TUx: Testing UX between Web Frameworks

Luis Carlos Alves
Henriques
Instituto Superior Técnico
Av. Prof. Doutor Cavaco Silva
2744-016 Porto Salvo
+351 214 233 200
luis.henriques@tecnico.ulisboa.pt

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.

General Terms

Performance, Design, Experimentation, Human Factors, Verification.

Keywords

Flat Design, Skeuomorphism, User Experience, Usability, Evaluation.

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 Windows 8 in 2012.

Since then Flat Design has become widely used in the newest websites, tools and applications, 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” by Transactions of the Lancashire and Cheshire Antiquarian Society [21]. In human computer interaction 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 style effects that were removed in the flat design to create this metaphors.

The concept that objects have a couple of characteristics that tell to the person or animal what they should do with them is old. The first person calling this characteristics as affordances was James Gibson [15]. 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 [17]. 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 [18] 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 [19] or Affordances in HCI from Kaptelinin et al. [20], 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 this paper we will present the work developed to validate our hypothesis and the conclusions that we reached in our study.

This Study was designed and developed at Webnographer 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.

2. Related Work

To get a better understanding of the context were we performed our study we did a research in two different perspectives. First we researched about design and usability where we could learn about affordances and their importance for usability and also the influence that aesthetics can have in the user performance. Second 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.

2.1 Design and Usability

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.

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 [15]. 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. 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 [17]. 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 [17]. Later in 1999 Donald Norman felt the need to clarify his work on The Psychology of Everyday things. This happened because people misunderstood affordances from Gibson [16], the real affordances, with Norman “affordances” that are actually perceived affordances (as he clarify on Affordance, Conventions, and Design) [18]. In his work Norman was talking about the reaction that the affordance will cause in the user even if is not real.

We had also an interesting work that tried to clarify and apply the concept of affordance (from Gibson) on the human computer interaction field. This work is called Technology Affordances and was developed by Gaver [19]. 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.

Recently a work from Kaptelinin and Nardi [20], 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.

In summary we could learn the importance of the affordances in the usability of the objects. By applying this concept we are able to make the user understand what he needs to do just by looking the interface without need of additional information. So since in flat design all the skeuomorphic details are being removed we think that this affordances are also being removed.

Related to the influence of aesthetics we have a few studies like Hartmann, J., Sutcliffe, A. & De Angeli, A. [3], which demonstrate no correlation between aesthetics and usability however we have much more studies that prove this correlation. In our work we summarize some of that works.

In the first study that we analyzed (Tractinsky, N., Katz, A.. & Ikar, D. [8]). This paper that is one of the first studies in this subject was good to understand the basis. This work was evaluating how the users would rate the usability of the system (an ATM machine layout) based on the aesthetics. They tested two conditions pre and post usage. After analyze they could prove that as they expected the user rated the interfaces with a worst aesthetics as less usable than the interfaces with better aesthetics. However they also found that after the actual usage the user tend to increase the usability rate for the less appealing interface. They think that this can be caused by the natural adaptation that the human as to something that is required to use. In the end they could prove that the system usability was affected by the aesthetics and not by the actual usability of the system.

The other work that we analyzed was Lee, S. et al. [5]. In this study they developed their work answering to the methodological limitations from the previous works by developing their work based in hypothesis divided in three parts: interaction before actual use, interaction after actual use and comparison of interactions before and after actual use. To apply the test the authors developed four systems with two variables aesthetics and usability. Besides the influence that perceived aesthetics has in perceived usability (already supported in the first study) we could learn that perceived usability also has influence on the perceived aesthetic, in other words they found that users tend to rate the less appealing interfaces as less usable. 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 analyzed the study performed by Tuch, A.N. et al. [9], in 2012. 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. As in the Lee study they also developed four interfaces with the aesthetics and usability variables. This study unlike the previous ones, demonstrates that perceived aesthetics does not affect perceived usability but in reality is perceived usability that affects perceived aesthetics. However they admit that their usability manipulation was stronger than the aesthetics manipulation which could had influenced the results.

Then on a study from Sonderegger, A. & Sauer, J. [7] 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. To perform the study they developed a test were the users would perform the task in two interfaces of mobile phones one with lower aesthetics and the other with high aesthetics. To measure the study the authors defined three categories: Perceived product attractiveness, Perceived usability and User performance. To assure the usability of the two mobile phone interfaces the prototypes were based in an already existent mobile phone. 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, a part of the influences of affordances and aesthetics in usability that was proved in the works discussed before, we could also identify two limitations that we consider be recurring in all works analyzed. The first one is the number of interfaces tested. In all of the studies the authors only test one interface. 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.

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, we do a comparison between the method used by Tullis, T. et al. [10] and Webnographer to show the main differences.

One of the most known usability test techniques is heuristic evaluation that was developed by Nielsen et al. [6]. 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 these problems was identified by Jiménez, C. et al. [4] 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. Basically the author say that the way that the heuristics are defined need to be clearer and standardized in order to allow people with less experience apply the evaluation without misunderstand the heuristics. However not only for this reason we think that this is not the best method to use in our study. Since the problem that we are studying is not only about the way the user percept the interface the expert can look to the interface not the same way as a normal user.

Then we analyzed a study that compared the lab testing with remote usability testing (Tullis, T. et al., July [10]). In this work the author do a comparative study to understand the advantages and disadvantages of each method. He also had to decide to compare the remote synchronous and asynchronous method. Basically the difference is that synchronous is done remotely but is still moderated like lab testing, which make the test very similar with lab testing. For that reason the author choose the asynchronous method. In the end after perform the tests the author could understand that the main advantages of the remote asynchronous test was the amount of users that can be tested without effort, the variety of users since they don't need to go to the lab and also the influence of the test environment. However this also have disadvantages being the main one the tester not being able to observe the user which makes him lose some information. However this information can be compensated by the post feedback from the user.

Based in the last paper we could prove that remote asynchronous method in general is a good solution to perform our study, since with this method we can test a variety of users bigger than with lab testing. In order to show the differences between the different remote evaluation methods, we analyze the three different types of remote tests founded during the research performed.

Relatively to the automatic evaluation De Vasconcelos, L.G. & Baldochi, L.A. [2]. This method is based in algorithms that analyze the interface trying to match possible usability problems. However we decided to discard this option from the beginning because this method focus on evaluating the structure of a website and our objective is the perception, which by comparing the two interfaces only changing the visual would probably result in the same level of usability between them. In other words the human factor would not be considered. By that reason we considered this a bad solution to our study.

Thus only remains for us to compare the synchronous method with the asynchronous. For that we analyzed the work from Anon [1]. According with his study the main advantage of the synchronous method is that we can replicate almost all of the techniques used in lab testing (like think aloud for example) with the improvement that we can test more diverse people wherever they are and the usability tests becomes cheaper. However the principal disadvantage against asynchronous method is that we can't achieve more users than in the lab tests because this tests such as lab tests, need to be moderated. Another advantage of the asynchronous method is that the data is more reliable because of the number and diversity of users. However this method have some problems being one of them the inexistence of direct observation of the user by the moderator of the test. But like Tullis, T. et al. [10] identified we can obtain good conclusions from the questions given by 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" [10].

Based in all of these findings even considering that the method described by Tullis [10] is very different of 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. And if we compare the two methods (the method used by Tullis [10] 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, because the setup is done only by the research with no need of setup from the participant side or client side. And another big difference is that the data that is collected by Webnographer tool (like time, interactions, etc) that allow us not only base our analysis on user feedback but also on the user behavior. However Webnographer also allow us record the user feedback through the questionnaire.

3. Proposed Solution

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 selected a real application from Simpletax[1] (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). Then by comparing the results from the two different usability conditions we can conclude if the results are similar between the two interfaces, relatively to the difference between flat and skeuomorphic. If they are similar then we can conclude that the style has influence on the usability. If not we will need to analyze the results to check how and why they are not different.

3.1.3 Webnographer Tool

After in this section we will briefly explain how Webnographer works and the main steps performed in a Usability test. The Webnographer 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. 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

[1] www.gosimpletax.com

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 current status of the interface being tested and what are the potential problems of this one. 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.

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.

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. 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. Relatively to the usability tests it is also give 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, this functionality was designed and developed by James Page.).

In conclusion we consider this tool a very good solution to perform our study based on the functionalities offered and our needs. First how easy is to use the tool from the participant perspective, since he has no need to setup anything to perform the usability test. Then 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. And finally the functionality that allow the changes on the interface shown to the participant, 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 analyze 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 [12] and Masson [22] 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 [22],[23]. 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 [12]. 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 [12].

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 [23]. 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 [12],[22],[23]. 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 [22].

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.

To perform this analysis we applied an implementation done by Matthew Leitch[2] 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). In the end we can put this

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

probability calculated before in terms of evidence according that indicate how strong our hypothesis is.

4. Case Study

The case study was done using Simpletax. This tool is a webapp where the main functionality is the tax return submission to HMRC[3]. We think that this application is a very good case study. The main reason is that the tool should be accessible for all kind of users.

On this topic we explain all the preparation done to create the usability test. We will also show the results and get the conclusions of the results that we got from the users performances.

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 to the user classify his/her satisfaction with the 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. 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:

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.

Step 13 – Click Edit Link For Tax Payer Details

The issue with this interaction is mainly lack of visibility. First it's a link without any affordance apart of 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.

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.

Step 22 – Click Add Expense Group

In this step we found the same problem as we found on 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 for functionality.

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.

Step 29 – Click "Check For Errors" (Submit Tax Button)

As the step before, the problem with this button is mainly the labeling. But in this one the problem is even worst. Basically the main function of this button is the tax return submission. In other words after fill everything that is required for the tax return we need to do our tax return submission. For that we should click this "Check for Errors" button to complete the process. It's true that this button also do the errors checking, however probably a button labeled as "Submit", that also perform an error checking, would be more clear than a "Check for Errors" that also do the submission.

After the usability test we perform a post task questionnaire (ASQ – After Scenario Questionnaire) from James Lewis [24], that is normally asked in the end of a task on Webnographer tool with the following questions:

- 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?

The goal of this questionnaire was to understand how satisfied the user is after performing the task. Additionally this questionnaire allow us to calculate a satisfaction score that will show the percentage of users for each score between 1 and 5, being 1 the lowest and 5 the higher.

Design Variations Developed

To setup the different conditions to perform the tests 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. Additionally we also decided not change the structure of the website between current

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

flat and current skeuomorphism. The result can be seen on Figure 1 and Figure 2.

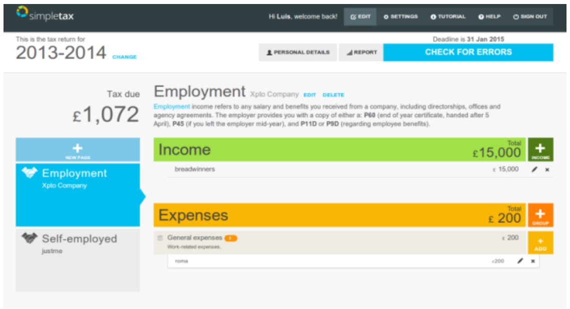


Figure 1 - Simpletax Dashboard with Flat Style

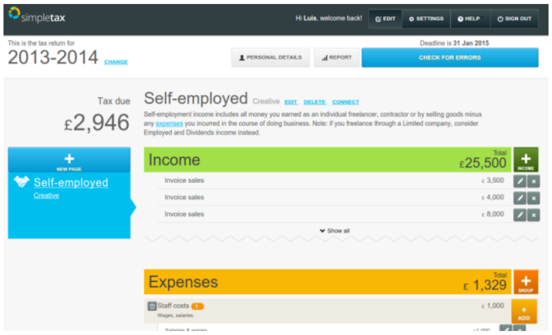


Figure 2 - Simpletax Dashboard with Skeuomorphism

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 we tried to improve the app usability.

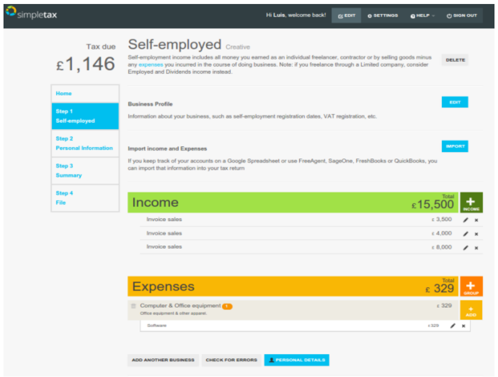


Figure 3 - Simpletax Dashboard with Flat Style and New Structure

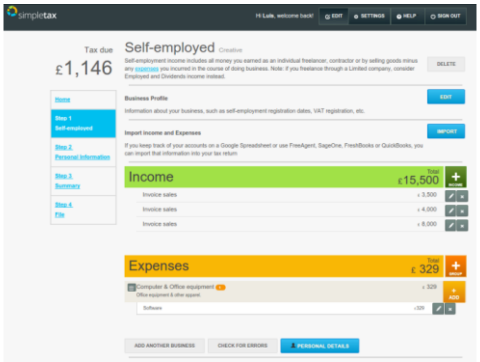


Figure 4 - Simpletax Dashboard with Skeuomorphism and New Structure

To apply this changes a script in JavaScript and using jQuery library was developed to change and add elements on the page to setup the new design prototype as can be seen on Figure 3 and Figure 4.

User Recruitment

For this test we defined a goal between 70 and 80 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.

4.2 Results

In this study we had 73 participants on the first test, 78 on the second test, 73 on the third one and 24 on the final test. Relatively to the last test the number of participants is lower due to the high drop rate that we got caused by the participants giving up the task or not answering the survey truthfully.

Then on table 1 show the interactions defined that need to be performed by the user during the task.

Table 1 - List of Interactions for the task in both Current and New Design

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

This is the list for all the steps identified with the cognitive walkthrough evaluation 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, being them “Continue to Summary” and “Submit to HMRC”, respectively. Also the big change between this two versions is the creation of multiple pages for the form. Now the user as 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.

Relatively to the Bayesian analysis we will focus in comparing the two conditions (Flat Design and Skeuomorphism) in both current and new design for an improvement of 1%.

4.2.1 Current Design Flat vs Current Design Non Flat

In this first iteration we tested the current design of the tool comparing the Flat Design against Skeuomorphism. After analyze and compare our results we found 6 steps where the users got an evidence of improvement of at least 80% for 1% improvement. The steps are the following:

Step 9 – Click Personal Details (87% evidence)

Step 12 – Save Personal Details (80% evidence)

Step 18 – Add Income (94% evidence)

Step 20 – Insert Amount (85% evidence)

Step 22 – Add Expense Group (97% evidence)

Step 29 – Check for Errors (93% evidence)

Conclusions

Relatively to the consequence of the change from Flat Design to skeuomorphism we had two cases in this iteration. One was on Step 20 (– Insert Amount) where the goal of the user was select a text box and insert an income value. 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.

The other case was the buttons that basically what we think is that the changes to Skeuomorphic were enough to give a different look and feel to the user. In other words just by adding gradients, bevels and other effects, the user is able to see this elements with less effort. A good example for this is the step 22 (Add Expense Group) that since is located in the end of a title bar (and not isolated), the difference between Skeuomorphism and Flat Design becomes evident as we can see on Figure 5 and Figure 6.

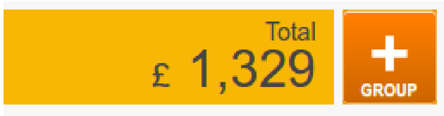


Figure 5 - Add Expense Group Button with Skeuomorphism

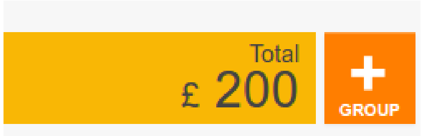
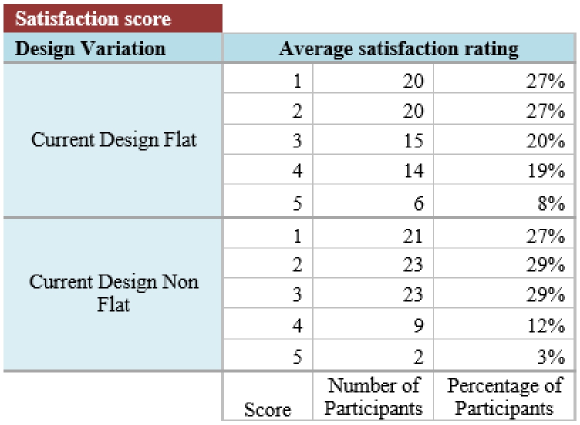


Figure 6 - Add Expense Group Button with Flat Design

Post-Questionnaire results

After the task we had the ASQ questionnaire to check how the users felt after using the tool. And surprisingly even having a slightly improvement on the success rate finishing the 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, which caused a strange look on the user interface.

Table 2 - Satisfaction Scores for the task in the current design being 1-low and 5-high



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

4.2.2 New Design Flat vs New Design Non Flat

For the second iteration with new design, as we did on the first, we performed the comparison of Flat Design against Skeuomorphism. After doing the analysis we found 3 steps with an evidence of improvement higher than 80% for 1% improvement. The steps are the following:

Step 9 – Click Personal Details (99% evidence)

Step 19 – Add Income (86% evidence)

Step 29 – Check for Errors (91% evidence)

Conclusions

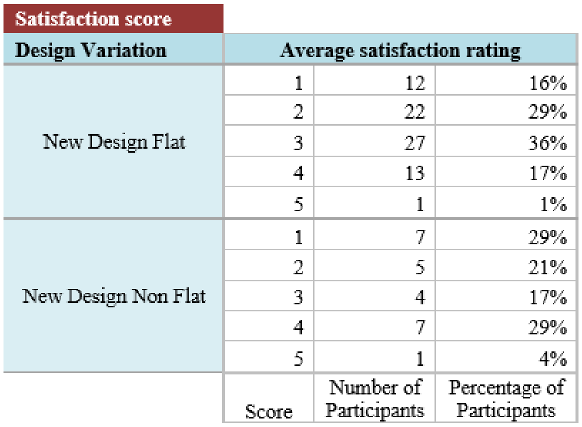
For this iteration we only found tendency of improvement in steps where the user needs to click a button. In step we got 86% evidence of improvement for the 1% improvement. However since the test has a low reliability, due to the number of participants in the second test, we are not sure if this is a real problem since we didn’t observed the same issue in similar steps.

Relatively to the steps 9 and 29 we got 99% (positive evidence) and 91% (strong evidence) evidence of improvement, respectively. We think that the probability is so high do to the buttons location. In other words since they are located in the end of the pages by changing to skeuomorphism they will become easier to see for the user.

Post-Questionnaire results

As we did on the first test we had a satisfaction survey after the task 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.

Table 3 - Satisfaction Scores for the task in the new design being 1-low and 5-high



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

4.3 Findings

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 interface).

After looking to this results the pattern that we identified is that the users were having problems whit the elements that are “hidden” in the interface due to the complexity. And when we added the skeuomorphism then they could found that elements easily. 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. 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 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 our conclusions for the first test are correct 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 were the user add the incomes and expenses, even without changing the appearance from flat 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 be 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

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.

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 analyze big amounts of data

Relatively to our work we were able to verify our 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, will be easier for the user understand what he needs to do even with the flat design.

6. 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.

First suggestion is test the “Almost Flat”, what we think is that with this new concept that is basically use flat but instead of remove all the style is maintain some components that maybe will be enough to give the affordance to the elements. One example of this concept is the Material design form 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.

Second suggestion for future work is testing interfaces from different fields, the objective is 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 to do this validation however due to unexpected issues we were not able to complete this third test.

7. 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.

[2] 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.

[3] 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

[4] 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.

[5] 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.

[6] 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

[7] 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.

[8] Tractinsky, N., Katz, A.. & Ikar, D., 2000. What is beautiful is usable. Interacting with Computers, 13(2), pp.127–145.

[9] 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.

[10] Tullis, T. et al., July 2002. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. Usability Professionals Association Conference.

[11] 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

[12] Wagenmakers, E.-J., 2007. A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review, 14(5), pp.779–804.

[13] Lewis, C. et al., 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI ’90. New York, New York, USA: ACM Press, pp. 235–242.

[14] Wharton, C. et al., 1994. The cognitive walkthrough method: a practitioner’s guide. , pp.105–140.

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

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

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

[18] Norman, D.A., 1999. Affordance, conventions, and design. Interactions, 6(3), pp.38–43.

[19] 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.

[20] 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.

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

[22] 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].

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

[24] 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/10.1145/122672.12