfaction score results we can see in one hand that the highest percentage of users is indifferent to the quality if the interface however we have a negative tendency. Relatively to the second variation we have two kind of users by having people really unsatisfied and people that are satisfied, however we still can see a negative tendency (The satisfaction score can be checked on Table 11)

ASQ Questionnaire				
Design Variation		**How would you describe how difficult or easy it was to complete this task?**	**How satisfied are you with using this application to complete this task?**	**How would you rate the amount of time it took to complete this task?**
New Design Flat	1	10%	14%	14%
	2	25%	30%	16%
	3	42%	23%	32%
	4	18%	26%	30%
	5	5%	7%	8%
New Design Non Flat	1	8%	21%	25%
	2	42%	25%	17%
	3	8%	8%	21%
	4	29%	33%	29%
	5	13%	13%	8%

Table 10 – Participants Answers After Scenario Questionnaire (the rate goes from 1 to 5 being 1 the worst rate and 5 the better)

Satisfaction score			
Design Variation	**Average satisfaction rating**		
New Design Flat	1	12	16%
	2	22	29%
	3	27	36%
	4	13	17%
	5	1	1%
New Design Non Flat	1	7	29%
	2	5	21%
	3	4	17%
	4	7	29%
	5	1	4%
	Score	Number of Participants	Percentage of Participants

Table 11 – Satisfaction Rates for the task in the new design being 1-low and 5-high

4.3 RESULTS DISCUSSION AND IMPLICATIONS

In this section we look into the results presented on the last section 4.2 and analyze them. In other words we will do a review in the comparison of the two styles variations on each interface design and then compare the results between them. In the end we will resume our main findings and our conclusions based in our results.

So looking to the general results we can observe that the main issues that we found are not in all the interactive elements. Instead what we have are problems in some specific steps of our task. And if we look to the current design test comparison. This does not mean that the other steps has no usability problems but that they are not relevant compared and does not seem to have a problem related to Flat Design as the steps that we highlighted.

Still looking only for the current design evaluation what we can observe is the users having problems with the interactions that need to be performed in complex interfaces (like click button to "Add income") or in elements that are hard to see because they are placed in the bottom of the interface (like the save button on the personal details popup on the current design of the tool on that time).

After looking to these results the pattern that we identified is that the users were having problems whit the elements that are "hidden" (not easily visible or having affordance to be interactive) in the interface due to the complexity. And when we added the skeuomorphism then the success rate increased on that elements. A good example to validate our conclusion is for example the "Add income" and "Add Expense Group" buttons on current design that both of them got an evidence of improvement around 95%. In contrast we have other parts of the Simpletax interface that didn't showed us evidence of improvement, for example the employment details form, even after adding elements that should improve the affordance. In this case we could observe that less than half of the users found the link to open the employment details popup. However if we check the success rates for people finishing the form filling with success we have a rate that is more than 80% (except for the last test that is the less reliable).

Comparing now with the results on the new design we can see that they are matching our conclusions for the first test if we consider the results on the evidence of improvement. Relatively to this test only two steps were improved by the changes that we made (step 9 and step 29). We also think that problematic parts of the interface like the panel where the user add the incomes and expenses, even without changing the appearance from flat to skeuomorphic, they got an improvement on the usability probably because of the flow generated by the interface change. In other words now on this screen all the interactive elements are located in the right side of the screen which make the user more focused on that side and helping him finding the elements. However like we said before, the last test is not the most reliable and for that reason we can't strongly validate our conclusions for the first test.

In conclusion the main finding that we got is when we have complex interfaces the skeuomorphism give to the user the affordances to distinguish the interactive elements in the interface.

However when we have simple interfaces, like forms for example, the actions that the user can do are so clear that the differences between skeuomorphism and flat are not relevant.

5 CONCLUSIONS AND FUTURE WORK

In this section we do a summary of the content of the work described on this document. We will also do a final review of how we developed and applied the solution proposed on chapter 3 and summarize the conclusions that we reached with our work. Finally we explain what can be improved on our work and future research that can be developed to validate the new kinds of Flat Design that are emerging.

5.1 DISSERTATION SUMMARY

In this dissertation the main objective was demonstrate that the style used on the computer interfaces as websites or applications can have influence on the user performance. In other words the usability can be affected by the styles that we use on the interface. In particular this dissertation focus on the negative influence that Flat Design can have in comparison with the design used until now and still used, Skeuomorphism.

To have a good understanding of the subject that we were working on during our work we did an investigation of the related work done until now related with the influence of the design and aesthetics on the usability of the interfaces. This related work is described on Chapter 2. First we researched about the concept of affordance and how important this concept is for the usability of an interface or object when the user is using it. Then we also describe some works that explain the relation between aesthetics and usability and how the aesthetics can influence the user performance while he is performing the task.

Since our work was focused on the usability testing we also wanted to prove that the method that we would use the remote usability testing non-moderated from Webnographer was a good solution compared to the normally used test methods. To do that we did a research about the usability test methods currently used to compare the existent test methods with the one that we would use from Webnographer. After comparing all the options we concluded that asynchronous remote usability method is a good solution and we explain why on section 2.3.2, additionally we compare Webnographer method with the approach used by Tullis et al [19].

Then on Chapter 3 we describe our proposed solution. To do that we explain our approach and some of the basics to apply research methods. On the used approach we present how we will validate our hypothesis to compare the usability between Flat and Skeuomorphism and the Webnographer tool that we will use to do the usability test. After that on research methods we explain some basics that we will need (according to the Webnographer tool) to build our usability test and evaluate the results.

Finally on Chapter 4 we explain in detail the work that was done to implement the usability test. On the section 4.1 we explain what was done to prepare the test. Then on section 4.2 we present and analyse the results that we got from the two interfaces evaluated (current design and new design of Simpletax tool) and we identify

the main issues found in this tests. Finally on section 4.3 we did the discussion of the results that we analysed before and we describe the main findings of our work.

5.2 CONCLUSIONS AND CONTRIBUTIONS

During this work, due to the research done relative to the context and related work, we were able to get a better understanding of the what is the affordances and how important they are to the usability of an interface or physical object. We could also learn about the influence that only the style or aesthetics used on the interface could have on the user performance using an interface. Also with the research about usability test methods we could understand better the options that we currently have and the advantages and disadvantages of each one. For example relatively to the method used, remote testing, this is a very interesting option nowadays, since with the globalization and the exposition that we have with the internet, is very important to do tests with users from different fields and cultures. The learnings gained about this method would not be possible without the opportunity of working at Webnographer with Sabrina Mach and James Page that developed their own method and tool to do remote evaluation.

Another interesting learning from this work, although not directly related to the main subject, was the statistical method used by Webnographer, and that we used to evaluate our results. The main advantage that we found with this method is that even with low rate of completions for the tests we can get a good level of certainty for the difference on the results. For example on the second test of the new design we had only 24 completions, however we were able to identify the improvement on 3 steps. However a disadvantage of this method is the computational power that we need to analyse big amounts of data

Relatively to our work we were able to verify the hypothesis that Flat design tends to be less usable than Skeuomorphism. However as we could understand with our second test this difference is relevant only with complex interfaces. In other words when we have interfaces that are relatively simple, like a form for example, it will be easier for the user to understand what he needs to do even with the flat design.

5.3 FUTURE WORK

After doing our work we could identify some work that can be developed to improve this study. We have two suggestions that would be interesting to apply. One would be test the new concepts of flat design, like the “Almost flat”. Another one would be replicate this test with tools from different fields.

Relatively to the first suggestion that is test the “Almost Flat”, what we think is that with this new concept that is basically to use flat but instead of remove all the style, we should maintain some components that maybe will be enough to give the affordance to the elements. One example of this concept is the Material design from Google. In their interfaces they are using shadows to give the notion of depth giving to the user the notion that

the element is a button and not a label. In our opinion this changes could maybe be the enough to improve the usability comparatively to the normal flat.

About the second suggestion for future work, testing interfaces from different fields, the objective is to validate that our conclusions are valid independently of the kind of interface that we are using. In other words that the flat design usability is not dependent of the field of the interface that is applied, like for example a medical tool. Actually on our work we started preparing a test with a tool from Prodsmart, that is a tool for factory management, but due to unexpected issues we were not able to complete this second test.

REFERENCES

[1] Anon, 2004. Here, there, anywhere. In Proceedings of the 5th conference on Information technology education - CITC5'04. New York, New York, USA: ACM Press, p. 132. Available at: http://dl.acm.org/citation.cfm?id=1029533.1029567

[2] Bargas-Avila, J.A. & Hornbæk, K., 2011. Old wine in new bottles or novel challenges. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, New York, USA: ACM Press, p. 2689. Available at: http://dl.acm.org/citation.cfm?id=1978942.1979336 [Accessed December 13, 2013].

[3] Ben-Bassat, T., Meyer, J. & Tractinsky, N., 2006. Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction*, 13(2), pp.210–234. Available at: http://dl.acm.org/citation.cfm?id=1165734.1165737 [Accessed November 11, 2013].

[4] De Vasconcelos, L.G. & Baldochi, L.A., 2012. Towards an automatic evaluation of web applications. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12*. New York, New York, USA: ACM Press, p. 709. Available at: http://dl.acm.org/citation.cfm?id=2245276.2245410 [Accessed July 25, 2013].

[5] Hartmann, J., Sutcliffe, A. & De Angeli, A., 2007. Investigating attractiveness in web user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07. New York, New York, USA: ACM Press, p. 387. Available at: http://dl.acm.org/citation.cfm?id=1240624.1240687 [Accessed December 12, 2013].

[6] Hassenzahl, M. & Monk, A., 2010. The Inference of Perceived Usability From Beauty. *Human-Computer Interaction*, 25(3), pp.235–260. Available at: http://www.tandfonline.com/doi/abs/10.1080/07370024.2010.500139 [Accessed October 22, 2013].

[7] Hassenzahl, M., 2004. The Interplay of Beauty, Goodness, and Usability in Interactive Products. Human-Computer Interaction, 19(4), pp.319–349. Available at: http://dl.acm.org/citation.cfm?id=1466559.1466561 [Accessed November 8, 2013].

[8] Jimenez, C. et al., 2012. Formal specification of usability heuristics. In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies - EAST '12*. New York, New York, USA: ACM Press, p. 55. Available at: http://dl.acm.org/citation.cfm?id=2372233.2372249 [Accessed July 26, 2013].

[9] Kurosu, M. & Kashimura, K., 1995. Apparent usability vs. inherent usability. In *Conference companion on Human factors in computing systems - CHI '95*. New York, New York, USA: ACM Press, pp. 292–293. Available at: http://dl.acm.org/citation.cfm?id=223355.223680 [Accessed December 13, 2013].

[10] Lavie, T. & Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies, 60(3), pp.269–298. Available at: http://www.sciencedirect.com/science/article/pii/S1071581903001642 [Accessed December 15, 2013].

[11] Lee, S. et al., 2010. Understanding user preferences based on usability and aesthetics before and after actual use. *Interacting with Computers*, 22(6), pp.530–543. Available at: http://www.sciencedirect.com/science/article/pii/S095354381000055X [Accessed October 22, 2013].

[12] Nielsen, J. & Molich, R. 1990. Heuristic evaluation of user interfaces. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* - CHI '90, Jane Carrasco Chew and John

Whiteside (Eds.). ACM, New York, NY, USA, 249-256. DOI=10.1145/97243.97281 http://doi.acm.org/10.1145/97243.97281

[13] Nielsen, J., 2000. Why you only need to test with 5 users. Available at: http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [Accessed December 5, 2013].

[14] Sonderegger, A. & Sauer, J., 2010. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, 41(3), pp.403–410. Available at: http://www.sciencedirect.com/science/article/pii/S0003687009001148 [Accessed October 10, 2013].

[15] Thüring, M. & Mahlke, S., 2007. Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, 42(4), pp.253–264. Available at: http://dx.doi.org/10.1080/00207590701396674.

[16] Tractinsky, N., 1997. Aesthetics and apparent usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97*. New York, New York, USA: ACM Press, pp. 115–122. Available at: http://dl.acm.org/citation.cfm?id=258549.258626 [Accessed December 13, 2013].

[17] Tractinsky, N., Katz, A.. & Ikar, D., 2000. What is beautiful is usable. *Interacting with Computers*, 13(2), pp.127–145. Available at: http://www.sciencedirect.com/science/article/pii/S095354380000031X [Accessed October 10, 2013].

[18] Tuch, A.N. et al., 2012. Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, 28(5), pp.1596–1607. Available at: http://www.sciencedirect.com/science/article/pii/S0747563212000908 [Accessed October 10, 2013].

[19] Tullis, T. et al., July 2002. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. *Usability Professionals Association Conference*. Available at: http://www.testapic.com/medias/RemoteVsLab.pdf

[20] Van Schaik, P. & Ling, J., 2003. The effect of link colour on information retrieval in educational intranet use. Computers in Human Behavior, 19(5), pp.553–564. Available at: http://www.sciencedirect.com/science/article/pii/S0747563203000049 [Accessed December 15, 2013].

[21] Van Schaik, P. & Ling, J., 2009. The role of context in perceptions of the aesthetics of web pages over time. International Journal of Human-Computer Studies, 67(1), pp.79–89. Available at: http://www.sciencedirect.com/science/article/pii/S1071581908001304 [Accessed December 15, 2013].

[22] An Essay towards solving a Problem in the Doctrine of Chances , communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. in the Philosophical Transactions of the Royal Society of London 53 (1763), 370–418

[23] Wagenmakers, E.-J., 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), pp.779–804. Available at: http://www.springerlink.com/index/10.3758/BF03194105 [Accessed November 20, 2014].

[24] Masson, M.E.J., 2011. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior research methods*, 43(3), pp.679–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21302025 [Accessed July 14, 2014].

[25] Gibson, J. J., 1977. The Theory of Affordances. In: Shaw, R. and Bransford, J. (eds) Perceiving, Acting and Knowing. Erlbaum, Hillsdale, NJ.

[26] Gibson, J. J., 1979. The Ecological Approach to Visual Perception. Boston: Iloughton Mifflin.

[27] Raftery, A., Bayesian Model Selection in Social Research (with Discussion by Andrew Gelman & Donald B. Rubin, and Robert M. Hauser, and a Rejoinder). Available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.198 [Accessed December 16, 2014]

[28] Norman, D. A., 1988. The Psychology of Everyday Things, Basic Books, New York.

[29] Norman, D.A., 1999. Affordance, conventions, and design. Interactions, 6(3), pp.38–43. Available at: http://dl.acm.org/ft_gateway.cfm?id=301168&type=html [Accessed December 9, 2014].

[30] Gaver, W.W., 1991. Technology affordances. In Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91. New York, New York, USA: ACM Press, pp. 79–84. Available at: http://dl.acm.org/citation.cfm?id=108844.108856 [Accessed December 17, 2014].

[31] Kaptelinin, V. & Nardi, B., 2012. Affordances in HCI. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12. New York, New York, USA: ACM Press, p. 967. Available at: http://dl.acm.org/citation.cfm?id=2207676.2208541 [Accessed December 17, 2014].

[32] Transactions of the Lancashire and Cheshire Antiquarian Society, Volume 7, 1890

[33] John Mullaly. 1998. IBM RealThings. In *CHI 98 Cconference Summary on Human Factors in Computing Systems* (CHI '98). ACM, New York, NY, USA, 13-14. DOI=10.1145/286498.286505 http://doi.acm.org/10.1145/286498.286505

[34] James R. Lewis. 1991. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. SIGCHI Bull. 23, 1 (January 1991), 78-81. DOI=10.1145/122672.122692 http://doi.acm.org/10.1145/122672.122692

[35] Meng-Yun Lin, 2013, Bayesian Statistics: technical report Nº2

[36] Phil Turner. 2005. Affordance as context. Interact. Comput. 17, 6 (December 2005), 787-800. DOI=10.1016/j.intcom.2005.04.003 http://dx.doi.org/10.1016/j.intcom.2005.04.003