Your Reputation Precedes You:

The influence of expectations on usability and visual appeal in a web environment

Milica Stojmenović

M.A. Psychology.
B.Sc. Highest Honours Psychology, Minor in Statistics, Senate Medal.

A thesis submitted to the
Faculty of Science, Engineering, and Technology
in fulfilment of the requirements for the degree of

Doctor of Philosophy

Swinburne University of Technology
Melbourne, Australia

Summited: February, 2016

Awarded: April, 2016

© 2016, Milica Stojmenović

Abstract

The purpose of this thesis was to examine the impact of expectation on the perceived and objective usability and visual appeal of a website. The problem is that many studies have been done on the relationship between usability and visual appeal but the results of these studies vary vastly. There are many factors that influence the results, including website domain, the type of task, if incentive is given, and metrics used to get the usability and visual appeal measures. However, no one has examined the impact of expectations on these two variables.

A set of five preliminary studies was completed in order to get a website data set that significantly varied in levels of visual appeal and usability. This resulted in four website versions: (1) easy and pretty, (2) easy and ugly, (3) hard and pretty, and (4) hard and ugly. Five levels of expectations were implemented: (a) easy and pretty, (b) easy and ugly (c) hard and pretty (d) hard and ugly, and (e) the control – no expectations. Three main computer laboratory studies (in the form of user-based usability tests) were completed to determine the effect of textual and verbal expectations on visual appeal and usability.

Results showed that while textual expectations were effective, the combination of textual and verbal expectations influenced participants the most. Specifically, when usability and visual appeal levels were congruent (i.e. both were either high or both were low), then expectations influenced them both equally as participants tended to agree with the expectations, pre- and post-use. In fact, when told that the website was going to be hard and ugly, participants were discouraged from using it, stating it was too hard to use, and struggled more when using it. Similarly, participants thought that the website was easier to use and prettier in the high expectations group than in the low expectations group. When the website and expectations of usability and visual appeal levels were incongruent (easy but ugly, and hard but pretty), participant reactions are less predictable. In fact, while evidence suggests that expectations still impact ratings of visual appeal and usability, participants might choose to focus on the good expectation more so than the bad or vice versa. Also, poor visual appeal made the website harder to use, and poor usability took away from the positive experience of the visual appeal.

Outcomes of this research suggest that web developers and project managers should focus on investing in marketing just as much as in the development of a pretty and usable website, given that prior expectations do influence both how users perceive and use a website.

Dedication

I would like to dedicate this PhD thesis to my father, Prof. Ivan Stojmenović. He was and always will be my greatest mentor and guardian angel. If it wasn‘t for him, I wouldn‘t have made it this far. He taught me everything I know and gave me the motivation to work hard and aim high. I was blessed to have had him in my life and to have had him as my father.

Voleću te večno.

In loving memory of Prof. Ivan Stojmenović.
(06/10/1957-03/11/2014)

Acknowledgements

I would like to thank my father, mother, and brother who helped, guided, and supported me, in every way possible. Thank you, mama, for always encouraging me to do my best and to never give up. To my brother, thank you for creating the website manipulations used in this thesis. This thesis would definitely not have been possible without the three of you – my musketeers and guardian angels. Hvala vam, ljubim vas i volim.

I would like to thank all my dearest friends and colleagues from the old EN505 crew, who were always understanding and helpful. Leonard, thank you for everything.

Also central to this thesis were my mentors, professors John Grundy, Vivienne Farrell, and Robert Biddle. Thank you both for saving me in my darkest hour, for your understanding, support, guidance, and hard work.

Thank you.

Declaration

This thesis:

1. contains no material which has been accepted for the award to the candidate of any other degree or diploma, except where due reference is made in the text of the examinable outcome;

2. to the best of the candidate‘s knowledge contains no material previously published or written by another person except where due reference is made in the text of the examinable outcome; and

3. where the work is based on joint research or publications, discloses the relative contributions of the respective workers or authors.

Milica Stojmenović

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1. Introduction

The Internet offers major opportunities for competitive advantage in ecommerce, and can provide a replacement for paper-based documents and personal services to the public. Poor websites will easily lose the attention of the user when alternatives exist. Visual appeal and usability have been identified as effectors of trust (e.g. Fogg et al., 2002), enjoyment, quality (De Angeli et al., 2006), and purchasing intent (e.g. Geissler, Zinkhan, & Watson, 2006), among others. They have, hence, often been the subject of research. The purpose of this thesis was to obtain a better understanding about the usability and visual appeal of websites, by examining the influence of expectations on these two variables.

In Human-Computer Interaction (HCI), the most widely used definition of usability is provided by the International Standards Organization (ISO). According to ISO 9241/11 (1998), usability is the extent to which a given product can be used by a specific group of users, to achieve specific goals with effectiveness, efficiency, and satisfaction in a specific context of use. Effectiveness relates to how well users can achieve specific goals. Efficiency is the time taken to complete a given task, where less time is better. Satisfaction is the user's experience of acceptability. The context of use involves a predefined group of users, in set environments, who perform specific tasks with the equipment in question. The ISO definition of usability was used in this thesis.

In the current HCI literature, visual appeal is housed under the field of _æsthetics.' Aesthetics is used to describe two different concepts: a pleasant experience and a visual property attributed to objects. However, measuring aesthetic pleasure/experiences would require physiological measures, such as heart rate, which are out of scope for this thesis. Instead, this thesis focused on aesthetics as a visual property of an object requiring judgment of the appearance (Feagin, 1995). In other words, an object's aesthetic appearance is subject to cognitive judgment, also known as aesthetic appraisal (Blijlevens, 2011). Henceforth, in this thesis, aesthetic appraisal is referred to as _visual appeal'.

This thesis examines these two critical areas and in particular, the effect of expectations on usability and visual appeal. This is done on a website genre where participants do not have highly developed mental models and the website is gender and age neutral. There are many business and service providers who would benefit from a positive consumer attitude towards their visual appeal and usability; this is certainly the case for government websites where public opinion is not generally favourable. In particular, the primary focus of this thesis was the examination of government websites (i.e. tourism and city council websites at first, and this was later reduced to just city council websites). Research has not been done on corporate websites that do not have an online shopping function, even though such websites are common online, and users do expect them to be pretty and usable with useful content (Burtuskova & Krejcar, 2013).

City websites, such as tourism and government services and information, are becoming important because people are relying on them more to plan their travels (Litvin et al., 2008). In fact, over a third of searches are related to finding travel

information (Forrester Research, 2006). Governments have been affected by the explosion of the Internet as well, and have encouraged consumers to move the majority of their enquiries to the web. This is due to the current state of affairs: long queues both physical and on the phone, massive amounts of information available, and many forms that need to be organised, stored and searched. This information needs to be easily accessed by diverse groups of people, including the elderly, disabled, and regionalised groups and group coordinators. In addition, multiple people can have the same question, causing the need for repetition of dealing with the same question that could be more economically answered online. Still, errors and misunderstandings occur therefore having the online interface readily available to everyone allows for individual difficult inquiries to receive greater personal attention.

The process to date for moving people from government offices and requiring face to face service has been guided through the advents of new technologies and opportunities. Initially customers were encouraged to outlets such as shopping centre offices, self-serve kiosks and automated phone enquiries. The independent enquiry process moving online is a natural progression. However, why would there be an expectation of a government website to be usable and pretty, given that the everyday consumer's real life experience of countless documents, queues, and being moved from one counter to another or being transferred on the phone is often unpleasant or 'unusable'. To encourage people to move to the web to free up the government resources and to assist with the organisation and provision of data that is readily available to the consumers, making the web appear to be a more usable self-reliant option can only hope to improve services. Therefore, government websites provide an excellent example of how moving people‘s interactions with service providers to an online environment is highly beneficial to all parties. Thus, the primary focus of this thesis was the examination of government websites (i.e. tourism and city council websites at first, and this was later reduced to just city council websites).

In order to observe the impact of expectation on usability and visual appeal, we manipulated expectations to create ―eognitive dissonance”: a disagreement of information in an individual‘s thoughts and environments, which may cause stress. When dissonance occurs an individual strives to achieve consonance by reducing the inconsistency. The cognition that is most resistant to change is the most recent one (Harmon-Jones, Amodio, & Harmon-Jones, 2009). Here, the most recent behaviour would be experiencing the expectation. Therefore, the theory of cognitive dissonance states that new information can impact expectations and these expectations can impact behaviour. Thus, in this study, we induced different expectations of the usability and visual appeal of websites to examine the impact on the perception and use of the website, as reasoned by the theory of cognitive dissonance.

To see if expectations influenced visual appeal and usability, we first carried out a series of preliminary studies to obtain a website that was relatively unfamiliar to participants, to make the manipulation of expectations easier. Then, the website was manipulated to create several versions of it, ranging in usability and visual appeal. The manipulations were user tested and verified. The main study was then undertaken with

participants interacting with the website to solve information retrieval tasks under controlled conditions. We used textual descriptions of the website to seed users with expectations in written form (hence, written expectations). The results showed that written usability and visual appeal expectations can influence objective and subjective usability, and visual appeal.

The next section gives a brief summary of related research. This is followed by a description of mental models. Next is an outline of the preliminary studies. The main method and results of our experiments are then provided and discussed. We then discuss key findings, conclusions, and future work opportunities.

Background

It has been suggested that the visual appeal of interactive technology is the first aspect detected and thus it influences a user's first impression (Tractinsky, Katz, & Ikar, 2000). The relationship between visual appeal and usability is said to exist (Tractinsky et al., 2000) because a similar trend with visual appeal is experienced in psychology, and again in marketing. In psychology, it has been found that beautiful people are perceived to have more socially desirable traits (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). A phenomenon has been observed, called the _halo effect', in which people and things are judged and characters are assumed based on their appearance (Thorndike, 1920). In HCI, the halo effect has been applied to interfaces because beauty is the trait that is seen first, and that it influences subsequent perceptions of characteristics (Dion at al., 1972). However, many studies have been conducted in the area and the relationship has been investigated in many ways, with little consensus on the direction of the relationship. Some studies have found that usability and visual appeal are affected before system use (e.g. Hassenzahl, 2004; Hall & Hanna, 2004), others have found that they are affected after (e.g. Van der Heijden, 2003; Quinn & Tran, 2010; Hartmann et al., 2007), and still others have found that they are affected throughout system use (e.g. Tractinsky et al., 2000; Chawda et al., 2005). Only a small sample of the most central papers illustrating each of these was summarized.

Tractinsky and colleagues (2000) examined the effect of visual appeal on the perceived usability of automatic teller machines (ATM) because previous research had shown that visual appeal and usability are highly correlated (Tractinsky, 1997). Tractinsky and colleagues (2000) found that interface visual appeal affected both pre- and post-use perceptions of usability. Aesthetic interfaces influenced satisfaction, and the perceptions of quality and performance (Tractinsky et al., 2000). Tractinsky and colleagues (2000) concluded that there is a strong relationship between a user's initial aesthetic perception and the perceived usability of a system, and that this relationship endures even after interacting with the system. This view is shared with Norman (2004), who proposes that aesthetic design may be more influential in affecting user preferences than usability, but this would depend on the context in which both are assessed.

Katz (2010) examined the relationships between the perception of visual appeal and user experience of a fictitious search engine, before and after use. The results showed that visual appeal did not affect performance or satisfaction but pre-use aesthetic perceptions were correlated with perceived usability. Aesthetic interfaces were not considered to be more usable after system use. However, one limitation to the study was that participants were given a strong incentive to perform as the top three performing participants received a monetary reward, thereby possibly increasing the importance of usability to participants, and decreasing the importance of visual appeal pre-use.

Tuch and colleagues (2012) found that visual appeal did not affect perceived usability; rather, usability affected perceived visual appeal after use, on a search engine. They independently manipulated visual appeal and usability, and used multiple measurements of both constructs, so that the results could be comparable to other studies. Perceived usability and visual appeal were measured using many scales, along with task completion time, number of clicks, and success rate. The results showed no effect of visual appeal on perceived usability. In addition, after use of a low usability interface, ratings of classical visual appeal were lowered. Also, and similarly as in the Katz (2010) paper, the participants had a large incentive: in addition to being paid, the top three performing participants would be given an iPod, making usability very important to them.

Overall, the relationship between usability and aesthetics has not yet been decisively defined and may vary depending on context (Lee & Koubek, 2010; Tuch et al., 2012), target audience (Di Angeli et al., 2006), tasks, and mood not being accounted for in many cases (Lee & Koubek, 2010). There is a lack of standardized guidelines on how to alter visual appeal without potentially influencing usability, along with a difference in scales and measures used to capture perceived usability and aesthetics. Expectations and their impact on usability and aesthetics have not yet been researched or accounted for (Lindgaard et al., in press). These factors make the results of the findings on the relationship between usability and aesthetics hard to compare and an overall understanding of the relationship is still lacking.

Research Questions

This thesis examined the influence of expectation on usability and visual appeal for an Australian city council website. It has been suggested that the visual appeal of interactive technology is the first aspect detected and thus it influences a user's first impression (Tractinsky, Katz, & Ikar, 2000). The relationship between visual appeal and usability is said to exist (Tractinsky et al., 2000) because a similar trend with visual appeal is experienced in psychology, and again in marketing. In psychology, it has been found that beautiful people are perceived to have more socially desirable traits (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). A phenomenon has been observed, called the _halo effect‘, in which people and things are judged and characters are assumed based on their appearance (Thorndike, 1920). In

Human-Computer Interaction (HCI), the term ‗halo effect' has been applied to a phenomenon that is not unlike the original effect that Thorndike thus named. Specifically, the visual appeal of an interface is seen first, and it influences subsequent perceptions of characteristics (Dion at al., 1972), including usability. In this thesis, expectations are predicted to impact usability and visual. Therefore, the core research questions in this thesis are as follows.

1. Do expectations influence website visual appeal?
2. Are perceived and objective usability influenced by expectations?
3. What effect does textual expectation setting have on visual appeal and usability?
4. What effect does verbal expectation setting have on usability and visual appeal?
5. What happens when visual appeal and usability levels are congruent (i.e. are both either high or both are low)?
6. What happens when visual appeal and usability levels are incongruent (i.e. one is high and the other is low)?

Methodology

A series of controlled computer laboratory experiments, acquiring and analyzing qualitative and quantitative data were performed in order to investigate expectations, visual appeal, and usability of websites. Five preliminary studies were conducted to acquire a data sample that was empirically tested to ensure that the websites varied in visual appeal and usability levels. This was done using expert- and user-based usability testing during which both subjective and objective usability was measured, along with visual appeal. Once the dataset was developed, to test the impact of expectations on visual appeal and usability, three main studies were completed in this thesis. Main Study 1 involved textual implementations of expectations in the attempt to persuade participants with respect to visual appeal and usability of the easy and pretty, and the hard and ugly websites (i.e. visual appeal and usability were congruent in quality: both high or both low). The results were mixed which lead to Main Study 2, in which expectations were implemented both textually and verbally, as a means to reinforce the expectation. Given the success of the verbally reinforced expectations, Main Study 3 was also done with verbal and textual expectations. It involved the mixed conditions, in which usability and visual appeal levels were incongruent: easy and ugly, and hard but pretty. The effect of the induced expectations was experimentally measured, with a total of 12 conditions, across four website versions. More information on these studies can be found in the fourth chapter which is the method chapter and in chapters 5, 6, and 7 which are the experimental chapters.

Results and Contributions

The main goal of the research was to add to the understanding of the relationship between usability and visual appeal in a web environment by examining the impact of expectations on the two concepts. This has not previously been done in the HCI community. The controlled experimental approach taken in this study is not new but the application of textual and verbal expectations in a website environment is. The findings demonstrate that both textual and verbal expectations impact usability and visual appeal, especially when visual appeal and usability levels are congruent. This is important as these expectations can increase or decrease the perception and ease of use of a website. Below is a more detailed summary of findings in each of the experiments. The specific findings and discussions can be found in the respective chapters.

A set of preliminary studies resulted in the development of an empirically chosen and tested website data sample that statistically varied in usability and visual appeal. The data sample consisted of one website that was "easy and pretty" (i.e. the original website), manipulated into being: "easy and ugly", "hard and pretty", and "hard and ugly". According to participant responses, this website data sample was from a less familiar genre (i.e. city councils), which was necessary in order to control for prior experiences and expectations.

Main Study 1 had a two-by-three design and tested the easy and pretty website and the hard and ugly website, with easy and pretty, hard and ugly, or no expectations (i.e. control group). Expectations were embedded in a written task description (hence referred to as 'written' and 'textual' expectations in this thesis). Quantitative results showed that both pre- and post-use visual appeal ratings were significantly different within the hard and ugly website, between the low expectations and control conditions. For objective usability, the average number of clicks per task significantly differed within the easy and pretty website, where participants interacted more with the website that had the positive expectation than with the low. However, perceived usability was not statistically affected. The results of this study demonstrated that it was possible to affect users' perceptions of visual appeal and alter some aspects of objective usability of a website through the creation of written expectations. The next study examined a different method of expectation implementation in order to see if perceived usability could be affected and more significant results could be obtained.

Main Study 2 had a one-by-three design and tested only the easy and pretty website version, with easy and pretty, hard and ugly, and no expectations. In addition to having expectations embedded in a written task description, they were reinforced with a confederate who acted like they were a participant, finishing the study. As the confederate was leaving, the real participant was arriving. The confederate would then give a short, scripted speech stating their 'opinion' on the website's usability and visual appeal, then leave. This opinion had the same meaning to the written version of the expectations. Pre- and post-use perceived usability and post-use visual appeal statistically differed between the easy and pretty and the hard and ugly websites. For objective usability, the low expectations group of the easy and pretty website differed

from the high expectations group in the average number of clicks per task, average completion time per task, and the average number of passed tasks (where the low expectations group struggled more). Therefore, the verbal implementation of the set expectation in addition to the written form was highly successful in influencing participants‘ perceptions and experiences with the city council website.

The next step was to examine the easy and ugly website and the hard and pretty website in Main Study 3. This study had a two-by-three design and tested easy and ugly, hard and pretty, and no expectations (i.e. control group). Similarly to Main Study 2, expectations were both embedded in a written task description and reinforced by the confederate verbally. Results were weaker than anticipated, since only pre-use perceived usability was found to vary in the hard and pretty website between the control condition and low usability, high visual appeal condition. However, the qualitative data seem to suggest that expectations did in fact influence participants. According to the cognitive dissonance theory, people can choose to reduce the dissonance by adding or increasing the importance of the consonant cognition, or by taking away or reducing the importance of the dissonant cognition. Thus, the lack of statistically significant findings does not automatically indicate that expectations are not influencing people‘s perceptions of visual appeal and usability. In contrast, the findings seem to support the theory‘s explanation in stating that people are complex and sometimes they ignore new information, rather than accept it. The next step would be to investigate what factors determine whether an individual will accept or reject the expectation.

For website design, the findings in this thesis suggest that how well a website is made is not the only factor that influences what people think about it. As demonstrated in this research through the use of a confederate, a bad reputation can turn people against your website, even if the reputation is not true. To overcome this, one could invest in marketing to give a website a more positive reputation right from the beginning. It will influence people before they use it and, according to the results of this study, last throughout use to influence their opinions after having used the website.

In this study, participants were forced to use the website, whereas in reality, there are thousands of websites to choose from and the competition can be fierce. If you advertise, there is a greater chance that people will (1) know about it, (2) know something good about it, (3) be willing to check your website out, and (4) like it a bit more after they use it.

Overall, this research contributes to an improved understanding of the relationship of usability and visual appeal by added understanding of the degree to which expectations affect these variables, in a web environment. The next step, given the rapid growth in popularity of tablets and cellphones in the last few years, would be to examine the applicability of the results on different sized screens.

Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 is the literature review which discusses the concepts of usability and visual appeal, along with the

findings in HCI pertaining to the relationship between usability and visual appeal. It outlines the gap in the current literature, from which the research questions were derived, for this thesis.

Chapter 3 gives the theoretical background, explaining the formation of expectations and how they can be used to predict behaviour. In particular, mental models, cognitive dissonance, the consistency theory, counter-attitudinal advocacy, and the post-dissonance theory are discussed. The theory-based research hypotheses are presented at the end of Chapter 3.

Chapter 4 is the methods chapter. It describes the approach taken to answer the research questions and test the hypotheses. Namely, it discusses the different data collection methods, and visual appeal and usability measurements and approaches used in the field. These are examined and the measures used in this thesis are justified.

Next, Chapter 5 describes the preliminary studies undertaken to obtain the varying websites required for further studies. There were five preliminary studies, including both user- and expert-based usability testing. Furthermore, this chapter contains the description of the processes for the selection, manipulation, and verification of the website data set. Each preliminary study has its own introduction, method, results, discussion and conclusion section.

The Main Studies 1 and 2 are described in Chapter 6 and involve congruent visual appeal and usability levels. Chiefly, only the easy/pretty and hard/ugly website versions were tested in this chapter. In Main Study 1, the impact of textually implemented expectations was tested. In Main Study 2, verbally reinforced expectations were implemented in addition to the written form. Each study has its own introduction, method, results, discussion and conclusion section.

Following this, Chapter 7 describes Main Study 3, where incongruent levels of visual appeal and usability were tested with both written implemented and verbally reinforced expectations. Specifically, the easy/ugly and hard/pretty website versions are used in this study. This study also has its own introduction, method, results, discussion and conclusion section

Chapter 8 is the general discussion. This includes a general summary of all the findings and of their implications. Other implications are also given, involving theoretical and practical implications. The discussion chapter also includes a summary of limitations, across the studies in this thesis.

Finally, Chapter 9 is the conclusion chapter. In this chapter, the summary of implications is given. Then, future work is suggested, including examining other theories, the use of confederates, other website domains, etc. Following these are the references and appendices.

Chapter 2. Literature Review

Today‘s online experience can be quite daunting when looking for information, with millions of websites to choose from. Some common questions that researchers in the field end up asking themselves are: How do people end up choosing a website to interact with? What are website usability and visual appeal? How do they influence perception and use of websites? Hundreds of researchers have published a great many papers but have not yet reached consensus on the relationship. Perhaps understanding the psychological issue of prior experience and expectations may help bridge the gap.

This thesis examines the effect of expectations on usability and visual appeal, in a website domain. The purpose of this chapter is to summarize and examine related work done in the field, to help answer some of these questions. To accomplish this purpose, this chapter is structured as follows. It begins with a brief history of the concept of usability in HCI, which leads to the definition of usability, as used in this thesis. The next section defines visual appeal by looking at its roots in aesthetics. Following these two definitions is a section on the findings to date that relate to the relationship between usability and visual appeal, and what literature might be missing from the existing pool to help understand the relationship. Then expectations are introduced, as a possible explanation for the gap. A summary of relevant papers of work on expectations in different fields is presented last.

Defining Usability

More and more, human task performance is supported by technology. The usability of technology became an issue in the late 1980‘s when the graphical user interface (GUI) became a widespread standard for computers and their popularity grew. Before then, UNIX used command languages as its only user interface style and the original interface did not permit anyone without training to use it. It was only as technology slowly advanced and user interfaces started including colour and icons that usability and usability testing became necessary so that the average person could use computers. In fact, the concept of usability only gained momentum in the 1990‘s, when the Internet and Windows ‘95 massively gained popularity. At that time, the purpose of usability was to ensure that the technology worked efficiently and effectively, allowing the user to finish specific tasks precisely and quickly. However, with the advent of the Internet, World Wide Web, and new modes of interaction, the concept of usability has been expanded to include broader aspects of the user experience. The term user experience is now used to describe two slightly different concepts: experience (such as aesthetics, fun, satisfaction) and usability (how easily a system is to use; Bevan, 2009a). User experience methods strive to improve user satisfaction while usability methods focus on improvement in human performance (Bevan, 2009a). Designers focusing on the concept of usability aim to improve productivity and reduce system errors (Wickens,

Lee, Liu, & Becker, 2004). Usability methods investigate the ease of use of products, and improvements so that handling the new technology is as intuitive as possible.

Recently, usability has also been defined as a subcategory of ―Quality of Use", which is a user's impression of the quality of a product and the efficiency with which it can achieve a specific goal (Bevan, 2009b). In this definition, usability evaluation deals with the operability of technology and usually results in a set of user requirements. These user requirements deal with the effort of use (Bevan, 2009b), navigation and visual elements and functionality (Yang, 2009).

The most widely used definition of usability is provided by ISO. According to ISO 9241/11 (1998), usability is the extent to which a given product can be used by a specific group of users, to achieve specific goals with effectiveness, efficiency, and satisfaction in a specific context of use. Effectiveness relates to how well users can achieve specific goals. Efficiency is the time taken to complete a given task, where less time is usually better (unless the system is a game and longer times for exploration and challenges could be considered better). Satisfaction is the user's experience of acceptability and fun. The context of use involves a predefined group of users, in set environments, who perform specific tasks with the equipment in question. The ISO definition of usability will be used in this thesis, where users are Swinburne university students, the environment is government websites in a computer laboratory setting, and the task is information retrieval. User experience was examined in this thesis as satisfaction, the subcategory of usability from the ISO definition.

In a recent paper, Thielsch, Engel, & Hirschfeld (2015) stated that usability was a main characteristic examined when evaluating websites. One of the core issues with usability is that it is an interactive construct but it is often judged pre-use when investigating first impressions (Thielsch et al., 2015). Thielsch and colleagues (2015) examined a concept they called ‗expected usability' and its relation to a concept they coined called ‗experienced usability' and objective usability. Their results showed that pre-use usability was not related to post-use usability or to objective usability. However, their definition of ‗expected usability' differed from the ‗expectations' in this thesis. Mainly, ‗expected usability' was defined as pre-use usability and its ‗experienced usability' was in fact post-use usability. Therefore, while their terms used appear new, the concepts were not. Other usability methods and approaches are in Chapter 4, the general methods chapter.

Website usability has been studied by HCI experts but usually it is in the context of a new website where usability is being evaluated. There, it is more often examined for development. For research purposes, usability is usually studied in conjunction with other concepts, such as visual appeal. Thus, visual appeal is discussed next.

Defining Visual Appeal

The definition of aesthetics and what makes a particular object beautiful have been debated on by philosophers for almost three thousand years (Reber et al., 2004; Taebi, Aldabbas, & Clarskon, 2013). In philosophy, aesthetics is a branch concerning

itself with beauty, taste (Lawrence and Tavakol, 2007), and sound (Taebi et al., 2013). Even mathematical principles have been applied to attempt to measure, explain, and to reproduce aesthetically pleasing artefacts. The golden ratio (Pacioli, 1509; Livio, 2002) is one of the most famous attempts at quantifying (i.e. proportioning) aesthetically pleasing characteristics. Pacioli (1509) equated beauty with divinity, which his comrade Leonardo da Vinci incorporated into his famous man in a box and circle which was an illustration in Pacioli's book (1509).

A literature review done by Bargas-Avila and Hornbaek (2011) found that aesthetics is the most researched topic in UX empirical studies. Yet, a recent review of empirical studies found that the conditions under which one variable influences the other are not yet known (Tuch et al., 2012). There is a lack of consensus in the literature on the definition of aesthetics and there are differences in terminologies (e.g. visual appeal, beauty, aesthetics, attracativeness, etc.; Chang, Lai, & Chang, 2007; Hekkert, Snelders, & van Wieringen, 2003) used to define the concept. It is linked with the concept of „beauty' when judging it (Zettl, 1999) and considering its importance (Tractinsky & Lavie, 2004). Visual aesthetics is a process by which people examine and react to a number of visual elements (Zettl, 1999), and a property attributed to objects. According to the Merriam-Webster Online Dictionary (MWOD), aesthetics is the philosophy of the nature, creation, and appreciation of beauty (MWOD, 2004). It is a particular understanding of what is pleasing to the senses, especially sight (i.e. the appearance), such as clear graphics (MWOD, 2012) and colour (Lavie & Tractinsky, 2004; Taebi et al., 2013).

While some say that „beauty is in the eyes of the beholder," (Titelman, 1996) others argue that it is a property of the object. For example, the processing fluency theory states that the more fluently an object can be processed, the more positive the aesthetic response (i.e. the prettier) would be to that object (Reber et al., 2004). An object is more fluently processed when it is prototypical (Thielsch & Hirsschfeld, 2010; Winkeilman et al., 2006). This was also hypothesized by Norman (2004) with the theory of emotion according to which very briefly shown complex stimuli can only be reacted to emotionally, as no time is allowed for cognitive assessment. Yet, emotions are out of scope for this thesis.

In the current HCI literature, aesthetics is used to describe two different concepts: a pleasant experience and a visual property attributed to objects. However, visual aesthetics needs a more precise definition (e.g. Moshagen & Thilsch, 2010) so that researchers can more readily operationalize and evaluate it and coordinate their work. Since beauty is said to be positive and intrinsic to the viewer (Santayana, 1896; 1955), it is seen to create pleasure without further thought about the object's usefulness (Maritain, 1966; Read, 1972). For instance, ice cream on a warm day is positive and intrinsic. The experience of pleasure is caused by the bodily sensation of eating it, and not by a visual property of the object (Reber et al., 2004). This experience of aesthetics has been referred to as aesthetic pleasure (Blijlevens, 2011), which is an object's by-product; the pleasurable feeling which serves as positive reinforcement to specific behaviours (Tooby & Cosmides, 2001). In the case of the ice cream, the pleasurable and

cooling experience of eating the ice cream on a hot day would reinforce the behaviour of getting ice cream to cool down on other hot days.

However, measuring aesthetic pleasure would require physiological measures, such as heart rate, ElectroEncephaloGraphy (EEG), Galvanic Skin Response (GSR), and body temperature, among others, which are outside of the scope of this thesis. Instead, this thesis will focus on aesthetics as a visual property of an object, and of the judgment of the appearance (Feagin, 1995). This concept has also been referred to as aesthetic appraisal (Blijlevens, 2011). According to Blijlevens (2011), aesthetic appraisal is the cognitive judgment of an object‘s aesthetic appearance. For example, people look at art in order to enjoy its beauty rather than to experience pleasure – meaning that people experience beauty as an object‘s property (Reber et al., 2004). For the purpose of this thesis, the definition of aesthetics will be aesthetic appraisal as introduced by Blijlevens (2011). In the context of the World Wide Web, aesthetics will be defined as a cognitive judgment of a website‘s aesthetic appearance and is mainly referred to as ‗visual appeal‘ throughout this thesis.

Interface visual appeal thus is a subjective judgement dependent on personality, cultural background, and gender (Filonik and Baur, 2009). It is also considered to be qualitative and affect-based, while usability is measured by more objective means, with efficiency as a primary measure (Butler, 1996). There are vast resources to guide improvement in usability however there is a lack of HCI guidelines for aesthetic design with what exists usually in the form of classical aesthetics (Lavie & Tractinsky, 2004). Classical and expressive aesthetics are two of Lavie & Tractinsky (2004) scales of aesthetics. In HCI, classical aesthetics refers to the aspect of screen space management such as contrast, repetition, alignment, and proximity (also known as the CRAP usability principles). Expressive aesthetics refers to the creativity and originality and novelty of the design (Lavie & Tractinsky, 2004). A study by Sonderegger, Sauer, & Eichenberger (2013) found that classical and expressive aesthetics were indeed different concepts and that both positively affected perceived usability. Yet, website visual appeal has seldom (if ever, to the extent of our knowledge) been examined in HCI without the accompaniment of usability or other HCI constructs.

In summary, while usability is objective with efficiency as a primary measure, visual appeal is subjective and is more affect-based (Butler, 1996). There are vast resources to guide the improvement in usability of information technology systems; however there is a lack of HCI guidelines for aesthetic design (Lee & Koubek, 2010). This thesis retained the ISO definition for usability and we focus on aesthetics in the sense of Blijlevens‘ definition of aesthetic appraisal, which will be referred to as visual appeal throughout this thesis. Given these definitions, the next section summarizes main findings in the literature on the relationship between usability and visual appeal.

Relationship between Usability and Visual Appeal

The usability and visual appeal of objects are age-old topics in HCI (Taebi et al., 2013). They are highly related – changing one may influence the other (Nielsen &

Molich, 1990). In the context of HCI, it has been proposed that "What is beautiful is usable" (Tractinsky et al., 2000) because users are more tolerant with errors when the system is good-looking (Hartmann, Sutcliffe, & Angeli, 2008). Visual appeal has been shown to be the differentiating factor when a user chooses a product between two equally functional items (Lindgaard, 2001; Crawford, 2004; Taebi et al., 2013). Moreover, they have a common characteristic in user satisfaction (Taebi et al., 2013), further complicating the relationship. Many studies have been conducted in the area and the relationship has been investigated in many ways, with little consensus or general agreement on the direction of the relationship (Sonderegger et al., 2014). Some studies have found that usability and visual appeal are affected before system use (e.g. Hassenzahl, 2004; Hall & Hanna, 2004), others have found that they are affected after (e.g. Van der Heijden, 2003; Quinn & Tran, 2010; Hartmann et al., 2007), and still others have found that that they are affected throughout system use (e.g. Tractinsky et al., 2000; Chawda et al., 2005). Hundreds of papers have been published on this topic but only a small sample of the most central papers illustrating each of these is summarized here.

The relationship between aesthetics and usability is said to exist (Tractinsky et al., 2000) because a similar trend with beauty is experienced in psychology, and again in marketing. In psychology, it has been found that beautiful people have more socially desirable traits (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). It has been hypothesized that there is a 'halo effect' where people and things are judged based on their appearance, and characteristics are assumed based on those judgments (Thorndike, 1920). In HCI, the halo effect has been applied to interfaces because beauty is the trait that is seen first, and that it influences subsequent perceptions of characteristics (Dion at al., 1972). Many studies have found that visual appeal has a positive effect on perceived usability (e.g. Kurosu & Kashimura, 1995; Tractinsky, 1997; Brady & Philips, 2003; Schenkman & Jonsson, 2000; Nakarada-Kordic & Lobb, 2005; Thuring & Mahlke, 2007; Hartmann et al., 2008; Sauer & Sonderegger, 2009; Sonderegger, Zbinden, Uebelbacher, & Sauer, 2012; Thielsch, Blotenberg, & Jaron, 2013). Other have found that visual appeal also influences trust and credibility (Robins & Holmes, 2007), and overall impressions (Tuch et al., 2010).

Tractinsky and colleagues' (2000) examined the effect of aesthetics on a user's perception of quality of interaction (in the context of Automatic Teller Machines: ATMs) because previous research has shown that aesthetics and usability are highly correlated (Tractinsky, 1997). They found that interface aesthetics affected both pre- and post-use perceptions of usability. In addition, aesthetic interfaces influenced satisfaction, the perceptions of quality, and performance (also referred to as objective usability in this thesis; Tractinsky et al., 2000). Other findings concur with these, in that they found an increase in performance with increase of visual appeal (Moshagen, Musch, & Goritz, 2009; Sonderegger, & Sauer, 2010). Another finding suggests that visual appeal may be more influential in affecting user preferences than usability (Norman, 2004). Tracinsky et al. (2000) concluded that there is a strong relationship between a user's initial aesthetic perception and the perceived usability of a system, and

that this relationship endures even after interacting with the system. This view is shared by Norman (2004), who proposes that aesthetic design may be more influential in affecting user preferences than usability, but this would depend on the context in which both are assessed. However, the ATM context Tractinsky and colleagues (2000) used in their study is somewhat limiting because ATMs do not usually vary vastly in aesthetic value as they are there for a very limited and goal-specific purpose. Gorgeous flowers and beautiful landscape pictures would be inappropriate for ATM machines, and ATMs were far more primitive in the late nineties than they are today. The difference between the high and low aesthetic conditions is thus questionable, especially because the aesthetic manipulation involved only the rearrangement of the spatial organization of the ATM buttons. Given that the authors did not present a clear definition of aesthetics, it is hard to determine if their manipulation really reflected a manipulation of aesthetics, or if it was a manipulation of usability. It is highly likely that the button arrangement merely reflected a difference in perceived order or even the habit of seeing the buttons in a particular order. Therefore, the interpretation of the results may not be justified by the stimuli. Thus, the finding that aesthetics influences usability both before and after use may be specific to the context in which this study was done. Moreover, the measures used to rate aesthetics and usability were non-validated single-item questions. While that paper‘s method may be flawed, its findings that aesthetics affects usability both before and after system use sparked the interest of many researchers to examine the relationship of visual appeal and usability.

However, recent literature demonstrated mixed findings with respect to the relationship between usability and aesthetics. For example, Sangwon and Koubek (2010) found that user preference was significantly affected by aesthetics but marginally affected by usability, pre-use. Yet, after use, user preference was significantly influenced by both usability and visual appeal. Katz (2010), for example, examined the relationships between the perception of aesthetics and user experience of a fictitious search engine, before and after use. The results showed that aesthetics did not affect performance or satisfaction but pre-use aesthetic perceptions were correlated with perceived usability. Other studies also found that visual appeal was unrelated to performance (e.g. Chawda, Craft, Cairns, Ruger, & Heesch, 2005; Hartmann, Sutcliffe, & de Angeli, 2007; Thuring & Mahlke, 2007). Also, visually appealing interfaces were not considered to be more usable after system use (Katz, 2010). Satisfaction was highly correlated with the relevancy of search results and with the pre-use perception of aesthetics. However, one limitation to the study was that participants were given a strong incentive to perform as the top three performing participants received a monetary reward, thereby possibly increasing the importance of usability to participants, and decreasing the importance of aesthetics after use.

Several other studies did not find a positive influence of visual appeal on perceived usability (e.g. Van der Heijden, 2003; Ilmberger, Schrepp, & Held, 2008; Van Schaik & Ling, 2009; Tuch, Roth, Hornbæk, Opwis, & Bargas-Avila, 2012). In particular, Ilmberger et al., (2008) Tuch et al., (2012), and Bartuskova & Krejcar (2013) found that visual appeal did not affect perceived usability; rather, usability affected

perceived visual appeal after use. User performance has also been found to decrease with high visual appeal, suggesting a negative relationship (Ben-Bassat, Meyer, & Tractinsky, 2006; Sauer & Sonderegger, 2009).

Ilmberger and colleagues (2008) examined which cognitive processes could explain the relationship between aesthetics and usability. They examined participant responses before and after use of an online shopping website, with conditions and websites differing in usability and aesthetic levels. The results revealed that usability influenced visual appeal, and not the other way around as Tractinsky and colleagues (2004) found. Ilmberger and colleagues (2008) also found that user mood influenced perceived usability, and that visual appeal can influence mood. Since visual appeal can influence mood, which influences perceived usability, Ilmberger and colleagues (2008) concluded that designers should strive to achieve both high usability and visual appeal. However, one limitation of the method used in Ilmberger and colleague‘s study was that visual appeal was manipulated only by adjusting colour, while usability aspects were manipulated in seven ways, including changing both screen elements and the depth and hierarchy of the menu. However, logos, video, and depth of field have been found to be significantly important design elements for visual appeal (Sutcliffe, 2001). Therefore, the usability manipulations may have been more severe and could have outweighed the aesthetic manipulations, influencing the results.

Tuch and colleagues (2012) found similar results with respect to the relationship between usability and visual appeal, on a search engine. They independently manipulated visual appeal and usability. Usability was manipulated by only changing the text in the interface, hindering the navigability and hierarchy of the website. Visual appeal was manipulated by changing the colours of the backgrounds to less pretty options. The authors also used multiple measurements of both constructs, so that the results could be comparable to other studies. Usability was manipulated by changing the labels and assigning different items to menu categories. This kept visual appeal constant because the versions all looked identical. Visual appeal was manipulated by changing the background colour, background texture, and decorative graphic elements of the page, keeping usability constant as all versions contained the same product items, and had the same depth and breadth of menus. Perceived usability and visual appeal were measured using many scales, along with task completion time, number of clicks, and success rate. The results showed no effect of visual appeal on perceived usability. In addition, after use of a low usability interface, ratings of classical visual appeal were lowered. However, similar to Ilmberger and colleagues‘ (2008) paper, Tuch and colleagues (2012) note that the usability aspect was strongly manipulated, while visual appeal manipulation was slightly less drastic between the conditions. Therefore, the usability manipulation could have _outshone‘ the aesthetic manipulation, causing the results. Also, and similarly as in the Katz (2010) paper, the participants had a large incentive: in addition to all participants being well paid, the top three performing participants would be given an iPod, making usability very important to them.

Summary. The relationship between visual appeal and usability is not yet well understood and the circumstances under which any of these results occur are not yet known (Sonderegger et al., 2014). For example, Katz (2010)'s results demonstrated significant correlations between perceived visual appeal and perceived usability and usefulness before system use, but not after. Yet, Tuch et al., (2012) found that visual appeal did not affect perceived usability; rather, usability affected perceived visual appeal after use. Still, others argue that both concepts are important in the UX of a product, but they influence user perception in different ways (Taebi et al., 2013). Visual appeal helps create the first impression which can lead to an automatic peak in interest towards the website. Upon use, usability becomes more important as it is the factor that keeps users on a particular website (Taebi et al., 2013). Yet, not many concur with these findings. Therefore, the relationship between usability and visual appeal has not yet been defined or generally accepted by researchers in the field. To examine why this disparity is occurring, the next section discusses the current literature's shortcomings.

Limitations of the Research to Date

The majority of the current reported research in these areas utilizes correlational data which makes it impossible to establish causality in the relationship between usability and visual appeal (Tuch et al., 2012). The results of the experimental studies differ, making it hard to deduce an overall understanding of the usability-visual appeal relationship. Further, the manipulations in these experimental studies require the manipulations of visual appeal and usability to be independently manipulated so that they do not influence each other. However, each study has different experimental manipulations that were neither systematic nor independent (Bartuskova & Krejcar, 2013) of both variables as a possible justification for the gap in the literature (Tuch et al., 2012). For example, Tractinsky et al. (2000) changed the visual appeal of their interface by moving some objects on the screen – yet object proximity and alignment are common features of usability that may have been altered as well.

In addition, in Tuch et al.'s (2012) paper, only the background colour was changed. Yet, altering the contrast of the background could change text legibility, making it less usable. Additional challenges in the existing literature include using different measures of both variables. Some researchers even use self-made, non-validated measures (e.g. Harman et al., 2007; Quinn & Tran, 2010; Chawda et al., 2005). Therefore, to help alleviate these issues, the work in this thesis strived to use only independent and systematic manipulations of visual appeal and usability, and used only validated scales for the two concepts.

Another limiting factor is the lack of control of a person's psychology. In particular, a visually appealing product can evoke a positive emotional response, which can in turn improve mood, and finally increase system ratings (Tractinsky, 1997). Tuch and colleagues (2012) thus suggested that future research should examine the impact of the affective experience in the aesthetic evaluation. Since this thesis did not focus on the emotional aspect of aesthetics but on the cognitive judgement, this was not done here.

However, this did come as an indication that there was an unaccounted, personal/internal aspect to the usability-visual appeal relationship. Additionally, a factor that can play a key role in the dynamic between usability and visual appeal is a user‘s experience (Tractinsky, 1997).

McLellan, Muddimer, and Peres (2012) found that prior experience and familiarity with a product increased usability ratings, regardless of product type. Prior experiences shape our mental models (discussed in Chapter 3) and mental models help us create expectations as to what is about to happen. Thus, what happens if participants came to the study with a previous bad experience or an overly good one with a similar system to the one being tested? They would be familiar with developed mental models and their own expectations. McLellan and colleagues‘ (2012) work suggests that they be more proficient in it and that in turn may impact their liking of the system. Thus, the work to be done in this thesis must be done based on first impressions, in an unfamiliar domain because controlling for or competing against developed expectations would be impossible at this stage. The gap in the literature may be filled by examining the initial impact of a controlled set of expectations.

Other website studies exist but their purposes and topics do not align with this thesis‘. They tend to strive to achieve better usability or to create websites that will maximize profits. For example, related to city websites, are tourism and hospitality website studies (e.g. Ip, Law, & Lee, 2011; Hashim, Murphy, & Law, 2007). Ip et al.‘s (2011) reviewed a series of website hospitality studies in order to create a website evaluation system that assessed features and effectiveness. Hashim et al. (2007) examined effectiveness, validity, and reliability of hotel websites to increase profits and to also create an evaluation method for hotel websites. However, these are not within the scope of the thesis, as we did not want to increase profit margins or create website evaluation methods. Instead, the work in this thesis strived to further the understanding of website perception and interaction by uncovering the influence of expectations on people‘s perceptions and interaction with websites.

Summary. Overall, the relationship between usability and visual appeal has not yet been decisively defined and may vary depending on context (e.g. Di Angeli et al., 2006; Hartmann, et al., 2008; Lee & Koubek, 2010; Tuch et al., 2012), target audience (Di Angeli et al., 2006), tasks, and mood not being accounted for in many cases (Lee & Koubek, 2010). Also, there is a lack of standardized guidelines on how to alter visual appeal without potentially influencing usability, along with a difference in scales and measures used to capture perceived usability and visual appeal. These factors make the results of the findings on the relationship between usability and visual appeal hard to compare and an overall understanding of the relationship is still lacking. Moreover, prior experiences have not been properly accounted for and the impact of expectations on usability and visual appeal has not yet been researched in the HCI community. The work in this thesis contributes to an improved understanding of the relationship of usability and visual appeal by determining the degree to which expectations affect this relationship, in a web environment.

Expectations

In this thesis, expectations are defined as beliefs about what will occur. They are sometimes referred to as predictions, assumptions, and surmises. Expectations are formed from mental models. Mental models are formed when people interact with objects, because they develop an understanding and internal representations of the objects (Rouse & Morris, 1986; Sarter, Nadine & Woods, 1991; Sasse, 1991; Sinreich, Gopher, Ben-Barak, Marmor, & Lahat, 2005). Thus, mental models are a combination of an individual's subjective perceptions, concepts, and ideas (Sinreich et al., 2005). In other words, they are an individual's summary of previous experiences. They are used when interacting with the environment because mental models allow people to understand and remember relationships between objects in the environment, and they also create expectations of what is likely to occur (Rasmussen, 1979; Rouse & Morris, 1986; more on mental models can be found in Chapter 3). Therefore, expectations are predictions of what is likely to occur, based on previous experiences for a given context.

Expectations differ from biases. Biases are unfair prejudices that can be for or against something or someone (Oxford Dictionaries, 2015). Expectations can differ on a case-to-case basis which differentiates them from biases in that biases tend to be constant for a particular subject. For example, a bias would be that government websites are bad, and they tend to be bad no matter what government website you use. However, the chances of an unbiased expectation to be the same each time is low. A website example would be that you could expect one government website to be bad, based on your previous experience with it, but expect a different one to be better.

Expectations in HCI

Very little has been researched on the impact of expectations on visual appeal and usability in websites. Every relevant piece of literature found is summarized here. Ludden, Schifferstein, & Hekkert (2012) investigated incongruent reality versus expectations, design, and emotion with tangible objects. In particular, the authors studied people's reactions when they came across something that would look like it would feel one way but felt like something else. For instance, they observed reactions to people interacting with a mug that looked heavy because it looked like it was made of metal but was in fact made of hollow plastic and was really light. Thus, Ludden et al. (2012) examined what happened when the expectation created by the appearance of the product was incongruent with the feel of the product. The mismatch between the object's expected and real tactile experience resulted in surprise which was then followed by either positive or negative emotions. A positive emotion as more likely to follow with repeated exposure and familiarity. However, Ludden et al. (2012) do not offer a means to predict the response of mismatched expectation-reality (i.e. cognitive dissonance – please see Chapter 3), and their work dealt with tangible objects.

Therefore, it may not be readily compared or used to predict someone‘s reaction to usability and visual appeal in an online environment.

Sokkar & Law (2013) addressed similar issues to the ones in this thesis but did not present a study of user behavior. They suggest a model with three phases to online shopping decision-making. One phase occurs before interaction, where expectations are thought to impact perceived qualities of e-commerce. Yet, no work was cited or done to support this claim. The work in this thesis examines and supports that aspect of their model.

Word-of-Mouth

Apart from the visual cues used by Ludden et al., (2012), there seems to be one other way in which expectations have been implemented in the literature: word-of-mouth (WOM; Granovetter, 1973). All communication has the common purpose of sharing information (Parush et al., 2011). WOM tends to be about people‘s experiences (Smith, 1993). In its initial definition, WOM included only verbal communication, in the form of face-to-face communication, and ‗hearsay‘ (i.e. what an unknown person said but the message got to you through someone you know). Recently, WOM has expanded to include textual and video references, such as user reviews and Youtube videos, respectively. People prefer WOM over standard marketing channels because WOM is easier to understand and more trustworthy (Smith, 1993). In addition, WOM product reviewers are regarded as the most credible, objective, and influential since they have been unbiased and unpaid reviews of things and experiences (e.g. Kamins et al., 1997). Yet, since people prefer WOM over other mediums (Herr, Kardes, & Kim, 1991), companies have recently largely adapted to using sponsored WOM when advertising. In marketing, WOM is managed by employing an agent to seed the message out (e.g. Youtubers are often sponsored to give positive reviews about a company‘s products). The next two sections describe the relevant studies found regarding textually and verbally implemented expectations, described respectively.

Textual Expectations

Generally, having polarized descriptions of upcoming tasks can be considered biasing participants. Yet, this occurs in life: social media and user reviews tell us what products are good/bad (e.g. Smith, Donnavieve, Menon, Satya, & Sivakumar, 2005). In this thesis, it is argued that positive or negative texts (as well as verbal communication) taint users with expectations that can alter their perception and use of websites.

Online marketplaces such as eBay incorporate both seller and buyer feedback into their business models. These reputations help both parties acquire trust in each other (e.g. Gefen et al., 2003). A buyer‘s trust of a seller depends on their perception of the seller‘s credibility and benevolence because credibility prevents adverse selection while benevolence minimizes potential moral hazard (Pavlou & Dimoka, 2006). However, buyers cannot reliably trust or ascertain a seller‘s credibility and benevolence with just a

numerical star rating. Instead, a much more reliable predictor is feedback left by previous buyers (Pavlou & Dimoka, 2006). User reviews can be considered online versions of word of mouth communication. Most online consumers actively look for and readily accept reviews because it effectively manages massive amounts of online information (Smith, Menon, & Sivakumar, 2005; Pavlou and Gefen 2004, 2005). One study found that people relied on peer and editorial reviews and recommendations more so than other means, such as paid ads, yet user reviews were seriously under-researched (Smith et al., 2005). These textual reviews can implicitly convey information about perceived quality, ease of use, and usefulness (Davis, 1989). In fact, positive feedback also increases trustworthiness (Lim et al., 2007) and price premiums for reputable sellers (Ba & Pavlou, 2002).

Amazon, a large online market, sued over 1000 of its users for writing false reviews online (Anand, 2015). Amazon‘s users used to get paid for leaving feedback on a particular product. Many users took advantage of this and posted online, textual reviews without ever experiencing the given products. It seems that Amazon views these reviews as influential and perhaps even critical for fostering trustworthiness in their users. Was their lawsuit a worthwhile activity? Indeed. The overwhelming majority (97%) of users rely on the textual feedback left by previous buyers before proceeding to purchase something from an unknown seller (Pavlou & Dimoka, 2006). They have been found to significantly influence both sales (Chevalier & Mayzlin , 2006) and consumer preferences (Vermeulen & Seegers, 2009), with about 80% of purchases being influenced by a recommendation (Voss, 1984). The more extraordinary the feedback, the greater the impact (Bikhchandandi et al., 1992). Textual feedback impacts prospective consumers because it covers many aspects of the object being reviewed (Duan et al., 2008a) offers evidence of a seller‘s history which is used to predict the seller‘s future behaviour in transactions (Pavlou & Dimoka, 2006).

Online user reviews can influence purchasing decisions and convey end-user experience, in the mobile app domain as well (Hoon, Vasa, Schneider, & Mouzakis, 2012; Vasa, Hoon, Mouzakis, and Noguchi, 2012). Moreover, textual user reviews impact the business performance of hotels in the online hotel booking domain (Ye, Law, & Gu, 2009). The impact of online peer reviews was also examined in the movie box office domain. Duan, Gu, &Whinston (2008b) examined the relationship between online textual reviews and sales. They found that users scanned and only read random samples of the reviews (Duan, et al., 2008b). Moreover, the textual movie reviews were found to both impact and were impacted by movie revenue. A better predictor of movie sales was volume of reviews, which the authors attributed to the spread in user awareness. Duan et al. (2008b) explained this result by the confounding variable of word-of-mouth.

While some online markets (e.g. Amazon, eBay, etc.) utilize user reviews to help future consumers make better decisions on who to buy from, some online retailers go further and include personal profiles with each review (e.g. Sephora, Epinions, etc). Smith et al. (2005) examined the impact of textual review user profiles in a simulated restaurant choice environment (in a controlled laboratory setting). Users had to select a

restaurant they would go to, based on online reviews, where the reviewer user profiles were presented and were trustworthy, untrustworthy, or not available at all. The authors chose unfamiliar cities and restaurants so that users could not rely on or be influenced by their previous experiences or expectations. Smith et al., (2005) found most users relied on the peer recommendation regardless of the person‘s profile. Participants that were not given user reviews relied on any other available cues, such as paid ads. Basically, any info is better than none when choosing a restaurant. However, when given the choice of both user reviews and paid advertising, users relied on the peer reviews. Smith et al., (2005) also found that longer search times neither not produced more accurate information, nor did they influence users more.

Pavlou and Dimoka (2006) argue that nuanced textual messages can significantly impact trust ratings not only when they are left by a neutral party, but also trust is intentionally impacted, such as in the case of marketing (Kim & Benbasat, 2003). This is a big problem in the Apple iOS App Store, where valuable personal information can be stored and profits are in the millions (Chandy & Gu, 2012). Deceitful text reviews give rise to two negative outcomes: trick people into downloading harmful spam with false positive reviews, and normal apps are avoided with false negative reviews (Chandy & Gu, 2012; Vasa, Hoon, Mouzakis, & Noguchi, 2012).

From the above mentioned studies, it is clear that textual expectation setting can influence the actions of consumers. In this thesis, it is investigated that positive or negative texts taint users with expectations that can alter their perception and use of websites (outside of the consumer domain). Thus, we examined if nuanced task descriptions could impart expectation and impact users. To the best of our knowledge, no prior work has been done on textual expectation setting for websites, and no one has examined its impact on the perception of visual appeal or usability. The next section examines related work on verbally implemented expectations, and their impact on people.

Verbal Expectations

The online environment is overloaded with information and yet, not much research has been done on online communication (Smith et al., 2005; this may not be the case for other fields). Given the importance of textual recommendations, feedback, and reviews, one would expect verbal expectations to have been thoroughly investigated but this is not the case, in HCI. The most prominent topic with respect to the impact of verbal communication deals with abuse, which is out of the scope of this thesis. Most research that has been done on word-of-mouth is over 20 years old. For example, Ellison and Fudernberg (1995) investigated two aspects of WOM: (1) given one better and one worse product, will reputation via WOM ensure that the better product is used and (2) between two equals, will WOM influence product choice? The major finding in the Ellison and Fudernberg (1995) paper was that people tended to either conform to or diverge from the information given via WOM. Some participants ignored the WOM information and went with their own experience. However, when the WOM was short

(i.e. little information was transferred), then over a longer period of time, everyone adopted the WOM/common belief. Given unequal payoffs, there were three possible outcomes: (1) diversity, (2) sufficient social learning for conformity towards the better choice (over time), and (3) conformity towards the worse choice. Given equal products, there are three aspects that predict the outcome: (1) only upon interaction with someone from the other choice is one able to switch to that choice, (2) other people‘s experience is regarded as equally important as self-experience, and (3) only current information is relied upon. These results suggest that social learning was taking place and that, over time, the majority of the population would conform to using the best product. The most prominent use of verbal communication used in a face-to-face situation occurs in psychology studies in the form of confederates.

Confederates

A confederate is an individual who is part of the experiment, and usually either acts like a participant or is someone in the background, and often interacts with and influences the participants. Confederates are not monitored by the researchers and are aware of the study‘s true purpose. Their specific roles are defined dependent on the experiment. While confederate use is not commonly found in HCI or usability studies, hundreds of experiments in psychology and sociology have been done using them. Two of the most famous experiments in psychology that used confederates to influence their participants are described here.

The first known study to use a confederate to sway participants was done in the 1950‘s by Solomon Asch. The Asch conformity experiments examine people‘s submission to the zeitgeist of the larger population, and the impact it has on beliefs (Asch, 1951; Asch, 1952; Asch, 1955; Asch, 1956). In the original study, a group of people (seven confederates and one participant) was asked to participate in a visual experiment examining perceptions. The study‘s main goal was to observe how the real participant would react to the confederates. The tasks involved viewing a line and noting its length, then identifying which of three differently sized lines was the same as the first line they had seen. Getting the answer incorrect was impossible, assuming normal vision. Answers were tallied aloud so that everyone heard each other‘s responses, and the real participant was the last to respond. Confederates were told to give the correct response for the first two tasks after which they unanimously switched to giving the obvious wrong answer for the majority of the remaining tasks. The results showed that 25% were not affected by the confederates, only 5% of the participants were entirely persuaded by the confederates, and the remaining 70% conformed on at least one task. Many individual differences were found between the participants who conformed: independence, confidence or lack thereof, desire for conformity, suspicion, doubt in their perceptions of the correct answer, and confusion. For some of the participants who readily conformed for over half of the tasks, the suggestive power of the unanimous confederate vote managed to persuade them into perceiving the incorrect answer as the correct one – unaware that they were incorrect answers, as they revealed

in post-task interviews. Others in the same situation, with lower levels of confidence, thought that they were misinterpreting the stimuli and were sure that the majority was correct. The remaining participants who conformed did so knowingly because they did not want to be the odd ones out.

This iconic study led to a myriad of others in psychology. The method and implications were ground breaking. The voice of the majority could sway the perceptions, thoughts, and actions of others. Another study with similarly legendary results was by Milgram in 1963. Rather than examining the impact of a group of people on a single person, he examined the impact of one person, an authority figure, on another. Milgram (1963) wanted to know what happened in Germany during WWII. Was everyone in the German military obedient to authority figures or were they all guilty of something horrible? Milgrim devised a study with Americans in which there were two confederates and one real participant. One of the confederates was the official _experimenter' in a white lab coat and the other confederate was acting like a participant alongside the real volunteer participant. The experimenter then assigned a role to each of the two participants, seemingly at random, but the confederate volunteer became a _student' while the real volunteer became a _teacher'. The two were separated into two adjacent rooms where they could communicate but not see each other. The student was meant to obey orders from the teacher. The teacher (i.e. the real participant) was given a set of paired words that the student was meant to learn. Subsequently, the teacher gave the first word and the student was meant to identify the second from four possible answers by pressing a button. For every incorrect answer, the teacher had to administer an increasingly more powerful shock. The teacher was told that the student had a heart condition (staged as part of the experiment), and the teacher was given a real, sample shock so they knew the shocks were real. A tape recorder was synchronized with the shock generator and every time the teacher activated the switch, an increasingly distressed recording would play (i.e. controlling the stimuli). Towards the higher voltages, the student (confederate) would also bang on the wall, and at a set point, all noise would cease from the room. If the participant wanted to stop the study, then the experimenter had a set of phrases they were allowed to use, in a specific order. These ranged from –please continue" to –you have no other choice, you must go on" (Milgram, 1963, pg. 374). The experiment ended if the participant wanted to stop the experiment after all the experimenter's verbal probes were said, or if the maximum voltage (a deadly dose) was reached and enabled three times in a row.

To gain an understanding of what the student expectation of teachers was, Milgram (1963) surveyed a small number of faculty and students and asked them to predict what percentage of teachers would harm their students. Participants expected that only 3% of teachers would be able to administer the lethal dosage. Most of the people in this preliminary study thought that people would stop when the confederate asked to be freed. What the results of the actual experiment found shocked not only psychologists, but people around the world.

While the participants were highly stressed and some said that they did not want to continue, an overwhelming 65% of participants went through to the end of the study

and administered the lethal voltage (Milgram, 1963). All participants paused and questioned the experiment; some were under so much stress that in addition to sweating and trembling, some (real) participants had seizures. However, the majority of the participants listened to the authority, who was just a scientist in a white lab coat (Milgram, 1974a). Milgram (1963, 1974) found that ordinary adults, doing their jobs, went to extreme lengths at the bidding of an authority figure – one that was unarmed and unthreatening.

Milgram‘s study has been repeated several times by researchers all over the world. A meta-analysis of these studies showed that the percentage of people who completed the study by administering the lethal dose of electricity ranged from 28% to 91% (Blass, 1999). In fact, a minority of people stand up to authority (Milgrim, 1974b).

Summary. These two studies on conformity, through use of confederates, and the aforementioned studies using texts, all strongly suggest that the influence of textual and verbal communication indeed impact a person‘s thoughts and actions. To the best of our knowledge, no prior work has been done on verbal expectation setting for websites, and no one has examined its impact on the perception of visual appeal or usability. Therefore, in order to help bridge the gap in the current literature, the work in this thesis examined the impact of textually and verbally implemented expectations on visual appeal and perceived and objective usability.

Summary

This research not only furthers our understanding of website usability and visual appeal, it also adds to guidelines and gives developers more insight in how to develop a website that will meet users‘ expectations of and strike balances between expectations, usability, and visual appeal. In addition, users should benefit from encountering poorer designs less frequently. In order to arrive at a better understanding of the relationship between usability and visual appeal, it is necessary to select a theoretical framework within which the phenomena under investigation can be discussed. Therefore, the next chapter examines the two major theoretical constructs that will inform the approach to the experimental research in this thesis. The concept of mental models is discussed first, followed by a discussion of the theory of cognitive dissonance, both of which are likely to influence the way people perceive the relationship between usability and visual appeal.

Chapter 3. Theoretical Concepts

The purpose of this chapter is to explain the theoretical concepts used to help guide the research and formulate the hypotheses in this thesis. The understanding of theory can justify the reasoning behind the research being done, determine the data collection approach, rationalize the data analysis, and it can help understand the findings. Theories are used to avoid arbitrary decisions and the findings are usually used to support, refine, or toss out the theory, in order to perpetuate research in a cohesive and generalizable manner. By definition, theories predict hypotheses that can be empirically tested and falsified (de Jong, 2014).

In this thesis, theoretical concepts attempted to: explain usability and visual appeal; explain the concept of expectations; create the research hypotheses, and to help understand the findings. Thus, theories for usability are discussed first, followed by theories for visual appeal.

As mentioned in the previous chapter, expectations are defined as beliefs about what is about to occur, also known as predictions and surmises. Mental models were used to explain the formation of expectations. Therefore, mental models are examined next in this chapter. The next few sections examine theories that strive to explain both congruent and incongruent cases of expectations and reality, including the Consistency Theory, the Counter-Attitudinal Theory, the Cognitive Dissonance Theory, the Self-Perception Theory, and the Post-Decision Dissonance Theory. The theory of Cognitive Dissonance was used to explain how people respond to new information that does not align with their mental models. Thus, the following section discusses the Cognitive Dissonance Theory. The research hypotheses in this thesis were based on Cognitive Dissonance, so this chapter concludes with the outline of these.

Usability Theories

Usability research was not originally embedded in a theoretical framework. In 1983, Card, Moran, and Newel had the goal to establish HCI as a strand of ~~"a~~pplied psychology concerned with the human users of interactive computer systems" (p.vii). However, research in psychology is heavily driven by theories. As a result, Card et al. (1983) strived to make HCI ~~"t~~heory-based, in the sense of articulating a mechanism underlying the observed phenomena" (p.13). They came up with an information processing model called the Model Human Processor, used to predict user performance and meant to help developers create better interfaces. However, computer scientists did not regard psychological theories highly. Thus, Newell and Card (1985) applied engineering theories of users (e.g. task analysis and calculations; Baecker, Grudin, Buxton, & Greenberg, 1995) to the field. The resulting model was GOMS, which examines: Goals, Operators (i.e. what users can do with the interface given their skills), Methods (i.e. approaches to achieving the goal), and Selection Rules (i.e. choosing a method; Card et al., 1983). GOMS attempts to predict expert users' approaches to

achieving goals and how long these would take. However, GOMS is not a _theory‘ because it does not explain the phenomenon of usability and it was later referred to as a _model‘ (Baecker et al., 1995). The model also had several flaws, some of which include that the model did not account for novice users, did not mimic actual use that entails errors (i.e. it only examined optimal performance), and does not address user preference (Olson & Olson, 1990). For these reasons, GOMS was not considered for use in this thesis.

In the decades since Card and colleagues‘ (1983) efforts to include theories in HCI, researchers in the field have rarely used theories to frame their approaches or findings, aiming for practical implications instead (de Jong, 2014). In fact, there are no widely accepted, holistic, usability theories (de Jong, 2014). Nielsen (1993) mentions a _usability theory‘ but described it as the approach of including usability principles in the design and testing of interfaces, which did not include an explanation of the concept, nor did it offer any predictions. Nielsen‘s widely used heuristics (1993) are used for usability assessment but they do not amount to a theory. In addition, he opposes the use of visual appeal, in the form of images and videos, stating that it distracts from the usability of a website (Nielsen, 2000). However, this idea is outdated since one of his main reasons was that Flash slows down downloading times. Internet speeds have become much faster in the last ten years and downloading times are no longer a concern. In addition, visual appeal is another concept central to this thesis and the dismissal of it in a theory or model make that model inapplicable to the research in this thesis.

As was illustrated in the literature review, usability is a complex concept and has evolved over time – the concept might not be easily encompassed into a single theory (de Jong, 2014). One related theory is the Technology Acceptance Model (TAM). TAM attempts to predict users‘ acceptance and use of a new system (Davis, 1989), without examining perceived and objective usability and the consequences of these (de Jong, 2014). Since both types of usability are central to this thesis, the model was not considered as it does not overlap with the scope of this thesis. Moreover, the goal of this thesis was not concerned with system acceptance. Therefore, no explicit usability theories were used to guide the concept of usability. The next section examines a similar situation with visual appeal.

Theories of Visual Appeal

Most of the theories that pertain to visual appeal are aesthetic theories. Aesthetic theories are used to evaluate art (Kairies, 2012). Some are based in mathematics and deal with the measurement and creation of something beautiful (e.g. symmetry, the golden ratio; Livio, 2002; Pacioli, 1509). Photographers and architects have their own principles for making aesthetic pictures and structures, including the rule of thirds (Meech, 2007) and the diagonal rule (Arnheimm, 1954).

Painters still have their own aesthetic theories, each comes with different type of artwork, and depend on the goal of the artwork (e.g. Imitationalism, Formalism, and

Emotionalism; Kairies, 2012). Artists painting with Imitationalism in mind try to create the most realistic paintings (i.e. follow the laws of gravity, proportion, lighting, colour, etc.). The theory of Imitationalism attempts to reproduce real objects and settings. A classic example is a yellowing green field with blue skies and white clouds above the field. Alternatively, another example of an Immitationalist picture would be a dark room with a small table in it; the table has a plate of grapes that are lit on side by a candle, and shadows hit the table on the other side. Nearly the complete opposite of Imitationalism is Formalism, which guides artists to create more abstract pieces. Artists painting formalistic paintings do not need to follow any laws of logic, nor do they paint recognizable things or settings. An example of a Formalist painting includes geometric shapes and primary colours, such as three vertical lines and four horizontal ones with red, blue, and yellow where the lines intersect to form squares. Following the Emotionalism view of beauty, an artist would strive to create a painting that can evoke a certain emotion. An example would be a painting of a man, sitting on a chair, hunched over with his head in his hands – trying to evoke sadness and despair. Another example would be a painting of mother and a child, laughing with a birthday cake – evoking happiness. Yet, none of these theories offer an explanation to the influence visual appeal would have over a different subject, such as usability.

Another theory related to visual complexity comes from Berlyne (1974), stating that a moderately complex stimulus is the most pleasing as too simple stimuli are boring and too complex stimuli cause stress. However, a precise definition and validity of measurement of visual complexity is lacking in the website evaluation domain. Psychologists in cognitive, neuro-psychology, evolutionists, and cultural-study fields have their own views for aesthetics: how it impacts the brain and how culture and evolution may influence the appraisal of it (Dutton, 2002). Nevertheless, the purpose of the thesis was not to understand what makes a website beautiful or how to create a visually appealing website (or painting or building). Instead, the purpose was to examine the effect of expectation on the evaluation of visual appeal. This thesis also did not use any theories for visual appeal as none added to the understanding of the concepts in this thesis.

While no theories were used for visual appeal, there were six theories relating to expectations that were examined for applicability and use in this thesis. They are discussed in the next section.

Theories of Expectations

Several theories on expectations are examined in this section. First, mental models are discussed because they were used to explain the formation of expectations. Then, several theories examining how beliefs and attitudes change based on our behaviours, or the behaviours of others, and are examined. These include theories that strive to explain both congruent and incongruent expectation-reality cases, including the Consistency Theory, the Counter-Attitudinal Theory, the Cognitive Dissonance Theory, the Self-Perception Theory, and the Post-Decision Dissonance Theory.

Mental Models

A mental model is a ‚subjective representation of external reality" (Toffler, 1970, p. 139). Mental models are a combination of an individual's subjective perceptions, concepts, and ideas (Sinreich, Gopher, Ben-Barak, Marmur, & Menchel, 2005). Mental models are used when interacting with the environment because they allow people to understand and remember relationships between objects in the environment (Rouse & Morris, 1986). This also means that mental models are used to create expectations of what is likely to occur in specific or previously encountered environments (Rouse & Morris, 1986). Craik (1952) was the first to refer to a mental model, explaining it as an ‚internal model of reality – this working model – [*that*] enables us to predict events which have not yet occurred in the physical world, a process which saves time, expense, and even life" (p. 82). In other words, mental models are produced as a function of experience, from which predictions about future events can be made. Mental models have also been referred to as ‚schemata' (Bartlett, 1932), which are described as cognitive look-up tables of previous experiences used to understand present events. Since mental models form and evolve with experience (Johnson-Laird, 1983), they can be induced experimentally (Gentner & Gentner, 1983). However, we have limited insight into our personal mental processes (Fischhoff, 1975; Nisbett & Wilson, 1977), making free recall of them during experimentation a hard task. Therefore, mental models are abstract constructs that are most readily inferred from observing verbal and/or nonverbal behaviour.

Scripts are in several ways similar to mental models. Scripts are contextualized pieces of information in the form of sequences of events that help an individual make sense of a situation, especially when there is some information missing and should be inferred (Schank & Abelson, 1977). For example, when seeing a man sitting in a restaurant being approached by a waitress, one would not be surprised to see that the waitress comes back with food, and would know that the man ordered it. Compared to a mental model, which is a general idea of what something is, its function and role in a context, scripts are procedural. Since the research in this thesis focused on the expectations of websites, which are objects and not events or procedures, this research used mental models as the basis of our expectations.

Mental models have practical applications in an HCI context (Ben-Ari, & Yeshno, 2006). Not only do websites designed to match users' mental models affect interactions (Bargas-Avila et al., 2007), they also increase memorability of item location on the webpage (Oulasvirta, 2004). In fact, user interface design guidelines exist that indicate the importance of designing according to users mental models (Apple, 1996; IBM, 2008).

In addition, Roth and colleagues (2010) examined participants' expectations for three specific genres: an online shop, a news portal, and a company webpage. Participants constructed examples of these three genres by choosing from a set of given elements and placing them on a website template. Roth et al. (2010) found that the participants had a common understanding of which elements belong to which genre of

websites. For example, using contexts from an experimental study that extends from this present research proposal, a city council website would not contain a shopping cart but a tourism webpage would. However, mental models evolve as a function of exposure to new stimuli. As web design develops and improves with time, and as users gain experience with different website genres, user expectations of element locations change as well (Shaik & Lenz, 2006; Roth et al. (2010). In this thesis, it is assumed that mental models assist users in making predictions, in the form of expectations. Prior to assessing the impact of expectation on the relationship between usability and visual appeal, it is essential to gain an understanding of what participants' mental models and expectations are of the experimental contexts, i.e. city council and tourism websites, and how they differ. In addition to assessing the perceived usability and visual appeal of the sample websites, gaining an understanding of the users' mental models was done as part of the first preliminary study.

This concept of mental models is associated with prototypicality (Lindgaard, Dudek, and Fraser, in press). In HCI, prototypicality is defined as "the amount to which an object is representative of a class of objects" (Leder, 2004, p. 496). According to Rosch (1975), objects are categorized according to the most representative prototypes of the categories. Prototypicality is the representativeness of an object (Leder, 2004). The more representative an object is of its class of objects, the more preferable it will be, because it conforms to expectations, or mental models (Whitfield, 1983). When people interact with objects, they develop internal representations of the objects (e.g. Rouse & Morris, 1986; Sinreich, Gopher, Ben-Barak, Marmor, & Lahat, 2005), called mental models. Mental models are a combination of an individual's subjective perceptions, concepts, and ideas (Sinreich, Gopher, Ben-Barak, Marmor, & Lahat, 2005). They are used when interacting with the environment because mental models allow people to understand and remember relationships between objects in the environment, and they also create expectations of what is likely to occur (Rouse & Morris, 1986). Prototypical objects are processed fluently, and are thus likely to elicit positive affect (Thielsch & Hirsschfeld, 2010; Winkielman et al., 2006). The processing fluency theory states that the more fluently an object can be processed, the more positive the aesthetic response will be to that object (Reber et al., 2004). Norman (2004) also hypothesized this with the theory of emotion according to which very briefly shown complex stimuli can only be reacted to emotionally, as no time is allowed for cognitive assessment. However, Thurgood, Whitfield, and Patterson (2011) have found that that people can recognize images flashed to them at less than one millisecond, suggesting that cognition occurs far faster and may indeed occur before or together with emotion.

In Summary. Mental models are an individual's summary of experiences, moulded into schemas that are used to set expectations of what is likely to occur (Rouse & Morris, 1986). In this thesis, the goal is to gain a better understanding of the role of expectation on usability and visual appeal. Therefore, mental models were used as an explanatory framework because expectations are grounded in them. In order to observe the impact of expectation on usability and visual appeal, we manipulated expectations

of both visual appeal and usability to be either congruent or incongruent with the actual website levels. Incongruent expectations were manipulated to create cognitive dissonance needed in order to examine the impact of expectations. Thus, the next few sections examine theories that strive to explain both congruent and incongruent cases, including the Consistency Theory, the Counter-Attitudinal Theory, the Cognitive Dissonance Theory, the Self-Perception Theory, and the Post-Decision Dissonance Theory.

Other Relevant Theories

Consistency Theory

When all information in a person‘s internal (thoughts, values, etc.) and external (other people and things we interact with) environments are all in agreement, it is said that the individual is in a positive mindset (i.e. no stress; Festinger, 1957). The theory states that people strive to maintain a maximum possible level of consistency in their internal and external environments, to limit stress. Should a disagreement arise, causing stress, people strive to diminish the inconsistency as a means of homeostasis, to return to the positive state.

Thus, when expectations are the same as the website usability and visual appeal, then there should be no stress and nothing unusual should happen and the true effect of expectations will not be evident. To learn more about the influence of expectations, one must examine the impact they have when they are incongruent with real levels of visual appeal and expectation. Yet, the consistency theory does not offer an explanation about what happens when there is an inconsistency. An extension of this theory is the cognitive dissonance theory, which explains what happens when there is an inconsistency in our opinions, behaviours, and/or environment.

As previously mentioned, Ludden et al. (2012)‘s paper (that examined what happened when the expectation created by the appearance of the product was incongruent with the feel of the product), found that the mismatch between the object‘s expected and real tactile experience resulted in surprise which was then followed by either positive or negative emotions. This phenomena could be explained by the cognitive dissonance theory, since the ‗surprise‘ described by the authors may have in fact been a type of stress caused by cognitive dissonance. In this thesis, the cognitive dissonance theory was similarly used, to base incongruent expectation manipulations. Incongruent expectations were included since they can elicit this ‗surprise‘ which reveals a reaction caused by the predisposition (i.e. the expectation – if it works) and reality (i.e. the actual website). Thus, the impact of expectations would be more readily evident because it would be the only differing factor.

Cognitive Dissonance Theory

Cognitive dissonance is a disagreement of information in an individual‘s thoughts, which may cause stress. The disagreement can occur between our thoughts (i.e.

memories, understandings, opinions, beliefs, etc), our behaviours, and our environments. In 1957, Festinger also proposed the theory of cognitive dissonance, which predicts that when an individual‘s thoughts are relevant but inconsistent or conflicting, a state of dissonance occurs. For example, you think your friend is smart and polite then those two thoughts are in agreement, or in other words, concurrent cognitions are called consonant. However, you just found out that your friend is an ex-criminal then this new piece of information is inconsistent, or in other words dissonant, with the previous two and would cause conflict. Therefore, in order for dissonance to arise, there must be a conflict between a person‘s prior beliefs and a situation in which those beliefs are challenged. Festinger referred to it as ‗cognitive‘ because it involves a level of awareness that goes beyond a sensory perception: recognition of the conflict requires cognitive insight. There are two definitions of cognitive dissonance used in the literature: a cognitive inconsistency and a negative emotional state caused by the conflicting cognitions. Dissonance can be experienced on a continuum, where it ―is greater when the number and importance of dissonant cognitions is higher and the number and importance of consonant cognitions is lower‖ (Harmon-Jones, Amodio, & Harmon-Jones, 2009, pg. 122). According to some researchers (Sakai 1999; Shultz & Lepper, 1999), the magnitude of dissonance to a particular topic can be measured by a formula that generalizes to D/(D+C), where D is the number of dissonant cognitions, and C is the number of consonant cognitions.

While it is relatively easy to predict an individual‘s actions when there is no disagreement, it is a lot harder to do so when there is dissonance. For example, if you thought that filling out a particular form would be complicated, you would not be surprised to see that it indeed is complicated and will most likely continue to fill out the form. However, if you expected that the form would be easy to fill in, and found out that it was indeed more difficult than expected, the prediction is less certain. Would you continue filling it out or leave it and try to get a friend to do it for you? The understanding of what occurs during such cognitive dissonance provides a better understanding of the degree to which expectation influences the relationship between usability and visual appeal.

When dissonance occurs, an individual strives to achieve consonance by means of reducing the inconsistency. Reducing dissonance is done in four ways: adding or increasing the importance of consonant cognitions, or taking away or decreasing the importance of dissonant cognitions. An example of adding a consonant cognition to the example used above relating to a friend who is an ex-criminal would be adding the thought that he is also a good-looking individual, thereby adding more consonant cognitions to increase the odds against the dissonant cognition. An example of increasing the importance of the consonant cognitions would be to stress the facts that he is smart and polite by thinking that he‘s successful at his job and that his family loves him. In the same example, taking away a dissonant cognition would be to think that he may have been wrongfully convicted. Decreasing the importance of a dissonant cognition in this example would be to think less of the crime he committed, reducing the possible act to an unpaid parking ticket. Dissonance reduction is measurable by noting

attitude change, usually in the direction of the cognition most resistant to change (Harmon-Jones et al., 2009). The cognition that is most resistant to change is the one associated with the most recent behaviour. In other words, a person‘s attitudes are most likely to change to concur with that person‘s most recent actions, so as to avoid further dissonance.

In this thesis, the cognitive dissonance theory was used to create dissonance in an individual by manipulating the expectations of websites in the incongruent conditions. Incongruent conditions occurred in Chapters 6 and 7, when the expectation of visual appeal and usability did not match the website‘s actual usability and visual appeal levels. For example, in Chapter 6, one of the conditions had high visual appeal and usability expectations, yet the actual website was hard to use and ugly. The reaction to this mismatch gave insight as to the impact of expectations on the perception and use of websites. The more astonishing the expectation, the greater the impact (Bikhchandandi et al., 1992), suggesting that the expectation levels needed to be really positive, or really negative.

Counter-Attitudinal Advocacy

Evidence that an individual‘s opinions and actions change is found in the Counter-Attitudinal Advocacy Theory (Zanna & Carlsmith, 1959). This theory states that if an individual publically says or supports (in some other way such as through an action) something that is incongruent to their prior beliefs, then their beliefs will change in accordance to what they said/did. This theory is said to be effective because the action that catalyzed the change was public and thus, it is harder to deny its occurrence (Aronson, Wilson, & Akert, 2010). For example, when participants were asked to do a boring task and asked to say, aloud, that the task was enjoyable, those that were paid next to nothing later revealed that they enjoyed the task (Festinger & Carlsmith, 1959). Those that were paid more did not change their opinions.

Much like the Cognitive Dissonance Theory, this occurs in order to reduce the dissonance produced by the incongruent behaviour. The difference between these two theories is miniscule: Counter-Attitudinal Advocacy suggests that there is an overt behaviour that the individual does themselves that acts as a catalyst for the dissonance and behaviour change. Cognitive Dissonance does not state that the dissonance must occur as an act of the own individual‘s behaviour. Therefore, since the expectation implementation in this thesis was an external factor, the Cognitive Dissonance Theory was used instead.

Self-Perception Theory

An alternative explanation to the reaction to stress caused by cognitive dissonance is the self-perception theory. The self-perception theory states that people examine their own actions in order to determine how they feel about something, particularly when they are forced to do something they are not fond of (Festinger & Carlsmith, 1959;

Bem, 1972). For example, according to the cognitive dissonance theory, people who are paid very little to lie would not like the experience because the lie was not justified (Festinger & Carlsmith, 1959). In contrast, according to the self-perception theory, the same scenario suggests that people like to lie because the reward was too small a motivation on its own, subconsciously suggesting that they must have lied because they enjoyed it (Festinger & Carlsmith, 1959). However, this theory assumes two things: (1) that people are stressed by the dissonance, and (2) that the dissonance was caused by their own behaviours.

The studies in this thesis did not ask participants to do anything that they normally would not, and the dissonance was not caused by their actions. Therefore, this theory was not used in this thesis.

Post-Decision Dissonance Theory

The Post-Decision Dissonance Theory states that the uncertainty we experience after having made a decision (or done something) arises from the possibility that the decision we made was the wrong one (Brehm, 1956). To reduce the dissonance, people will change their opinions on the matter to agree with the action (Knox & Inkster, 1968). For example, after betting on a horse to win a race, people thought that the horse they chose was more likely to win (Knox & Inkster, 1968). Similarly, comparing the ratings of household appliance appeal pre- and post- ownership, the appeal of the appliance increased after it was given to consumers as a gift (Brehm, 1956). Therefore, based on this theory, expectations may also impact the perception of visual appeal and usability after system use. For example, if participants thought that visual appeal was low pre-use, and it indeed was low, then the ratings might be even lower post-use. In addition, the Post-Decision Dissonance Theory might predict that system use would not necessarily change participants‘ opinions in the incongruent conditions (i.e. the expectation does not match the actual website levels). In this case, expectations would still impact ratings post-use and these ratings would remain the same (i.e. post-decision dissonance would combat the experience of having a different system to what was expected). Participants were not asked about their confidence in their ratings and thus, their uncertainty cannot be commented on in this thesis.

Summary of Theories Used. Several theoretical concepts were examined above that were used to help guide the research. No theories were used to explain the concepts of visual appeal and usability. Since we induced different expectations of the usability and visual appeal levels in this thesis, we needed theories that would help define and explain expectations. Therefore, mental models were used to describe the formation of expectations. Cognitive dissonance fulfilled the need to study incongruence between induced expectations and the stimuli, and was thus used to guide this research. The next section describes the research hypotheses, derived from cognitive dissonance.

Research Hypotheses

Based on the cognitive dissonance theory, there were four research hypotheses developed to help in answering the research questions from chapter 1. Cognitive dissonance is a disagreement of information causing stress, and people strive to reduce the stress by changing the way they think about the issue. The cognition that is most resistant to change is typically the one associated with the most recent behaviour/event (Harmon-Jones et al., 2009). In the case of this thesis, the most recent behaviour would be experiencing the expectation which would vary depending on the experimental condition. Therefore, if expectations influence visual appeal and usability, then participants should agree to the expectation given, and the perceived variables should be reported as either higher or lower than the control condition, in accordance with the expectation level. By this logic, the following was hypothesized.

The *first research hypothesis* states that when expectations of both or either of the two variables (i.e. visual appeal or usability) are set to be high, then participants will perceive and rate the appropriate variable(s) to be higher. If the expectations were set to be high for both variables (i.e. pretty and easy to use), for example, then participants would rate both visual appeal and usability even higher than their actual website levels (i.e. higher than the control group). If only one variable was said to be high and the other to be low (e.g. pretty but hard to use), then the one with the high expectation would still be perceived to be higher than the website‘s actual level. Higher ratings were expected because participants would be swayed to increase their ratings to reduce the inconsistency in order achieve consonance.

The *second research hypothesis* uses the same logic as the first hypothesis and states that when expectations of both or either of the two variables are set to be low, then participants will perceive and rate the appropriate variable(s) to be lower than the actual website levels. Lower ratings are expected because participants decrease their ratings to reduce the cognitive dissonance and agree with the low expectations. For example, if the expectations were set to be low for both variables, then participants would rate both visual appeal and usability lower than their actual website levels. If only one variable was said to be low and the other to be high (e.g. ugly but easy), then the one with the low expectation would still be perceived to be lower than the website‘s actual level. Lower ratings were expected because participants would be swayed to decrease their ratings to reduce the cognitive dissonance in order achieve consonance.

The last two research hypothesis state that participant performance (in the form of objective usability measures) is also affected by expectations. In particular, the *third hypothesis* states that participants will perform better (i.e. efficiently and effectively) when the expectations are set to be higher. The cognitive dissonance theory states that the most recent action tends to be the most prevalent disposition (especially when there was no prior disposition) and therefore, people will act according to it. When the expectation is high, the participants may be affected accordingly and not only perceive it to be better but also experience the website by using it with greater efficiency and effectiveness (i.e. objective measures of usability). Similarly to the third hypotheses, *the*

fourth hypothesis states that participants will perform worse (i.e. struggle more) when the expectations are set to be lower. These are hypothesised because participants who perceive it to be either easier or harder to use may reflect their perceptions in how they use the website as a confirmation bias.

Summary

The goal in this thesis was to gain a better understanding of the impact of expectation on usability and visual appeal. Theories and theoretical concepts were used to help explain the origin of expectations and how they can impact us. Specifically, mental models were defined in this chapter as cognitive summaries of our experiences that give us expectations of that context (Rouse & Morris, 1986). Thus, mental models were used to explain the origin of expectations. Moreover, expectations are predictions of what is likely to occur. Congruent expectations and actual usability and visual appeal levels would not be surprising and thus the true impact of the expectations would not be clear. Thus, in order to observe the impact of expectations on usability and visual appeal, we manipulated expectations of both visual appeal and usability to be either congruent or incongruent with the actual website levels. Incongruent expectations were manipulated to create cognitive dissonance needed in order to examine the impact of expectations. From the theories discussed, the theory of cognitive dissonance fulfilled the need to study incongruence between induced expectations and the stimuli, and was thus used to guide this research. Using the theory of cognitive dissonance, four research hypotheses were proposed. In summary, these were:

(1) When expectations of both or either of the two variables (i.e. visual appeal or usability) are set to be high, then participants will perceive and rate the appropriate variable(s) to be higher.

(2) When expectations of both or either of the two variables are set to be low, then participants will perceive and rate the appropriate variable(s) to be lower than the actual website levels.

(3) Participants will perform better (i.e. faster and with less error) when the expectations are set to be higher.

(4) Participants will perform worse (i.e. slower and with more errors) when the expectations are set to be lower.

The next chapter examines the methods by which these four hypotheses were examined. It details the approach taken and outlines the experimental studies in this thesis.

Chapter 4. General Method

In order to find answers for the research questions and hypotheses in this thesis, this chapter discusses the general approach taken to examine the impact of expectations on visual appeal and usability on a website. As mentioned in the introduction, visual appeal is the cognitive appraisal of the website‘s aesthetic appearance. Usability was defined according to the ISO definition, which in short is ease of use.

While this research is based on mental models and distributed cognition, both theoretical constructs, it took a more practical approach, empirically examining the effect of expectations on usability and visual appeal. A series of controlled laboratory studies was performed that include observation of participant behaviour, brief semi-structured interviews for participant feedback, and questionnaires in the form of usability and visual appeal scales. The data analysis was both qualitative and quantitative. Expectations, usability, and visual appeal levels were all controlled and manipulated according to the experimental conditions in each study. In addition, the laboratory environment was controlled and access to the usability lab was limited to only the participant, confederate (if applicable), and researcher.

This chapter starts with a discussion on data collection methods. This is followed by an explanation of the type of data used. Specifically, qualitative and quantitative analyses are examined to inspect what each yields and how these were applied in this thesis. Since several usability tests were done in this thesis, the next section examines usability evaluation approaches, including expert- and user-based testing. After the usability approach follows the usability metrics used. Then, visual appeal evaluation is discussed. Following this, participant selection is discussed. Next, the statistical approach is elaborated on. Finally, an outline of the studies that compose of this thesis is given as an explanation on how the research questions were investigated.

Data Collection Methods

There were several options on how to collect the data necessary to analyse the impact of expectations on visual appeal and usability of a website. Given that the subject domain is the interaction of humans with websites, the most appropriate data collection methods would need to correspond with typical HCI methodology. Typical HCI data collection methods include observation, questionnaires, interviews, and experimentation (Sharp, Rogers, & Preece, 2007). Each of these concepts is discussed below.

Observations

As an HCI method, observation is defined to be a surveillance of a user in a specific context at the beginning of the project, or if the interface is designed, then the observation is done with the user to inspect the success of the prototype (Sharp et al.,

2007). Observation can be direct or indirect. Direct observation includes the researchers being next to the phenomena being studied, and examining it in real time. The session may or may not be recorded and notes are generally taken. In contrast, indirect observation usually occurs when the researcher is not in the same area as the phenomena or the observations are happening at a later date, if the phenomenon was recorded. Direct observation allows for unobstructed access to the phenomena, and in the case of this thesis, to the user as they are interacting with the system. However, it can be considered intrusive to participants and researchers can influence participant behaviour by _hovering' around them (Mayo, 1933, Crystal, 2003). Indirect observation is less intrusive but depending on what areas are being recorded or can be seen by the researchers, some major aspects can be obscured and missed from the analysis. Either form of observation can occur in the field (i.e. in situ) or in a controlled laboratory. Observations of in situ phenomena would often be more suited for exploratory research. Simply observing people's natural reactions to a website would not yield enough information to determine with certainty that one aspect influenced another. Thus, while all of the variations to observational methods can yield important results, the results of observation alone will not be able to indicate if expectations influence usability and visual appeal, given that the purpose of this thesis is not exploratory. We are looking for more insight and specific outcomes where some variables would have to be controlled for, which goes against purely observing a phenomenon in the field or even in a laboratory. The method applied in this thesis needed to be more structured and allow the researchers to gain access to specific concepts and perceptions, and not necessarily observe participants without a specific goal.

Questionnaires

Questionnaires are also frequently used as a data collection method. They involve a list of questions on paper or online that the participant needs to answer. Questionnaire length can vary and questions can be closed or open-ended, objective or subjective. Closed questions are typically more structured and easily compared between participants because the answers are limited and participants usually need to tick the boxes that apply or select the radio button that corresponds to their answer. Closed questions can be created to be objective. For example, selecting which age bracket you belong in. Open-ended questions are more subjective in nature, as they allow for a longer, more descriptive response. Open and closed ended questions are examined on a case-by-case basis to enable the relevant information required to be extrapolated.

In HCI, many usability questionnaire questions have closed-ended questions in the form of Likert scales in which participants are asked to judge various aspects of the interface (e.g. Garcia, 2013; Brooke, 1986; 1996). In Likert scales, all items are assumed to be equally important (van Alphen, Halfens, Hasman, & Imbos, 1994). They may also include open-ended questions to allow for the user to express their opinion on the interface and provide feedback on any outstanding aspects that were missed in the Likert scales. Questionnaires are often mailed (either by courier or email) if they are not

administered in person. While questionnaires are quick to administer, this method would be more suited towards acquiring general information as it does not allow for proper control of website usability, visual appeal, or expectations.

Interviews

Interviews, in general terms, are questionnaires administered verbally with the participant. They are often face-to-face, and can be structured, semi-structured, or unstructured. A structured interview entails a predetermined set of questions that the interviewer asks the interviewee, with little room for further enquiry by asking a sub-question that was not pre-set. A semi-structured interview also entails a predetermined set of questions but it allows for further enquiry via sub-questions in order to gain more information or clarification from a participant. An unstructured interview has no predetermined questions and appears as more of a conversation between two people where the interviewer asks ad-hoc questions, depending on what the interviewee says.

In any of these three methods, the participant is asked to recall details about the interface, however this may be inaccurate due to the failure of the participant to remember exact details or unwillingness to reveal the truth. For HCI studies where the interface testing is usually impersonal, it is more likely that if a participant responds incorrectly, that it is due to their inability to remember rather than that they are lying. In any case, and as was the case with questionnaires and observation, the data collected from interviews does not allow for any certainty in which factors influenced the participant‘s responses. Thus, the experimental data collection method is discussed next.

Experiments

The experimental method attempts to show causality – that one variable impacts another. In order to do this, participants are randomly assigned to conditions in which all factors are controlled. One of the conditions does not receive the experimental treatment for purposes of comparison. Statistical tests are done to determine if the dependent variables indeed influence the dependent variables.

Experimental tests are typically done in a laboratory or controlled environment. A recent study examining the effects of visual appeal on usability found that the physical context of use (informal: home vs. formal: office) did not affect the relationship or judgement of visual appeal or usability (Sonderegger, Uebelbacher, Pugliese, & Sauer, 2014). Hence, a formal setting, such as a computer laboratory or classroom on a university campus, would not alter participants‘ judgment when assessing a website‘s usability or visual appeal.

Moreover, experimental studies tend to aim at furthering theory rather than at replicating reality (Mook, 1983; Plot, 1991). Some experiments are created to examine what happens under certain circumstances, such as the confederate studies done by Asch and Milgram mentioned in the literature review. Such experiments attempt to explain phenomena that occur in real life but do not copy the environment (Druckman

& Kam, 2009). An artificial laboratory setting would not impact the generalizability of the study, as it would help examine a certain aspect of the phenomena in greater detail, without confounding variables.

Thus, in this thesis, an artificial/laboratory setting was used to control and eliminate confounding variables, as expectations are not tangible and need to be inferred. Usability and visual appeal were manipulated to be bad, providing participants no evidence that they were good. The laboratory setting created conditions in which the only factor that could influence their decision was expectation. This is the approach that was necessary in this thesis in order to examine the impact of expectations on visual appeal and usability in a web domain. However, to gain the most information from participants, this thesis includes observation, logging of activity data during the controlled experimental testing, and the completion of questionnaires to get ratings for visual appeal and usability. This process was followed by short semi-structured interviews at the end of each participant‘s sessions to probe for more detailed feedback.

Longitudinal vs. cross-sectional studies

Observations and experiments can be one-offs (i.e. cross-sectional studies: a study done at a single point in time) or they can be longitudinal studies. The work in this thesis was not a longitudinal study as repeated exposure develops mental models and we would lose control over the expectations if we allowed users to develop their own. Literature on first impressions suggests that opinions do not drastically change even with repeated exposure (e.g. Staw & Hoang, 1995) and that users decide on whether they like something in milliseconds (Lindgaard et al., 2006). The expectation effect would differ with continuously forced use of a disliked website, altering user opinion. This study was not designed to get accurate usability measures. It was to see if expectations impacted the perceptions and use of a website.

Data Types Collected

In any of the data collection methods discussed above, both qualitative and quantitative data can be collected. Qualitative data relies on descriptions and justifications whereas quantitative data is numerical and can be statistically computed. Thus, on one hand, qualitative analysis can usually detect reasons for a phenomenon or it can give rise to explanations for why something did not occur where it should have. On the other hand, quantitative data gives a degree of certainty that the phenomena being observed did not occur by chance. Therefore, the research in this thesis used both qualitative and quantitative analysis methods to determine what was happening with regards to expectations, usability, and visual appeal, why it was happening, and if it was definitely happening due to the impact of expectations.

The next section describes the measures examined and used for measuring usability. Specifically, once a website was chosen for use in this thesis (Chapter 5, Preliminary Study 1), the website needed to be examined for its current and actual

usability level. In addition, participants in the Main Studies (Chapters 6 and 7) interacted with the website via usability test. Therefore, the next section examines usability testing methods, and justifies the methods used in this thesis, since more than one was applied.

Usability Evaluation Methods

Over the last thirty years, the exponential growth in use of household computers, homemade webpages, and numerous handheld devices (among other technologies) have increased the need to test the usability of these products to increase user efficiency, effectiveness, and satisfaction. In turn, several usability evaluation methods have been developed, which for the purpose of this thesis is split into two main categories: user- and expert- based evaluation. The research in this thesis used both expert- and user-based usability evaluation methods. Once the Gold Coast city council website was selected, three experts evaluated the usability of it to get an understanding of the usability level. This was then followed by user-based usability testing to confirm the experts‘ findings. The following examines both expert- and user- based usability evaluation methods, respectively. The most widely used methods are discussed and the methods chosen for use in this thesis are justified and elaborated on.

It is important to note that these evaluation methods are usually meant for finding areas where the interface could be improved, at different stages in the development of a system. This was not the case in this thesis, as the website was live and we first evaluated it to gain an understanding of its usability level, and then it was manipulated to create four different versions of the website (high/low usability and high/low visual appeal), at which point it was evaluated again to see that the manipulations were properly done. The purpose of the usability testing was therefore not _what can we fix to make it better‘. Instead, it was _how usable is it?‘ and _is it as usable as we need it to be?‘

Expert-based usability evaluation has been successfully applied to many areas of research and development, including the game development community (Choi, 2009). Experts are knowledgeable in usability testing, and are also referred to as evaluators and researchers. In this type of evaluation, usability experts discover usability problems by inspecting the existing user interface or prototype with a set of questions, guidelines, or heuristics (Jaspers, 2009). Expert-based usability evaluation methods are usually done when there is not enough time for user-based testing, or when users are not available for testing. In addition, these tests are sometimes conducted to identify major usability problems before user-based tests are done (Petrie & Bevan, 2009). Expert-based methods consist of guideline review, Heuristic Evaluation (HE), usability inspection, consistency inspection, and walkthroughs (Maguire, 2001; Gray & Salzman, 1998; Nielsen, 1993; Nielsen, 1994). The most widely used expert-based methods are HE and Cognitive Walkthrough (Polson, Lewis, Rieman, and Wharton, 1992), and the next two sections examine them, respectively.

Heuristic Evaluation

HE has been referred to as expert evaluation (Seffah & Metzker, 2009) and it is considered to be the most popular expert-based method (Nielsen, 1994). Usually, the evaluators are HCI experts, but novices can also be involved with this testing. HE is conducted by following a set of pre-defined, recognized set of usability principles (the heuristics) as reference for potential usability problems (Jaspers 2009; Seffah & Metzker, 2009).

There are a few widely recognized principles; some of them include: the use of simple and user-understandable language, consistency, clearly marked exits and shortcuts, and understandable error messages (Nielsen & Molich, 1990). However, some experts often forgo these heuristics and depend on their own experience and understanding to evaluate the system (Petrie & Bevan, 2009). In general, HE is usually conducted with a small team of evaluators. However, the larger the expert team, the better because more usability problems are discovered with each set of eyes (Nielsen, 1994). It is suggested that each expert should go through the interface twice so that the expert can get an idea of the system‘s scale and navigational organization the first time and the second run-through is for the actual evaluation (Jaspers, 2009) where experts note which heuristics are violated. Once issues are identified, the heuristic violations are often ranked in order of severity (Seffah & Metzker, 2009) so that developers can focus on the most important ones first if the deadline is fast approaching. Finally, the results of all of the HEs done by the experts are compiled and a report is made to summarize the issues for the project managers.

Many slight variations to the HE procedure exist. Sometimes, particular elements are analyzed while other times, the expert is given a set of tasks to inspect that the user would usually have to do (Petrie & Bevan, 2009). The variation involving tasks has similarities with the Cognitive Walkthrough method (described in the next section) because both are focused on task completion. However, experts using the HE check for compliance to guidelines while attempting to complete that particular task and this is not the case with a Cognitive Walkthrough. Another variation to HE, called Cello, involves the evaluators working together to find usability issues, and then individually rating the importance of them (Petrie & Bevan, 2009). One more analytical evaluation method that is similar to HE is Combined Heuristic Evaluation (Bekker, Baauw, & Barendregt, 2008). Combined HE is a mix of two analytical methods: four heuristics developed by Malone and Lepper (1987) and Nielsen‘s (1994) ten usability heuristics (Bekker et al., 2008). However, since the heuristics used to do the HE are not predetermined by the method, the Combined HE is really just an HE with a specific set of heuristics, and not necessarily a new method altogether.

Even though HE is an efficient usability method that produces many usability issues at a low cost, it has a few limitations. For example, different experts evaluating the same project come up with different problems, meaning it is not 100% repeatable (Jaspers, 2009). Also, expertise influences the outcomes of a heuristic evaluation, as expert usability assessors will find more usability issues than the novice evaluators

(Jaspers, 2009). Another issue with HE could be that experts are expensive to hire. However this is a drawback of all expert-based usability evaluation methods, including the Cognitive Walkthrough, examined next.

Cognitive Walkthrough

In walkthrough evaluations, experts use their experience to evaluate the 'learnability' of the design (Polson et al., 1992) and find problems while working through specific tasks (Petrie & Bevan, 2009). Completing a task usually takes more than a single screen from the interface and puts the evaluator in context, which makes this type of usability evaluation quite effective. The Cognitive Walkthrough only includes experts and they try to reason out and analyse the cognitive processes needed to solve tasks (using the system being tested) by attempting to mimic what a user would do (Jaspers, 2009; Petrie & Bevan, 2009). By doing so, the experts are required to think as the user would think (the cognitive part of the walkthrough). During this process, experts evaluate whether the interface's layout allows the user to complete tasks and if it provides appropriate feedback (Wharton, Rieman, Lewis, & Polson, 1994). The Cognitive Walkthrough analysis used to require detailed documentation, but this is no longer required and it is now simpler to conduct (Spencer, 2000).

In order to conduct a Cognitive Walkthrough, researchers must first define the user context and which tasks the system needs to enable the user to do. Then, each step in the sample task (or tasks) is constructed and given to experts to evaluate. Problems a user would encounter when interpreting labels and the consequences of certain actions are noted (Jaspers, 2009) and repaired.

The Cognitive Walkthrough method effectively reveals severe usability problems (Sears, 1997), especially when the task descriptions are detailed (Sears & Hess, 1999). This is beneficial to researchers when there are time or financial constraints, and it is important to find only the most severe usability issues. However, providing detailed descriptions of how users accomplish tasks so that experts could walk through it, requires the researchers to create the correct action sequence for each task, making this method somewhat time consuming to set up (Jaspers, 2009). Although, Cognitive Walkthrough may be applied to a designer's preliminary idea or in an early design stage because the complete system may not be required for the completion of a given task (e.g. logging in requires only the first couple of screens).

Several variations to the Cognitive Walkthrough exist, including the combined and pluralistic walkthroughs. Both of these variations include users as well as experts. In the pluralistic walkthrough, users work together with experts to discuss issues that arise when they work through tasks using the system (Bias, 1994). Partala and Kangaskorte (2009) came up with the combined walkthrough in which experts evaluate a user's behavioural, affective, and cognitive responses during information retrieval tasks (i.e. searching for specific information, such as a book reference number) in a usability test. Task times and completion rates are used to measure behaviour. Arousal is measured to get a rating of emotion. However, the purpose of measuring arousal and emotion is not

clear as the concepts do not necessarily aid in finding usability issues and requirements. Rather, they are more suitable for user satisfaction. Cognition is analyzed after the usability issues are found and the users are asked to revisit them and Think Aloud (the Think Aloud method is discussed in a later section). These integrated methods in the combined walkthrough allow researchers to get a multifaceted explanation of the interaction between the user and the system, allowing for the more efficient identification of usability issues (Partala & Kangaskorte, 2009). However, these two variations of Cognitive Walkthrough are not pure expert-based methods as users are involved, making them hybrid methods. The main disadvantages to these two variations are that involving both users and experts at the same time is both time consuming and expensive.

Another, less used, method by Farooq and Zirkler (2010) is the Application Programming Interfaces (APIs) Peer Review, which is also adapted from Cognitive Walkthroughs. The goal of an API Peer Review is to uncover usability issues in a specific software feature (Farooq & Zirkler, 2010). During such a test, the designer walks the usability expert through a series of steps that a user would do to perform a specific task, as in a Cognitive Walkthrough. At each step, the expert's job is to comment on encountered problems and the usability engineer takes these errors and converts them into API usability problems. Costs are slightly reduced because although the researchers still need to be paid, money is not spent to recruit users and there is no need for equipment for the API Peer Reviews (Farooq & Zirkler, 2010), other than the interface being tested. Therefore, the API Peer Review is very similar to the Cognitive Walkthrough, with the exception of experts being guided by designers/engineers. This takes less time (i.e. no time is wasted due to exploration). However, this can also hide some problems. For example, designers make it easier to navigate, reducing the chances of finding errors. Also, if a user needs a lot of exploration time to navigate, this would indicate that there are some usability issues that need to be fixed. There are other advantages and disadvantages to expert-based usability evaluation, discussed next.

Advantages and Disadvantages of Expert-Based Usability Evaluation

Besides the advantages and disadvantages given above for each of the expert-based usability evaluation methods, there are some general ones that encompass all expert-based methods. In general, they are a little simpler and thus a little quicker to carry out than user-based evaluation (Petrie & Bevan, 2009). Expert-based usability evaluation methods can also be employed early on in the system development process, to evaluate ideas as well as systems, before they are given to users for testing. In addition, an expert-based usability evaluation may encompass a wider range of users (i.e. an expert can take the role of countless personas) than user-based evaluation using a small number of users (Petrie & Bevan, 2009).

However, there are some disadvantages as well. Mainly, because some expert evaluation methods are employed early on in the development process, not all usability issues can be uncovered, particularly not the ones that appear in later versions (Jeffries

& Desurvire, 1992). Another criticism is that expert-based usability evaluation methods cannot always be applied successfully to complex interfaces (Slavkovic & Cross, 1999) because experts in usability are usually not experts in the given field and they cannot perfectly mimic the user.

Expert-Based Methods Used

The initial goal was to evaluate a website's usability level, to ensure that the initial usability level was high (i.e. before manipulations to create different versions of the website). To enable this, three researchers (experts in HCI) conducted an expert-based usability test of the city council website of Gold Coast. The method applied was the Cello, which involved the evaluators working together using Nielsen's heuristics (Nielsen, 1994) and the HE++ (Chattratichart & Lindgaard, 2008) to evaluate the website's usability and to find potential usability issues. Then, the experts rated the importance of the issues for severity. A long report of the issues found was not necessary as no one was going to fix them – they were just for the purposes of finding out how usable the website is and what needs to be done to manipulate it to make it harder to use for the other versions of the website. To verify the usability level, a user-based usability test was done as well, discussed next.

User-Based Usability Evaluation and Testing Methods

User-based evaluation methods involve direct user participation during the testing and development of a product (Bastien, 2010) and are sometimes referred to as a performance-based evaluation (Bailey, Wolfson, Nall, & Koyani, 2009) and User-Centered Design (UCD; Hussain, Slany, & Holzinger, 2009). When creating an interface and if possible, user testing and evaluation should be used at all stages of development and if this is not doable, then it should at least be implemented at the final stage (Petrie & Bevan, 2009). Given that the development of the website was not under question in this thesis, user testing was done in order to confirm the website's usability level, as found by the expert-evaluation.

Typical user-based tests start with a description of the population (Karat, 1997; Medlock, Wixon, Terrano, Romero, & Fulton, 2002). Usability testing should be done with the target audience because it best gives an accurate account of the level of usability, required for developers and management (Dumas & Redish, 1993; Petrie & Bevan, 2009). Then, an outline of realistic activities that the system needs to allow the users to do is created (Bevan, 2009c). The two most common ways to engage participants with the system are to either let them investigate them on their own, or to involve users in these typical tasks with the product (Karat, 1997; Bastien, 2010). The investigative methods are also referred to as formative, and focus on uncovering usability issues by observing the user's behaviour and noting intentions and expectations (Rubin, 1994; Petrie & Bevan, 2009).

While minimal guidance is given to users while testing the system (unless they are stuck; Petrie & Bevan, 2009), some usability tests are done in pairs of users (Bastien, 2010). Although, individual user testing has been shown to identify more usability errors because pairs can overcame some of the problems by working together (Bastein, 2010). In any of these cases, the user‘s actions and comments (during the task or in retrospective interview) would be observed and recorded. Depending on the project goal, attention would be given to measuring task completion times and rates, and types and frequency of errors (Rubin, 1994; Bastien, 2010). Analysis of these data is done with the purpose of identifying system errors and their causes (Jaspers, 2009) and implementing solutions in order to improve the product.

There are many user-based methods and each researcher (or usability expert) has their own preference as to which specific method they like to use. However, project budget, time, and user availability can restrict practitioners in method selection. The following sections examine specific user-based usability testing methods, and discuss some of their advantages and disadvantages, starting with Think Aloud.

Think Aloud

This method is the most used user-based method (Nielsen, 1993). Originating in cognitive psychology, it was designed to obtain information about reasoning and expectations when performing specific tasks (Jaspers, 2009). In the Think Aloud method, participants verbalize their thoughts as they perform specific tasks with an interface (Van Someren, Barnard, and Sandberg, 1994), which gives researchers access to their expectations of the system, and also shows where the system fails. As in most other user-based methods, the participants in the Think Aloud test should be representative of the population (i.e. a sample of the prototype‘s target audience). Similarly, the tasks should also be representative of the ones the system is intended to do. During testing, if the participant stops talking, the researcher prompts the participant to continue. Although necessary in order to capture the thought process, this method might disturb the cognitive processes (Boren & Ramey, 2000). However, if these prompts are kept to a minimum (in length and in quantity), then the cognitive processes are not interrupted (Jaspers, 2009).

There are three main variations to the Think Aloud method: concurrent think-aloud protocols (the most commonly used), retrospective think-aloud protocols, and constructive interaction (Van den Haak, Jong, & Schellens, 2009). Participants undergoing the concurrent think-aloud protocols must verbalize their thoughts while they work with a particular test object. Although, evidence exists that the concurrent think aloud method interrupts cognitive processes and can slow down use (Boren & Ramey, 2000). Therefore, in retrospective think-aloud, participants first work silently on a test object and then verbalize their thoughts afterwards (sometimes with the help of the video recording of them doing the task). The third method, constructive interaction, involves a pair of participants who work together and verbalize their thoughts as they interact with each other. Participants usually enjoy the constructive interaction method

the most (Van den Haak et al., 2009). Concurrent think-aloud protocols gives more insight into the participant's thoughts and expectations, and generates more usability issues than retrospective think-aloud (Van den Haak, Jong, & Schellens, 2003). Since the purpose of the usability tests in this thesis did not include finding issues that needed to be fixed, and having users work together could result in influencing their experience and thus mental modals, the concurrent think-aloud was not used. Instead, the retrospective think-aloud was used in this thesis, with the option of allowing users to comment on anything they wanted during the test but they were not encouraged to do so.

One limit to retrospective think aloud is that it comes after the testing and participants could forget certain aspects that they would have liked to comment on. This makes retrospective think aloud somewhat unreliable for task-related performance, but useful for reflection on other issues that could have arisen (Jaspers, 2009). In a more recent study, all of the think aloud methods were found to be equally productive in generating data, making them equally effective and interchangeable (Van den Haak et al., 2009). Thus, the retrospective think aloud method was used in this thesis, for all user-based usability tests.

Think aloud may uncover less usability issues than HE and Cognitive Walkthrough, but the issues it does uncover are typically more severe (Jeffries, Miller, Wharton, & Uyeda, 1991). This might occur because HE and Cognitive Walkthrough are expert-based usability evaluations and experts are trained to find usability issues (enabling them to find many of them), whereas users find fewer issues with the system but those issues that they do find are usually show stoppers. However, this is counterintuitive as expert-based usability evaluation should take place before user-based testing for the exact purpose of eliminating severe usability issues. Perhaps users and experts have different standards and the definition of a 'severe' usability issue differs between these two groups. Overall, the Think Aloud methods are widely used because they effectively detect usability issues and their causes. Other user-based methods are being used as well, and the next section describes the FirstClick and Clickstream methods.

FirstClick and Clickstream

One method used to examine software and user webpage performance (but can also be applied to print-outs of interface screens) is called the FirstClick test. It is done by presenting users multiple tasks (only the first screen for each task is presented) and recording whether participants' first clicks are the correct clicks, leading them towards completing the task (Bailey et al., 2009). There are a couple of strengths to this method. One is that many more scenarios can be tested in the same amount of time as a traditional test and the more tasks there are, the more usability issues are found. Another is that by analyzing many different first clicks, researchers get to understand user expectations, and lower fidelity prototypes (e.g. homepage) are needed (Bailey et al., 2009). The major drawback of the FirstClick method is that it does not assess the entire

interface. The first click can really only assess the first step of a given task on a single page. In fact, it is best used to test terminology on an interface before it is implemented. This was not needed in this thesis, as the entire website needed to be tested for a better feel of the usability level. Thus, there is another method, called Clickstream, which addresses this issue.

The application of Clickstream by researchers allows them to determine whether participants could successfully navigate through the prototype (Bailey et al., 2009). In this method, an entire scenario is given to participants and all of their selections (clicks) are recorded and analyzed in order to uncover pitfalls in the interface. This is an effective method to use if at least one section of the interface is designed and can be tested, whereas the FirstClick method is a good way to test separate pages, if there are several designs available for the same interface and different ideas need to be tested. In some ways, the FirstClick method is the first step of the Clickstream, which goes on to test the entire task or interface. This does not differ drastically from a regular task-oriented usability test, but it does keep track of the sequence of clicks the user makes more rigorously. Since the research in this thesis did not benefit from knowing if participants were following the optimal path, this method was also not used.

Rapidly Testing and Evaluating Usability

If time is short, then there are methods that are created to help speed up the usability testing cycle. Rapid Iterative Testing and Evaluation focuses on the fast overturn of usability issue correction and further development of the technology. Specifically, Rapid Iterative Testing and Evaluation focuses on quickly changing features in the interface and rapidly checking the effectiveness of these changes once usability issues are uncovered (Medlock et al., 2002). As is the case with most user-based usability methods, data is collected primarily through the observation of participants attempting to complete tasks. Once real usability issue are revealed (and not just a user hiccup), developers come up with solutions and implement them before it is tested again. In order for this to occur, participant testing must be spaced out to allow for enough time to review results (Medlock et al., 2002). Also, the management and design teams need to work closely together to make sure that the solutions found to the usability problems are feasible. Once the solution is implemented, another participant is emgaged to test the efficiency of the changes and to find more usability issues. Rapid Iterative Testing and Evaluation is effective for finding and dealing with usability problems (Medlock et al., 2002). However, it is more of a process of spacing out participants than it is a usability evaluation method. The actual usability evaluation used by the participants is not mentioned and can thus vary, depending on the researchers. This makes Rapid Iterative Testing and Evaluation a structure to which usability evaluation methods can be applied. However, yet again, this method has to do with iterative development which is not what was needed in this thesis.

Extremely Rapid Usability Testing focuses on rapidly collecting data, with a lesser focus on the implementation of solutions to the usability issues. Extremely Rapid

Usability Testing is another method for participant testing in which the data collection methods (questionnaires, interviews, usability testing, etc.) are applied, in a rather unconventional settings such as exhibitions and shows (Pawson & Greenberg, 2009). In Extremely Rapid Usability Testing, participants are drawn to the researchers because the experiment is on display next to many other booths on an exhibit floor, filled with company representatives trying to advertise themselves. This type of environment quickly attracts participants and they rapidly provide feedback (Pawson & Greenberg, 2009). This method is therefore suitable for the usability testing of a product between versions (e.g. product X version 2.4 is ready for testing and the results of the testing will be implemented in version 2.5 of the product). While this is an interesting approach to user-based testing, advertisement, exhibitions, and rapid version testing were not in the scope of this thesis.

The Remote Asynchronous Usability Testing Method

There are two types of remote testing: synchronous and asynchronous (Bruun, Gull, Hofmeister, & Stage, 2009). In remote synchronous testing, the researcher collects data in real time, from a different location. Laboratory testing is typically synchronous since synchronous testing is cost effective, not always tied to a room or building, and it saves time (Bastien, 2010). Asynchronous testing involves postponing the data compilation until it is convenient for the researcher.

Remote synchronous testing separates users spatially from evaluators while asynchronous separates them both spatially and temporally. This means that participants are not in the same room as researchers during testing in the synchronous testing case, but can be electronically connected and view participant progress during testing online or at another computer. In the asynchronous case, researchers are at a different location and are completely uninvolved with the testing so they can look at the data at a different time. However, rather than being considered a usability testing method, remote synchronous and asynchronous testing could be regarded as way of applying the actual evaluation and testing methods (with each of these methods, most other usability evaluation methods could be used). For example, Clickstream can be done remotely and asynchronously, just as well as it could be done synchronously.

While being in the same room as the participant (i.e. direct observation) gives the researcher a better view of system use, it may also influence the participant‘s behaviour, because people react to others‘ body language and facial expressions (Crystal, 2003). The influence of the experimenter on the participant is known as the Hawthorn effect, where participants change and often improve their behaviour as a reaction to being watched (McCarney et al., 2007; Fox et al., 2008). However, evidence exists to suggest that when participants are engaged by the task at hand, they tend to solely focus on it and pay no attention to their surroundings (Csikszentmihaly, 1990). The downfall is that knowing how engaged they are in their tasks means that you need to interrupt them in order to find out. To avoid interference, direct observation was avoided in this thesis, replaced by remote observation.

Remote testing can be considered a dimension of testing rather than an actual usability testing method. There are three ways for remote asynchronous testing to be done: forum-based online reporting and discussion, user-reported critical incidents, and diary-based longitudinal user reporting (Bruun et al., 2009). However, these could also simply be put together into another dimension of testing: how to report (i.e. the user still needs to use another method to evaluate). That being said, forum-based online reporting and discussion requires the user to take notes on usability issues found, rate their severity, and post it in a forum for discussion with other users. In this method, data is automatically saved for future analysis.

In user-reported critical incidents, users report problems they find with the system online immediately and directly to the researchers. In diary-based longitudinal user reporting, participants do tasks with the system under study for several days. During tasks, participants take notes on usability issues they find and only submit them after the testing period has passed (Bruun et al., 2009). Users find half as many usability issues when applying these three methods, although in less time, than a typical usability test described earlier (Bruun et al., 2009). Thus, remote asynchronous methods are appealing for usability testing if enough participants are taken for the testing to compensate for the small amount of usability issues found. However, acquiring participants can take time and this takes away from the advantage of taking less time. While the number of participants influences the number of usability issues found, the severity of these issues also need to be examined, as fewer but more severe issues can be just as important as more usability issues that are less severe.

Advantages and disadvantages to user-based usability evaluation

Testing a product with real end-users is a good way to uncover pitfalls and get feedback. User-based methods tend to find severe usability issues (Jeffries et al., 1991). A major disadvantage is finding users to participate because this usually takes time and compensating them for their time takes money. Another potential disadvantage is that all users generally receive the same questions for feedback comparison purposes and this limits the test scope. Finally, researchers must assume that participants are being truthful in their feedback throughout the test and after (Gould, 2009), and that their memory serves them well. Therefore, while user-based usability evaluation is important for product quality, they take time and money to implement.

Implications for Thesis

In this thesis, the user-based usability tests were conducted using remote synchronous testing. This was done in a usability lab with participants in an observation room, connected via phone and the researcher in a control room that had access to the participant‘s screen at all times. This was done to avoid influencing the participants‘ responses by being in the same room as them. Yet, a researcher was there to help if the participants got stuck or had any questions. The retrospective think-aloud was used,

with the option of allowing users to comment on anything they wanted during the test. In other words, participants were allowed to work through the usability test and were asked to verbalize their thoughts afterwards.

Usability Measures

Two types of usability needed to be examined in this thesis: objective and subjective. Objective usability tends to be measured during use, while subjective usability is measured before or after use. In other words, subjective usability relies on asking the participant for their feedback on how they perceived the usability (i.e. satisfaction). This could be done by asking how they perceived or experienced the usability, or it could be done by asking them to fill in a form or scale. Objective usability, commonly referred to as ‗performance', is not perceived and user feedback is seldom required. Instead, actions the participant does with the interface are recorded. Time taken to complete a task (i.e. efficiency) and success of task completion (i.e. effectiveness) are classic examples of objective usability.

In order to acquire usability measures, a participant must interact, to a certain extent, with the interface. There are typically three types of tasks: viewing, browsing, and goal-oriented. Viewing an interface usually occurs in a matter of seconds and simply exposes the participant to screenshots of the interface (usually the homepage), without any interaction. A browsing task is one that involves the user interacting with the website on their own time and pace, without a specified purpose. In other words, the user does not pursue information set by an experimenter – there are no defined tasks other than to view the website and to acquaint themselves with the interface. A goal-oriented task involves the participant interacting with the website in order to complete a specific task, such as purchasing an item or getting specific information. Goal-oriented tasks have also been referred to as ‗information retrieval' tasks (Zhang, 1996). The work in this thesis includes both viewing and information retrieval tasks. Viewing the website was necessary in order to acquire the pre-use perceived usability and visual appeal ratings. Goal-oriented system use was required in order to obtain objective usability measures, and for the participants to become more experienced with the interface so that they could rate the website on perceived usability and visual appeal post-use as well. The following two sections go into more detail about objective and subjective usability measures, respectively. This is followed by visual appeal measures.

Objective Usability Measures

As previously mentioned, the objective usability measures in this thesis came in the form of performance measures, obtained per task. In general, effectiveness and efficiency, being part of the usability definition used in this thesis, were measured. Effectiveness was measured by examining success rates and answer correctness. Success rates were measured by averaging the total number of passed tasks (a binary value, on a pass/fail basis) per participant (i.e. the proportion of correct responses).

Mathematically, taking the sum of binary numbers amounts to the same as calculating the mean, especially given that statistics were done on the group mean (i.e. mean of means). The correctness of the task answer was evaluated as follows. If the participant found and wrote the appropriate answer, then the task was correctly answered. Otherwise, it was incorrect.

Efficiency was measured by counting the number of clicks per task, the number of times a menu was hovered over thereby revealing the menus content, task completion time, and in the preliminary studies, the number of hints offered by the researcher per task. The drawback to these measures is that it is time consuming to do them manually for each task. Several programs exist that can keep track of these, but false positives do occur. For example, if a participant is clicking a text to highlight a section of it, those clicks would be included in the total click count, yet we are not interested in those as they are not navigation clicks. In addition, hovering over a menu item to expand it is not a common feature that is registered by usage trackers. Also, the tasks were presented in random order and if the user accidentally skipped a task and had to go back and do it – these actions would not be properly coded. Participant audio and video were not recorded, so catching errors post-test using the video of the screen interactions would be hard, leading to a greater possibility to improper data classification. Thus, all objective data was counted manually. Specifically, the number of clicks and hovers were counted per task, and time to complete the task was recorded. Time was measured in seconds, per task, from when the participant read the question and began to look at the website for the answer, to the time when the participant was on the final webpage from which they concluded the task answer. For the preliminary studies, hints started off as vague and became increasingly specific over time if the participant continued to be lost. Participants were allowed three hints per task, at which time the task was automatically considered a fail. For more details on hints, please see Chapter 5. All of these measures were used to gain an understanding of Gold Coast‘s city council‘s website‘s objective usability level.

Eye-Movement Analysis

Recently, eye tracking technology has become popular when evaluating usability. Since self-reports do not always accurately correlate with performance measures, the eye tracker is an objective measure where participants‘ gazes are recorded by cameras on the desk or laptop computer. This method is often used to compare the eye-tracker‘s recordings to where participants say they look **(**Albert & Tedesco, 2010). In addition to this function, eye trackers give information on how attractive or visible a feature is in a given design by measuring how long participants look at it. This is especially the case where participants cannot vocalize their thoughts, such as with childhood development studies (e.g. Olah, Elekes, Brody, & Kiraly, 2014; Davidse, de Jong, Shaul, & Bus, 2014; Leppänen et al., 2014). Moreover, whether or not a feature was actually seen as reported by participants can also be tested using an eye tracker **(**Albert & Tedesco, 2010). If a participant is stuck in a particular area of the interface, the amount of time

and number of times they look at certain areas of the screen can tell developers where users are finding errors (Coltekin, Heil, Garlandini, & Fabrikant, 2009) or where they are anticipating the answer to be. However, if a user is lost and does not know where to click, then the eye tracker is not needed to confirm that they are lost.

Eye movement recordings in conjunction with traditional usability tests are referred to as the Empirical-Evaluation-Based Methodology (Coltekin et al., 2009). When applying this method, eye and mouse movements, response times, accuracy, and self-reports are recorded while participants does tasks. Gaze plots and fixation patterns are marked by the eye tracker, which connects traditional usability measures with users' interface interaction processes. In addition, the empirical-evaluation-based methodology brings both quantitative and qualitative information to the analysis, making it very versatile and useful (Coltekin et al., 2009).

However, given that the purpose of this thesis was not to find usability issues, or to find particular areas of interest, the eye-tracker was not necessary. Subjective usability measures were examined, however, in order to determine the influence of expectation on participants' perception of usability. Therefore, the next section discusses subjective usability measures.

Subjective Usability Measures

As is the case with most subjective measures, or matters of opinion, the general approach is through verbal protocol, and scales that have both Likert and semantic rating items. In fact, scales are one of the most precise, non-invasive, measurements of subjective perceptions, such as usability and visual appeal (Hirschfelda & Thielsch, 2014). Only scales that were validated, accepted, and used in the field were chosen in this thesis. Moreover, one-item scales were not considered because while most usability information is retained in the item, some information is lost and reliability is compromised (Christophersen & Konradt, 2011; Sauro, 2013). Thus, to supplement the quantitative information provided by the objective usability, questionnaires were also used in this thesis for subjective ratings of both usability and visual appeal.

One such usability metric is the Questionnaire for User Interaction Satisfaction (QUIS; Chin et al., 1988; Harper & Norman, 1993). As the name suggests, QUIS is a questionnaire that asks participants about their satisfaction with the interface at hand. The original QUIS (there are several versions, including online versions) was very long, with 80 nine-point Likert scale questions that participants needed to fill in. Question topics varied from demographic, to overall system satisfaction, to specific evaluations of interface characteristics such as terminology. For example, one question asked participants the degree to which they found the messages on-screen clear or confusing (Harper & Norman, 1993). This also indicates that this questionnaire would not be easily used for pre-use measurements, where terminology and function would have to be guessed. In addition, some of these questions are too detailed for the purposes of this thesis, in which we just need to see what their opinions are of the overall usability. We

could have chosen a sub-set of questions that did relate more so to our study. However, the particular subset may not have been chosen to accurately represent the main usability topics (i.e. construct validity, generalizability).

Another, slightly shorter, questionnaire is the Software Usability Measurement Inventory (SUMI), which is a 50-item questionnaire that measures users' perception of an interface. However, the SUMI is also a post-study questionnaire, and its licence costs several hundreds of dollars each month. The data collection in this thesis took place over two years, rendering this questionnaire unusable.

A shorter usability assessment questionnaire is the Post-Study System Usability Questionnaire (PSSUQ). The PSSUQ has 16, 7-point Likert scaled questions, measuring user satisfaction. However, since the questions are all phrased in a positive manner (i.e. positively biased; e.g. "It was simple to use this system"), then usability tends to be rated slightly higher with this questionnaire (Garcia, 2013). Thus, it was not further considered for use in this thesis.

One of the most widely used metrics is the System Usability Scale (SUS; Brooke, 1986; 1996). It is a short, 10-item questionnaire and it is known for being a "quick-and-dirty' tool for assessing website usability (McLellan, Muddimer, Peres, 2012). The ten questions are statements and the participant needs to indicate the degree to which they agree with the statement via five-point Likert scales (i.e. agree, disagree). Each participant fills in all ten questions. According to Brooke's scoring method, half are negative statements (e.g. "I thought that the website was cumbersome to use'), meaning that the ratings need to be inversed for the negative statements. For example, a rating of "4' on a negatively phrased question becomes a "2' for the purposes of statistical calculations. The SUS questionnaire has been used in studies for both interface designs and implementations, and for subjective appraisals of usability without the purpose of altering the website design (McLellan et al., 2012).

The SUS scale contains more than one-item, has been validated, used repeatedly by researchers in the field, and is quick to administer. By having negative questions, it also makes certain that participants are actually reading the items, as consistency in their responses would signal their lack of attention to the questions. The SUS scale was thus used in this thesis as a means to acquire subjective usability ratings, both upon initial viewing of the website (pre-use) and after having interacted with the city council website for about an hour (post-use).

Visual Appeal Evaluation Approaches

Unlike with usability, one does not need to use an interface in order to accurately assess its visual appeal. First impressions are accurate measures of it because visual appeal can be assessed in a fraction of a second (Lindgaard, Fernandes, Dudek, & Brown, 2006). The approaches to measuring visual appeal do not vary widely; they all involve viewing the interface, and different research will vary with respect to exposure times. Instead, there are many different assessment tools and finding a standardized measure of visual appeal is not easy because there are no guidelines on which measures

to choose (Augustin, Wagemans, & Carbon, 2012; Faerber, Leder, Gerger, & Carbon, 2010). Measures differ in the literature because the definition of visual appeal has not been standardized (Augustin et al., 2012; Markovic, 2012), as has the definition of usability. As mentioned earlier, some definitions are focused on the cognitive processes, while others examine affective responses (Leder, Belke, Oeberst, & Augustin, 2004).Visual appeal can be measured by observing body language and facial expressions, by asking participants to verbalize thoughts, by using standardized scales (Hirschfelda & Thielsch, 2014), and by monitoring physiological data.

Observing body language can be subjective if not done by a professional, and it assumes that the participants will have a big enough reaction to the website that they will outwardly show it. Facial expressions (e.g. for micro expressions; Freitas-Magalhães, 2012) are not easily suppressed and are accurate in determining true and false emotions. A true emotion, for example, is when someone is genuinely happy. An example of a false emotion is when an individual fakes the happiness with a forced smile. Micro expressions are also good predictors of when an individual is lying about something they said. However, none of which are within the scope of this thesis.

Expressions such as ~~"~~it's breathtakingly beautiful" suggest that there is a physiological reaction to a visually appealing stimulus. Physiological measurements such as heart rate and body temperature, psychophysiology data such as galvanic skin responses (GSR), and neurophysiologic measures such as thermoregulatory sweat testing (TST) and sympathetic skin response (SSR; Tarchanoff, 1890), can be used to measure psychological changes resulting from exposure to stimulus. The most common psychophysiological measure is the polygraph, most commonly used as a lie-detector. The electro-encephalogram (EEG) measures brain wave patterns but requires hardware to be attached to the head and special analytical software. In sum, each one of these physiological measures is slightly invasive as participants often have to be hooked up to equipment. In addition, equipment often requires software and licences, which can take time to get and learn to use. Also, these devices measure physiological responses, which do not align with the definition of visual appeal used in this thesis. In this thesis, visual appeal was defined as a cognitive judgment of an object's aesthetic appearance (Blijlevens, 2011). Thus, the remaining visual appeal measures that are discussed are in the form of questionnaires and scales.

One of the most widely recognized set of visual appeal scales are the two by Lavie & Tractinsky (2004), for classical and expressive aesthetics. In HCI, classical aesthetics refers to the aspect of screen space management such as contrast, repetition, alignment, and proximity (also known as the usability CRAP principles). Expressive aesthetics refers to the user's judgements of creativity and originality of the design. A recent study by Sonderegger, Sauer, and Eichenberger (2014) found that expressive aesthetics was significantly different concept from classical aesthetics. The difference between the different aesthetics measures can be attributed to the similarity of classical aesthetics with usability principles. Since classical aesthetics contains elements that are considered usability characteristics, defining usability and visual appeal as independent constructs becomes difficult. Furthermore, the lack of independent definitions makes

independently manipulating usability and visual appeal of the website data sample impossible. Thus, Lavie & Tractinsky's scales (2004) were not used in this thesis.

Another visual appeal assessment tool is the Visual Aesthetics of Websites Inventory (Moshagen and Thielsch 2010). However, this tool was too long (18 items) to sustain their concentration and interest in the study, and it needed to be given to participants twice in one session (once at the beginning and once at the end). Therefore, the Visual Aesthetics of Websites Inventory – Short version (VisAWI-S; Moshagen & Thielsch, 2012) was used instead. The VisAWI-S is a four-item visual appeal scale, where each item is a seven-point Likert scale. The four items are simplicity, diversity, colourfulness, and craftsmanship. Simplicity refers to whether or not the website looks like the website was cohesively constructed. Diversity asks about the website's layout and if the user finds it interesting. Colourfulness refers to the colour composition. Craftsmanship asks if the website was professionally designed. The full scale can be seen in Appendix B. These four questions are quick to fill in, and the questionnaire is a good substitute for its much longer version (Moshagen & Thielsch, 2012). This scale does have a couple of limitations. As Moshagen and Thielsch (2012) stated in their own paper, the scale was developed in the German language which may make it hard to maintain the original meaning when translated into different languages (as often, nuances are lost in transition). In addition, the validation of the scale was done with Germans. These may be limiting factors, since other cultural and linguistic differences were not explored. However, the authors did provide the translation in English which was easily understood. This scale was for visual appeal assessment by participants and it was used as a guideline for manipulating the website to create uglier versions (i.e. low visual appeal). More information about the manipulations of the website is in Chapter 5.

Study Participants

A random sample of student volunteers was used in this thesis. Random participants were used because finding participants that have certain expectations would be difficult, as they would have to be aware that they may be biased in some way. In addition, some websites (e.g. government) have a poor reputation and it would be harder to get people who believe otherwise. It also lends itself to the possibility of having other confounding variables. Instead, random students were used because they do not threaten external validity (e.g. Svahnberg, Aurun, & Wohlin, 2008; Druckman & Kam, 2009), and expectations were controlled experimentally.

Visual appeal, perceived and objective usability needed to be measured and analysed. Each participant needed to be able to view all cues that were on the website in order to be able to give accurate feedback on the visual appeal and usability. Therefore, normal, 20/20, vision was a requirement for participant selection. Moreover, in order to get a diverse sample of participants to better represent the population, this study was advertised to participants varying in their demographics including, age, gender, and English language proficiency. A large sample size was required to enable statistical analysis.

A grand total of 223 Swinburne University of Technology staff and student volunteers participated in the user- and expert-based tests in this thesis. All of these participants had 20/20 or corrected to 20/20 vision, and were screened for colour blindness. In addition, all participants were technology-savvy, regular Internet users. The university students used were of mixed backgrounds, varying in gender, age, cultural backgrounds, and in English proficiency levels. The visual appeal and usability relationship was not affected by age (Sonderegger, Sauer, & Eichenberger, 2013), so participants 18 years or more were used. Students were chosen as the test subjects because access to them was more readily granted by the university‘s ethics board. In addition, there are thousands of students at Swinburne University, so acquiring a random subsample of them for use in this thesis was deemed doable at the start of the thesis. Also, the laboratory used was on campus which made it more readily available to students already there. Additional information on these participants can be found in Chapters 5, 6, and 7.

While there were over 200 participants, these were distributed over eight studies. In the main studies, there were 10 participants were condition (total of 140). According to Nielsen and Landauer (1993), five participants will find approximately 75% of all interface errors, with 15 participants finding 100%. The decision to engage ten participants per condition was based on the premise that ten was the required number to find nearly all usability issues (Nielsen & Landauer, 1993). While the participants in this thesis did not need to find any usability issues, they did need to be able to assess the usability with accuracy. Therefore, each condition in this thesis had ten participants.

Statistical Analysis

Qualitative analysis was done to uncover patterns in the data and gain a deeper understanding of participants‘ reasoning and reactions to the expectations. Quantitative statistical analysis was used in addition to the qualitative analysis to ensure that the patterns being uncovered were not happening due to chance or error. In a sense, statistics were done to validate the importance of the findings in this thesis.

There are many statistical tests that can be applied to data to test for differences between measures of centrality. However, not all of them can be applied to all data. In fact, the majority are very specific tests and are only applied in certain cases. In order to determine which statistical tests to apply, certain criteria need to be examined. Therefore, the first step in statistical analysis is usually testing assumptions, which occurs to determine if the data is normally distributed and to see if the variance is homogenous.

The normality assumption was tested using Shapiro-Wilk (Shapiro & Wilk, 1965; Razali & Wah, 2011) and skewness and kurtosis measures (Cramer, 1988; Cramer & Howitt, 2004; Doane & Seward, 2011). If the normality assumption was not violated, then parametric tests are used (e.g. the parametric Levene‘s test, Martin & Bridgmeon, 2012) to examine the homogeneity of variance assumption. However, in this thesis, the normality assumption was violated in every study and thus the non-parametric Levene‘s

test was used (Löfgren, 2000; Nordstokke & Zumbo, 2010; Nordstokke et al., 2011). Given that assumptions for normality and constant variance were not unilaterally met, that some variables were binary (passes), some were discrete (clicks and hovers) and others continuous (time), and that sample size per condition was relatively small (n=10), ANOVAs could not be applied to the data. Therefore, non-parametric tests were applied, chiefly Kruskal-Wallis for main effects, Fisher‘s Exact Test and Wilcoxon Mann-Whitney for pairwise comparisons. Spearman‘s Correlation Coefficients were also used to examine other relationships that may exist between variables.

Outliers were looked for (Hoaglin, Iglewiez, & Tukey, 1986) but were not found, since the main ratings occurred on a short, restricted scale. The IBM SPSS statistical software was used to calculate all necessary values. Using the results from the SUS and VisAWI-s, beanplots were created to gain a general understanding of the data. Beanplots are a more advanced form of bar graphs, where the distributions are shown on both sides of the middle bar (Kampstra, 2008). They give the population spread which allows for more accurate concussions to be drawn. They can also visually present more complicated results.

Stimulus Type

This thesis required two controlled stimuli. The first stimulus was a website data sample that varied in visual appeal and usability. The second stimuli were the textual and verbal expectations. Both of these stimuli are explained here.

Website. A live website (i.e. fully functional website that was already online and accessible to everyone) was chosen as the base for the data sample in this thesis. This was done for two main reasons. The first reason was that constructing a website would be time and monetarily consuming and unnecessary given that the goal of the research was not to create a perfect website. The second reason was that the interaction from a live graphical user interfaces influences users‘ experiences of the interface (Miniukovich & De Angeli, 2015). Static screenshots that are commonly used do not reflect the transition and lag time (or lack thereof) of a live interface. These factors influence perceptions of system quality (Miniukovich & De Angeli, 2015) and may interfere with ratings of visual appeal and usability. To maximize ecological validity, seven out of eight studies in this thesis used live interfaces. More information on these studies is presented next.

The website was chosen from an unfamiliar domain so as to control previous experiences participants would have. A series of preliminary studies was done to ensure that the website was indeed from a less familiar domain (i.e. Australian city council websites). Moreover, the website was manipulated to vary in usability and visual appeal, and these manipulations were also verified in the preliminary studies.

Expectations. The participants were randomly chosen from Swinburne University and the website genre was unfamiliar to them, therefor participants did not have many

(if any) prior experiences or expectations of city council websites. Thus, as mentioned in the Introduction, both textual and verbal expectations were implemented in this thesis. Generally, having polarized descriptions of upcoming tasks can be considered biasing participants. Yet, this occurs in life: social media and user reviews tell us what products are good/bad (e.g. Smith, Donnavieve, Menon, Satya, & Sivakumar, 2005). The expectations were explicitly stated and polarized (overly good or bad) in order to remove the possibility of them not being understood. Extraordinary expectations create greater changes in an individual because they can cause positive and negative dissonance (Bikhchandandi et alk., 1992). Thus, we examined if nuanced task descriptions and verbal feedback could impart expectation and impacts users.

Outline of Studies Conducted

To examine the research questions and test the hypotheses, a series of controlled laboratory experiments was conducted as outlined in this section. Firstly, a series of preliminary studies was done to obtain the website that would be used in future testing, and obtain four versions of the website, varying in usability and visual appeal, to be used for the different conditions. The four conditions were high in usability and high in visual appeal (HuHv), high in usability and low in visual appeal (HuLv), Low in usability and high in visual appeal (LuHv), and low in usability and low in visual appeal (LuLv). The original website was evaluated via expert- and user-based usability tests, and the different versions of it (high/low in usability/visual appeal) were tested via user-based tests. Once an appropriate data set was ascertained, the main studies were conducted in order to gain an understanding of the effect of expectations on usability and visual appeal. The first main study in this thesis was done with the HuHv and LuLv websites, with high, low, and no (i.e. the control) written and verbal expectations. The second main study then repeated the testing but with the remaining two websites (HuLv and LuHv) with congruent, incongruent, and no expectations. Each study is explained and elaborated on below, and even more so in its proper chapter.

Preliminary Studies

The purpose of the preliminary studies was twofold. First, it was to obtain a website genre that was unfamiliar to participants to enable the control of expectations. Second, it was to find a website dataset that varied both in visual appeal and usability. Therefore, five preliminary studies were done in advance of the main studies, outlined next.

Preliminary Study 1. The first study was done to examine two different genres (tourism and city councils) in order to determine which genre was less familiar to participants. A controlled laboratory experiment with 30 participants was done in which the collected data included subjective visual appeal, usability, and expectations

questionnaires. Data was statistically analyzed. The city council websites had more random results, indicating that they were less familiar. This was established by comparing results of an expectation questionnaire (checklist of items that would appear on a city council or tourism website, and semantic differential scales for opinion of the genres) to actual website facts and to their ratings of perceived usability and visual appeal of 52 websites. In addition to having inaccurate expectations of what items appear on city council websites, their expectations indicated that city council websites would be cumbersome to use and uglier than the average website. Yet, their actual website ratings (upon viewing and rating several in each genre) showed that the prettiest website was in fact a city council website. Therefore, city council websites were chosen because they were deemed less familiar (so that expectations could be manipulated with greater ease) and the Gold Coast city council website was chosen because it was rated as the prettiest website overall. It was estimated by the researchers that it would be easier to manipulate the prettiest website to be uglier than it would be to do so vice versa.

Preliminary Study 2. Three researchers performed an expert-based usability test to establish a usability level for the chosen website. This was done using a variation of the expert-based heuristic evaluation method described earlier, in a lab setting. The researchers found that the website was easy to use.

Preliminary Study 3. This study was done to confirm that the Gold Coast‘s city council website‘s usability level was considered to be high by users. This was done via controlled laboratory experiment with ten participants in a user-based usability test. Data collected included subjective visual appeal and perceived usability questionnaires, retrospective talk aloud, and objective usability measures. The results confirmed the expert-based usability results; it was usable. Thus, the original Gold Coast city council website was deemed to be the high usability and high visual appeal (HuHv) website, used in future studies.

Preliminary Study 4. Phase four of the preliminary studies included the manipulation of the original Gold Coast city council website to create three other versions (hard and pretty, easy and ugly, and hard and ugly). Once the new website versions were ready they were each tested with users to ensure that the manipulations were sufficient and varied significantly from the original website. This was also done via controlled laboratory experiment with ten participants in each condition (total of 30). The collected data included visual appeal and perceived usability questionnaires, retrospective talk aloud, and objective usability data. Statistical analysis was applied to the applicable data. The website manipulations showed that visual appeal was significantly worse but the results were not as clear for usability. To address these results, another preliminary study was added.

Preliminary Study 5. The purpose of the fifth preliminary study was to re-manipulate the usability of the website and re-test it using a different set of ten participants. The method was identical to Preliminary Study 4. The results showed that the Lu website was now significantly harder to use than the Hu website. The dataset was ready for the first main study, outlined next.

Main Study 1

The purpose of the first main study was to see if written expectations influenced the visual appeal, perceived and objective usability of the Gold Coast city council website. To examine this influence, a controlled laboratory experiment was done in which the collected data included subjective pre- and post-use visual appeal and perceived usability questionnaires, retrospective talk aloud, and objective usability data. The first study included two levels of visual appeal (high, H, and low, L) two levels of usability (H and L), and three levels of written expectation (H, L, and none, N). Data was statistically analyzed. Some significant results were obtained but results were overall mixed as participants were not consistently influenced by expectations.

Main Study 2

Given that the first main study did not have optimal results, the purpose of this study was to re-examine the impact of expectations on usability and visual appeal, but with verbally reinforced expectations. This was done using the exact same approach as in Main Study 1, but with the addition of a confederate. In psychology, a confederate is a member of the research team who acts as if they were a participant, and interacts with the real participants in order to influence their opinions (Asch, 1956). Therefore, this study included a confederate to verbally reinforce the written descriptions used in the previous study to see if the combined implementation of expectations would have a greater impact on visual appeal and usability. The experiment only included the HuHv website (i.e. the original Gold Coast city council website) with the HuHvHe and HuHvLe conditions. Data was statistically analyzed.

Main Study 3

Thus far, the visual appeal and usability levels studied were congruent with each other. In other words, they were both either high or low. The purpose this study was to examine the influence of expectations on visual appeal and usability when those two factors are incongruent, in the Gold Coast city council website. The exact same method was used as in Main Study 2 (with a confederate). To study this, the easy but ugly, and the hard but pretty versions of the website were used, and each website was subjected to three expectation conditions: high usability and low visual appeal (HuLv), low usability and high visual appeal (LuHv), and no expectations (Ne) which was the control condition. This way, the expectations for usability and visual appeal were either both

congruent or both were incongruent with the actual website levels. Data was statistically analyzed. More details can be found in Chapter 7.

Summary

In this thesis, five preliminary studies and three main, experimental studies were done, in a controlled, laboratory setting. Participants for each of the studies were volunteers from Swinburne University of Technology. They were randomly assigned to conditions. Qualitative and quantitative data were collected through observations, interviews, and questionnaires. These included the SUS questionnaire for perceived usability, the VisAWI-S for visual appeal, and several performance measures for objective usability. In the preliminary studies, usability was further assessed using expert- and user-based testing.

Qualitative analysis was done to uncover patterns in the data and gain a deeper understanding of participants‘ reasoning and reactions to the expectations. Once the data was compiled, all statistical assumptions were checked and non-parametric tests were applied in addition to qualitative analysis to ensure that the patterns being uncovered were not happening due to chance or error.

To avoid the influence of previous experience and confounding expectations, an unfamiliar website genre was chosen for the experiments. Then a single city council website was manipulated to four versions, varying in usability and visual appeal levels. Expectations were also controlled for – manipulated and implemented both textually and verbally, in order to examine their impact on participants when interacting with and evaluating usability and visual appeal.

The preliminary studies were done to obtain the website that would be used in future testing, and obtain four versions of the website, varying in usability and visual appeal, to be used for the different conditions. The main studies examined all four website versions, with textual and verbal expectations that were congruent or incongruent to the website‘s actual usability and visual appeal levels. These results were compared to the control conditions to reveal that, indeed, expectations impacted both the perception of and use of website usability and visual appeal.

The following chapters discuss each of these studies in more detail. Each chapter has its own introduction, method, results, and discussion sections. Chapter 5 discusses the five preliminary studies. Then, Chapter 6 discusses the first main study which is divided into two sub-studies. Specifically, the chapter starts with Main Study 1 which pertains to the HuHe (easy and pretty) and LuLe (hard and ugly) websites and expectations, implemented in written form. The second part of Chapter six (i.e. Main Study 2) uses only the HuHe website with a confederate to verbally implement the expectations in addition to the text. Then, Chapter 7 discusses Main Study 3 in which the mixed conditions were studied (i.e. HuLe and LuHv). After these chapters are the discussion and conclusion chapters.

Chapter 5: Website Acquisition and Transformation

This chapter describes the process for acquiring the stimuli required for the formal studies that investigated the impact of expectation on visual appeal and perceived usability. The goal was two-part. First, it was to acquire a website genre that undeveloped mental models to participants. This was done in an attempt to control for previous experiences and expectations with the websites. Then, the goal was to develop four stimulus websites: one that was high in usability and visual appeal (HuHv), one that was low in both of these (LuLv), and the two mixed versions of high usability and low visual appeal (HuLv), and low usability and high visual appeal (LuHv).

The process involved 5 phases. Phase 1 comprised a preliminary study to select a suitable website domain with undeveloped mental models, for the subsequent formal experiments. Participants rated a sample of city council and tourism websites on visual appeal. These ratings identified the appropriate website genre and website. Phase 2 involved an expert-based usability test for the selected website. This was followed by a user-based usability test for the selected website in Phase 3. The outcome of Phases 1, 2, and 3 established levels of visual appeal and usability for the website. These levels provided the basis for manipulating the usability and visual appeal characteristics of the website to create the four stimulus versions mentioned above. Phase 4 was designed to create and validate the four versions to ensure they provided the appropriate levels of usability and visual appeal. Since some of the manipulations in Phase 4 were not significantly different, the website was re-manipulated and re-tested in Phase 5. This process resulted in four validated stimulus websites exhibiting the required characteristics and levels of usability and visual appeal.

Phase 1: Preliminary Study Introduction

The purpose of Phase 1 was to select a suitable stimulus website genre and website for the subsequent formal studies. Given that the thesis‘ purpose is to acquire an understanding of the influence of expectation on the relationship between usability and visual appeal, we needed to control for expectations as much as possible. In order to achieve this, we needed a website genre that had less developed mental models, to exclude the influence of past experiences. Therefore, the purpose of Phase 1 was twofold: (1) to find a domain with a less developed mental model and (2) to find a website in that domain that did not meet the domain‘s expectations.

To acquire such a domain and website, participants rated a sample of city council and tourism websites for visual appeal and perceived usability. These two domains of city websites were selected because they provide a neutral context, without gender-preferences, ethnic, or age discrimination that can bias expectations. In addition, city websites tend to display very similar information. These websites provided a spectrum ranging from serious styles (city council) to pleasurable styles (tourism). Other website genres were examined as well, such as shopping, entertainment, and other government

websites such as the tax office website. However each of these other website genres came with potential confounding variables. For example, what kind of shopping would it be? If we chose a clothing website, would males want to even do the study and if they did, would they struggle more than if it was a more tech-oriented site? To avoid the possible influence of gender and previous experiences, shopping websites were not used. Entertainment websites may also come with biases. For instance, if it was a movie theatre website, introducing movies with different genres and age restrictions might influence participants in unaccounted ways. Additionally, _entertainment' on its own is meant to evoke emotional responses, which were out of scope for this thesis but would need to be controlled. Using the tax office website may have had a confounding expectation, given the likelihood of a previous negative experience. Therefore, a website genre needed to be chosen which did not have obvious limitations and that participants (i.e. university students) would not be experienced of extremely familiar with. Hence, city council and tourism websites were chosen. The most populated and largest Australian cities in 2012 were examined for use in this thesis. City websites were chosen only if the city had both a city council and a tourism website.

Participants also filled out a questionnaire on expectations to see which website genre they were less familiar with. These ratings enabled the researcher to identify the appropriate website genre and website to use in the rest of the thesis.

Method

Participants

A sample of 30 (23 males, 7 females; 28 aged 18-30 years, two 31+) Swinburne University student volunteers participated, all with 20/20 or corrected to 20/20 vision, and screened for colour blindness. According to the demographics form participants filled out, all participants were technology-savvy regular Internet users. Some 16 were of Australian descent, and the rest were international students of varying English fluency levels.

Twenty-seven (of 30) participants claimed to use the Internet regularly for studying, 25 for social purposes, for entertainment, 20 for tracking news, 17 for banking, 14 for shopping, and four used the internet for traveling purposes. Nearly half (14) of the participants stated that they were not very familiar, 15 stated that they were somewhat familiar, and one said that s/he was very familiar with the purposes of city councils. All were tested in individual sessions taking approximately one hour, and all were given a $20 gift voucher at the end of the session.

Apparatus and Location

Participants were tested using a Hewlett Packard desktop computer, running Intel® Core™2 Duo CPU with 3GB of RAM, and a screen resolution of 1290 X 720. A program running on Firefox Mozilla was developed to present the websites and collect

users‘ scale ratings and responses to the expectations questionnaires. Microsoft Excel and SAS were used to analyze the data. The study took place in a quiet computer lab at Swinburne University of Technology.

Materials

All documentation pertaining to the preliminary study is in Appendices B and C. An informed consent and project information form were prepared and approved by Swinburne‘s SUHREC ethics committee. A demographic questionnaire was administered to determine the participants‘ background information (e.g. age, gender, and education). As mentioned in the Method Chapter (Ch. 4), the System Usability Scale (SUS) and the Visual Aesthetics of Websites Inventory – Short version (VisAWI-S) scales were used for each website. Each participant filled in all questions on these two scales, for each website.

The most populated and largest Australian cities in 2012 were examined for use in this thesis. Cities were chosen only if they had both a city council and a tourism website. Thus, a sample of 26 live Australian city council and 26 live tourism websites was used. Three screenshots, making up a set, were taken of each website: (1) home page, (2) a main menu page, and (3) one from deeper in the menu and hierarchy. For city council websites, this included the homepage, the About page, and an Accessibility or Safety webpage. For tourism websites, this included the homepage, the Accommodation page, and the natural attractions or parks webpage.

A computer program was created to display the stimuli and collect participant responses. The program initially presented a set of instructions from which pressing the ―Start” button displayed a fixation screen containing a ‗+‘ in the centre, shown for one second (see Figure 5.1), in an attempt to avoid any carryover effect between website sets. Then, the program displayed a set of webpage screenshots, such as shown in Figure 5.1, and each screen was displayed for two seconds. At the end of each set of three webpages, the participants completed the VISAWI-S and SUS scales on the same screen. Pressing ―Next” activated the next set of webpages, commencing with the ‗+‘ fixation screen. Participant‘s results were stored in a comma-delimited file.

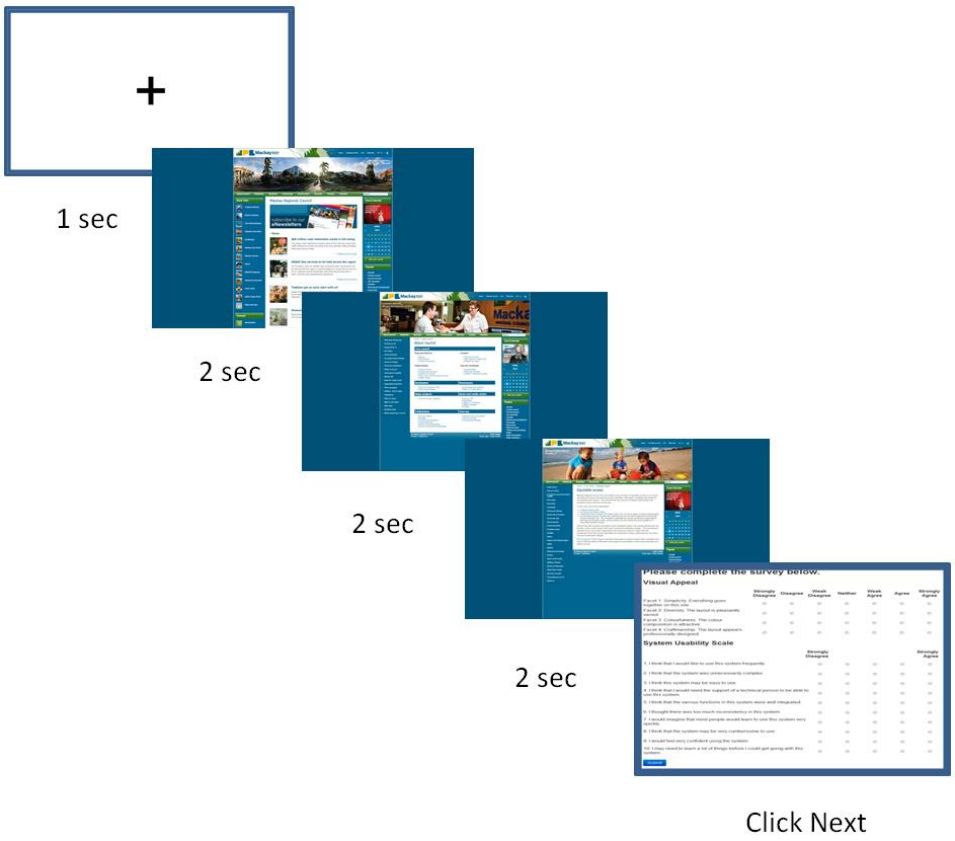


Figure 5.1. Example of the website set of screenshots and rating scales.

A one-page, two-part Expectations Questionnaire was given to participants to elicit their a priori expectations regarding city council and tourism websites. The two genres were examined for features that were both common and specific to each genre. Thus, the first half of the expectations questionnaire was an item checklist containing 23 items, such as a shopping cart or a photo gallery, asking participants to tick all items they would expect to see in the given website genre. This part was included to identify if participants were more familiar with one genre and could therefore identify its components more readily. The second half contained nine graded semantic differential scales addressing participants‘ expectations of visual appeal, usability, and attitude towards websites in each of the two web genres. For example, an item from the scale asked participants if the genre was enjoyable. The nine semantic differential scales ranged from one to five, with a score of one meaning that participants agreed most with the word closest to the one and a score of five meaning that participants agreed most with the word closest to it. The expectations questionnaire was presented twice, once at the end of both genres. Ten dollar iTunes gift cards were used as the advertised payment method to thank participants for their time.

Design

This study adopted a within-group design. Stimuli were shown in two blocks, each comprising the 26 image sets belonging to one of the two genres. The order of the

blocks was counterbalanced to avoid serial order effects. Within a web genre, webpage sets were presented in a different random order for each participant. The order of webpages within each website set was constant, as mentioned above. Participants were presented with a practice website set prior to the actual website to familiarize them with the task procedure as shown in Figure 5.1. Practice data were excluded from the data analysis.

Procedure

Participants were briefed on the purpose and the procedure of the session. They were then asked to read and sign the consent form. The researcher then described the activity in more detail using a standard script. Participants were told that the instructions would be repeated on the computer screen in front of them and that they would be able to read them at their own pace. Completion speed was not emphasized and participants were told that there were no right or wrong answers. Pressing the –Start" button on the instruction screen initiated the webpage practice set. After the practice round, participants were given the opportunity to ask questions. At that point, pressing –Next" initiated the formal study. Upon completing all 26 websites in one genre, participants completed the Expectations Questionnaire before proceeding to the second web genre repeating the same procedure. At the end, participants were thanked for their participation again, given a gift card, and excused.

Results

The results are presented as follows. The Expectation Questionnaire results are summarized first. The SUS and VIsAWI-S results are presented last.

Expectations Questionnaire

The first part of the Expectations Questionnaire contained a 23-item checklist asking participants to identify which item they thought would appear in city council and tourism websites. The researcher first counted the actual occurrences of the items from the checklist in all the websites (see Appendix C4 for the occurrence rates per item), per genre. For example, shopping carts appeared in 0% of city council websites and 20% in tourism websites. The actual-occurrence frequencies were compared between genres via two-tailed, paired, Student t-tests and no difference was found. Therefore, there was not enough evidence to suggest that the website items differed significantly between the website genres. The participant results of the expectation checklist were averaged and the two genres were compared via paired t-tests. A difference ($p<.01$) was found in participant expectations of which items occur in which genre, suggesting that participants have different expectations of the two genres. In addition, when the participant responses were compared to the actual frequencies via t-tests, the city

council websites differed ($p<.05$) from the actual ratings whereas the tourist websites did not.

The semantic differential scales were averaged per item in the scale and the two genres were compared using t-tests. Six out of the nine items were found to be significantly different between genres. Specifically, tourism websites were rated to be more enjoyable and prettier than city council websites (all comparisons significant with $p<0.001$). City council websites were rated as more complicated, boring, stressful, and overall worse than tourism websites (all significant, $p<0.001$). Therefore, the results show that the city council websites were expected to offer a worse experience, usability, and visual appeal than the tourism websites.

Perceived Usability and Visual Appeal Correlation

The perceived usability and visual appeal ratings were first aggregated across all websites and both genres. The data were analyzed for outliers (±3SD); none were found. The data were normally distributed and variance was homogenous (please see Appendix C4 for details). Negative statements in the SUS scale were reversed to enable comparison. Pearson‘s Product Moment Correlation Coefficients determined that perceived visual appeal was positively and significantly correlated with perceived usability ($r=.662$, $p<.001$).

Website Visual Appeal

As mentioned earlier, the sample website was chosen based on the visual appeal ratings. The statistical significance of the perceived usability ratings was not taken into account during website selection because ratings were based on short viewing times and objective usability of the chosen websites would be checked in the following two phases.

For city council sites, visual appeal scores ranged from 3.97 to 5.97, and for tourism sites, they ranged from 2.56 to 5.93. Since the initial purpose was to acquire the highest and lowest rated websites in visual appeal, the three best and three worst rated sites from both genres were taken into consideration for further analysis, and can be seen in Figure 5.2. The full list of 52 websites and their ratings can be seen in Appendix C3. The Gold Coast city council website was rated as the highest in visual appeal and the Toowoomba tourism website was rated as lowest. This was not consistent with the findings of the Expectations Questionnaires which found that there is an expectation that tourism websites would have higher visual appeal.

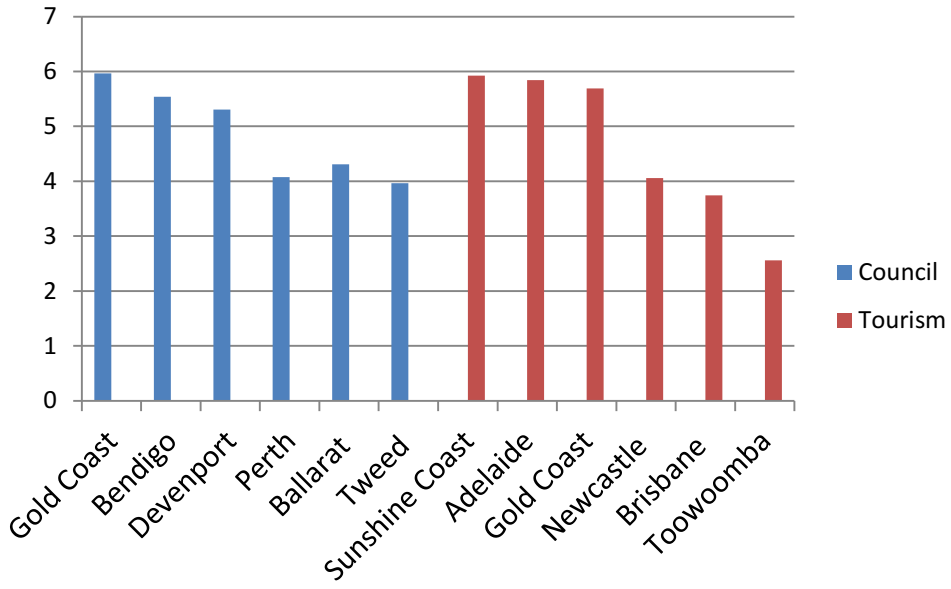


Figure 5.2. Three best and worst rated websites on visual appeal for each genre.

Student *t*-tests performed on the VIsAWI-S data revealed that city council websites differed statistically from tourism websites in both visual appeal ($M_{council}$=4.77, $M_{tourism}$=5.07, p<.001) and perceived usability ($M_{council}$=3.67, $M_{tourism}$=3.77, *p*<.05). Within a genre, statistical differences were found between the top and bottom three tourism websites (*p*<.001) and the top and bottom three council websites (*p*<.001). According to Hirschfield & Thielsch, (2014), the results of two studies suggest that a reasonable cut off point for web developers to strive for on the VisAWI-S is 4.5. Regardless of website genre, this concurs with our findings, since the ratings of the top three websites (all above 4.5), were different from the bottom three (all three were below 4.5).

From the top three city council websites, Gold Coast differed from both Bendigo (*p<.01*) and Devonport (*p*<.001), but the difference between Bendigo and Devonport was not statistically different. There were no significant differences amongst the bottom three city council websites (*p*>.05 for each comparison). Similarly, there were no statistical differences amongst the top three tourism websites. However, for the bottom three tourist websites, Toowoomba differed statistically from both Brisbane (*p*<.001) and Newcastle (*p*<.001), and Brisbane and Newcastle were not statistically different.

Discussion

The purpose of this preliminary study was to select a suitable stimulus domain and website to be used in the remainder of the studies. To control for expectations, we needed a website domain that had less developed mental models, to exclude the influence of past experiences. We found that while the actual occurrence rates of website items from the Expectations Questionnaire did not differ between genres, participants had different expectations of the item-occurrence between the genres, and

the actual versus perceived item checklist values differed for the city council genre. This would suggest that their mental models were less developed for city council websites and that they were guessing when filling in the questionnaires. Thus, the city council website genre was selected for future phases. In addition, participant expectations for the look and feel of these two genres differed, further suggesting that they were less familiar with the genres. Specifically, the visual appeal expectations for city council websites were generally lower than for tourism websites.

One exception was the Gold Coast city council website, which was rated as the highest in visual appeal across both genres. Therefore, this discrepancy between expectations and ratings suggested that the mental models were less developed for the Gold Coast city council website. In addition, 96.6% of participants indicated that they were unfamiliar or only somewhat familiar with city council websites. Of these, the Gold Coast website was unexpectedly highly rated, suggesting that it did not match the expectations for city council websites. Therefore, consistent with the aim of Phase 1, the Gold Coast city council website was empirically selected as the website suitable for future experiments in this thesis.

The lowest rated website in visual appeal was the Toowoomba tourism website. However, the two genres did statistically differ in both perceived usability and visual appeal. Therefore, introducing a different website for the HuLv and LuLv versions of the study would present several confound variables such as information content, placement of objects, image variations, etc. Moreover, it was deemed to require less effort to manipulate a website that was high in visual appeal to make it low, than to manipulate a website that was low in visual appeal to be higher in the trait. Thus, the researchers chose to use only the city council Gold Coast website, and to manipulate it into the remaining website versions for the rest of the studies in this thesis.

Limitations and Future Studies

Threats to construct validity. The expectations questionnaire used in this study was not validated or standardized. The researchers created it solely for the purpose of this preliminary study. This does pose a potential threat to construct validity. However, given that the measures were not created to elicit expectations, and were only used to calculate error rates, then this questionnaire should not pose large risk to construct validity.

Threats to statistical validity. The statistics done in this preliminary study were basic. More in-depth statistics were not necessary since this was a preliminary study and the results were not used to answer the research questions.

Threats to internal validity. In this study, iTunes gift cards were used as the advertised payment method to thank participants for their time. This may inadvertently have biased participant selection towards Apple product users, as other operating system

users may not have had use of the reward. It was therefore decided to offer a generic type of gift card in future studies.

Threats to external validity. Given that the scales were not standardized, obtaining these results in other studies may be difficult. However, the questionnaire is in Appendix C1 and future researchers can use it to examine existing expectations in websites.

Conclusion

Consistent with the aim of this preliminary study, it enabled the empirical selection of a website suitable for manipulation and use in the main studies. While this preliminary study focused on visual appeal, the usability level of the chosen website was assessed in Phases 2 and 3, via analytical and empirical usability tests respectively.

Phase 2: Heuristic Evaluation Introduction

The purpose of Phase 2 was to gain a better understanding of the Gold Coast city council website's usability level by analytically assessing it. The researchers used the "HE++ Evaluation method" described in Chattratichart and Lindgaard (2008) to examine the interface for usability issues. This involved examining the website according to the seven problem areas, entailing graphics, information content, formatting and layout, system efficiency and functionality, navigation, wording, and help and error messages. Each of these areas was examined using Nielsen's heuristics (Ravden & Johnston, 1989; Nielsen, 1993). These consisted of system status visibility, correct use of language and concepts, user control to leave a page at any time, consistency of concepts and language, absence of error, recognisability of items, efficiency of use, ease of recovery from error, and help availability.

Method

Participants

Three Swinburne University researchers participated in the heuristic evaluation (one male and two females). The recommendation for heuristic evaluation is to have three to usability experts evaluate the interface for optimal results (Nielsen & Landauer, 1993). This is recommended because one expert may miss some issues whereas having more than five does not ensure that additional information will be found (Nielsen & Landauer, 1993). Thus, two were professors at Swinburne University of Technology and the third was a PhD student from the same university, all of whom were selected as they were experts in the field of usability. One of the professors was an HCI expert with a psychology background, and the other one was a HCI expert with a technological background. They completed the evaluation together in a two-hour session.

Apparatus and Location

A Macbook Air was used, running 1.3GHz dual-core Intel Core i5 with 4GB of RAM, and a screen resolution of 1024 X 640. The Gold Coast city council website was downloaded using PageNest in June, 2013. The study took place in a quiet and private office at Swinburne University of Technology.

Materials

The researchers used the HE++ Evaluation method (see Chapter 4) and Nielsen's heuristics (also see Chapter 4) to guide them in their usability evaluations. The website under evaluation was the Gold Coast city council website, as described in the first preliminary study.

Procedure

The researchers worked together, using a single computer screen to view the website as they systematically worked through the HE++ guidelines. They considered and discussed each of the seven problem areas using Nielsen‘s heuristics. Each of the problem areas was evaluated separately. Each time, the researchers started at the home screen and examined several webpages until they were all in agreement with the assessment of that problem area. One of the researchers wrote down all the comments and assessments. This procedure is typical of a heuristic evaluation process as described in Chapter 4.

Results

The heuristic evaluation results revealed that Gold Coast‘s website usability level was high. No major usability issues were found and the seven minor ones that were found were easily overcome. The first problem area according to the HE++ method was graphics, and four minor issues were found. Graphics, including symbols, buttons, links, icons, and maps, were first examined. The icons in the menu were used inconsistently. Specifically, some of the icons that appeared next to the links in the menu disappear on pages further in (e.g. library), while others retain their icons on the webpage associated with the link. In addition, some of the metaphors used for the icons were not clear. For example, there was one speech bubble for a link that led to a page that allows an individual to contribute their ideas for the city, and there were two speech bubbles for a page that gives information on how to give the council feedback. The meaning of the speech bubbles was not intuitive. Moreover, the information icon at the bottom of the website was not clickable, which was inconsistent with the other icon‘s functionality. Colour contrasts were well balanced between the background, icons, text, and images. Some consistency issues were found regarding what constituted a link. For example, some titles were clickable while others were not links (see Figure 5.3 for an example). These links also varied in formatting, depending on where they were located on the website, with colours including magenta and black.

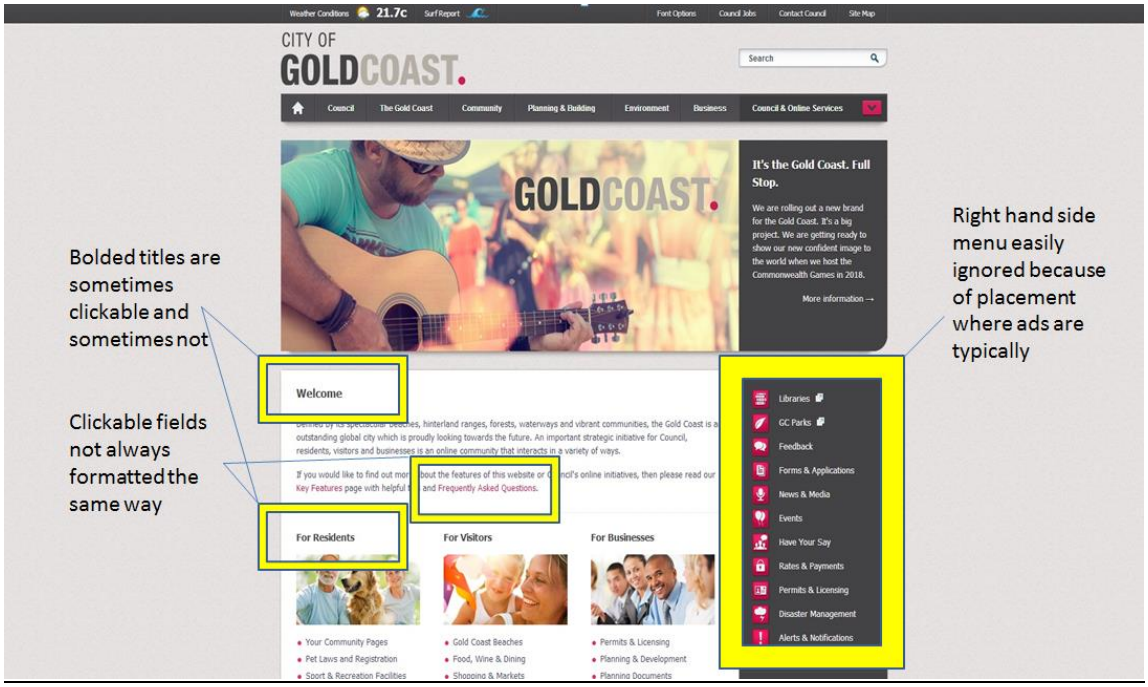


Figure 5.3. Example of possible issues found during HE.

The second area examined was information content and nothing was found to be confusing or needed improvement. Next, one minor issue was found in the formatting and layout of the website. These included font size, white space, and alignment. The font size was adjustable by clicking a button, making it easy to enlarge the font if we needed to. There was enough white space to keep similar information together and different things apart. If two areas were close to each other but irrelevant, then there was a faint line between them to distinguish them. All elements were properly aligned, including text, with one minor exception. On the Library page, deeper into the interface, the first icon on the menu on the right was unaligned by a millimetre with the rest of the icons. Otherwise, the general formatting was consistent throughout the pages in the website. The layout was clean, not cluttered, and was very well organized.

The next elements to be inspected were system efficiency and functionality, including download time. The website was fast to download, taking a second to download over a high speed Internet link. All links within the website downloaded within seconds as well, and all the links functioned. The fifth problem area analyzed was navigation and two minor issues were found. First, there were two main menus, one was on the top of the website, immediately below the city name and it was constant throughout the interface. The other main menu was on the right side which was counter intuitive to one of the researchers at first. This area is generally reserved for ads. However, the researcher later noted that the placement of the menu made sense since the city council website had no ads. Second, the items on this menu were not constant, changing with respect to where you were in the interface, giving the submenu of the main menu at the top. For example, if –Council" was selected in the main menu at the top, the side menu to the right would reveal the subcategories of the –Council" tab. There is a third menu at the bottom of the website that displayed links to council

information pages from anywhere in the interface. In addition, the website had breadcrumbs, telling the user exactly how they got to the page they were on and provided an easy way back to previous pages. In summary, the researchers found that the website was easily navigated.

No issues were found with the wording used in Gold Coast's city council website. The category names and language used were appropriate, without jargon or spelling or grammatical errors. The last area the researchers examined was the help and error messages. While help was not needed for the navigation of this website, when the researchers purposefully searched for a page that was not in the website, the error message displayed was "404 - Page Not Found." There was a list of possible reasons why the page was not found, in addition to links to the home page, the site map, and contact information. Thus, there were no problems found with this area of the HE++.

The severity of the issues found was judged by the researchers to be low and thus the researchers did not want or need to change anything on the website. The potential usability issues found here were tested in the tasks set for participants in Phase 3.

Discussion

The results of the heuristic evaluation done by the three usability experts showed that Gold Coast's website usability level was high. No major usability issues were found. Colour contrasts were well balanced between the background, icons, text, and images. The general formatting was consistent throughout the pages in the website. The layout was clean, not cluttered, and was very well organized. Two minor issues were found with navigation. First, there were two main menus, one was on the top of the website, the other to the right side (where ads are usually found). Second, the items on this right-hand side menu were not constant, giving the submenus. However, once you understood that the menu to the right gave a submenu, then it became easier to use the website. The wording used in the website was clear with little to no jargon, spelling or grammar mistakes found. All menus were properly and clearly labelled. Thus, there were no severe problems found in the website.

Limitations and Future Studies

Threats to construct validity. The measures used in this study were structured and widely used, especially Nielsen' heuristics. Therefore, this study does not have any validity issues when it comes to the constructs in it.

Threats to statistical validity. No statistics were done in this study. However, heuristic evaluations are usually not directly accompanied by statistical testing. Instead, other methods are done with statistics to validate the findings. This is done in the next two preliminary studies.

Threats to internal validity. The three researchers examined the website together. This may have helped them overcome some issues that a single individual may have been stumped by. Yet, grouped expert usability tests are a widely accepted method, as discussed in Chapter 4. Therefore, this study should have no internal validity issues.

Threats to external validity . This study is not inherently generalizable. However, the applicability of the results is tested in the next preliminary study, done with users.

Conclusion

The three researchers used the heuristic evaluation method to examine the usability of the Gold Coast city council website. Since no major errors were found (just the inconvenience of having the second menu bar on the right hand side of the page), the website‘s usability was deemed high. The next step in the usability testing process was to verify these usability findings with user-based testing. Thus, the next phase describes the user-based usability test done on the Gold Coast city council website.

Phase 3: User-Centred Usability Test Introduction

The purpose of this study was to verify the Gold Coast city council website's usability level with users. This was done through a series of 15 randomly ordered information retrieval tasks using the website. The user performance measures were measured with objective usability measures: (1) the number of clicks, (2) the number of times a menu was hovered over thereby revealing the menus content, (3) time per task (4) the number of hints per task, (5) correctness of task answer, and (6) success at completing the task. The number of clicks and the number of hovers were counted, and time to complete the task was recorded. Time was measured in seconds, per task, from when the participant read the question and began to look at the website for the answer, to the time when the participant was on the final webpage from which they concluded the task answer.

Help was given to participants in the form of hints that ranged from vague comments such as "try somewhere else" to very specific comments such as "try clicking [*the name of a button, link, or icon*] and see if that will give you more information". Hints usually started off as vague and got increasingly specific with more time that went by and if the participant was still lost. The first hint was given at one minute into the task, if the participant was not on the right path. The second hint was given 30 to 60 seconds after the first hint, depending on if the participant was still struggling or on the wrong path. The third hint came again at 30 to 60 seconds after the second, also depending on the progression of the participant through the task. The first two hints were counted as 'passes', and three or more hints were automatic effectiveness fails, given that the third was a specific hint that told participants where to click in order to answer the question.

The correctness of the task answer was evaluated as follows. If the participant found and wrote the appropriate answer, then the task was considered as correct. Otherwise, it was incorrect. A task was considered passed if less than 3 hints were given and if the webpage from which the participants concluded their answer was correct. An incorrect answer could have been passed, if the participant was on the correct final webpage given that the incorrect answer was not due to a usability problem, but either to the participant's hurry to finish the task quickly or to a language barrier for the international students. All of these measures were used to gain an understanding of the Gold Coast city council website's usability level.

Method

Participants

A sample of 10 (8 males, 2 females; 8 aged 18-30, two 31+) Swinburne University student volunteers participated, all with 20/20 or corrected to 20/20 vision, and screened for colour blindness. All participants were technology-savvy regular

Internet users. Six participants were not very familiar with the purposes of city councils, three were somewhat familiar, and only one was very familiar. Half were of Australian descent, and the rest were international students of varying English fluency levels. All were tested in individual sessions taking approximately one hour, and all were given a $20 gift voucher at the end of the session.

Apparatus and Location

Participants were tested using a Hewlett Packard desktop computer, running Intel® Core™2 Duo CPU with 3GB of RAM, and a screen resolution of 1290 X 720. The website was presented on Firefox Mozilla. Microsoft Excel and SAS were used to analyze the data. The study took place in the Usability Laboratory at Swinburne University of Technology. The usability lab consists of two rooms that are connected via one-way mirrored glass, with the experimenter being able to view the participant. There are two connected computers, on in each of the two rooms, so that the experimenter can view and record participant-computer interactions.

Materials

All documentation pertaining to Phase 3 is in Appendix B. The same informed consent, demographics questions, SUS, and VisAWI-S used in Phase 1 were also used here. The Gold Coast city council website was downloaded in June 2013 and was tested using a list of tasks prepared by the researcher for this usability test. The list of 15 randomly ordered tasks was given to the participants at the start of the study. Tasks were all goal-oriented tasks, in the form of information retrieval. An example of a task is: "How many beaches are located in the Gold Coast?" All tasks were on the same page. The tasks can be found in Appendix C6. Morae software was used in the Usability Laboratory to record participant interaction with the website.

Procedure

Similarly as in Phase 1, participants were briefed on the purpose and the procedure of the session. This briefing used a documented training script to ensure consistency. They were then asked to read and sign the consent form and fill in their demographic information. Once their details were filled in, participants started on the tasks. The order of tasks was randomized and each task was to be started from the homepage. Task completion speed was not emphasized and participants were told that there were no right or wrong answers. Once the participants finished all of the tasks, they were asked to fill in the SUS and VIsAWI-S. At the end, participants were thanked for their participation, given a gift card, and excused.

Results

The results are presented as follows. The visual appeal results from the VIsAWI-S scales are presented first. Then the perceived usability data from the SUS scale results follow. The objective usability results are presented next. The correlations between visual appeal, perceived and objective usability are presented last.

Visual Appeal

The average visual appeal rating for the Gold Coast city council website was 5.775/7. This is similar to the score (5.966) the website received in Phase 1. However, these scores were given by participants after website use, whereas the scores given in the preliminary study were based on first impression only. In addition, there were only a third of the participants in this phase as there were in Phase 1.

Perceived Usability

The average score for perceived usability was 3.92/5. The score that the website received in Phase 1 was similar (4.29). As was the case with visual appeal, these scores were given by participants after website use, which is in contrast to the first preliminary study, where participants rated the website based on first impressions. Again, there were only ten participants, so the difference in rating could be due to the smaller sample size as well.

Objective Usability

The objective usability measures were: (1) the number of clicks, (2) the number of times a menu was hovered over thereby revealing the menus content, (3) time per task (4) the number of hints per task, (5) correctness of task answer, and (6) success at completing the task.

On average, 13.6 out of 15 tasks were completed, giving a success rate of 90.7%. While the overall success rate was high, these included tasks in which participants received one and two hints. The average number of hints given per task was less than one (i.e. 0.41, with a maximum of three). This suggests that participants did not receive help. When filtering the data by tasks that received no hints, one hint, two hints, and fails (i.e. either did not complete the task or were given three hints). The average number of tasks that were completed: without any aid was 9.8/10, with one hint was 2.9/10, with two hints was 0.9/10, and the average number of tasks per participant that received three hints or that received no hints and were failed was 1.4/10. Thus, the majority were completed without hints, with a very small number of failed tasks (just over one; see Appendix C7 for details).

During the usability evaluation, the researcher noted that non-native English speakers seemed to struggle a little with task completion. To investigate whether

English proficiency influenced the results, these two groups were compared. However, the average completion rate for both groups was 13.6, which was identical to the entire sample mean. Looking at how many tasks were completed by native English speakers and non-native speakers, the results are as follows. There were five participants in each group. Native speakers completed on average 11.4 tasks without help, and completed on average 1.6 tasks with one hint, 0.6 tasks with two hints, and failed 1.4 tasks. Non-native speakers completed 8.2 tasks on average without help, and 4.2 tasks with one hint, 1.2 tasks with two hints, and failed 1.4 tasks. To ensure that the native and non-native speakers indeed differed in their responses, t-tests were run on the data. The results showed that there was no difference between the number of fails (p-value = 1) and the number of times two hints (p-value<0.05) were given. However, number of tasks that received no hints (p-value<0.05) and the number of tasks receiving one hint (p-value<0.05) were significantly different for the two types of participants. Given that both of these were treated as passes, the non-native English speakers were deemed to yield the same results as the native English speakers. Therefore, all of the data was kept and none of the participants were replaced.

To acquire a more detailed understanding of the usability levels per task, the number of clicks and hovers, and task completion time were measured and averaged as well. Per task, the average number of clicks was 3.96, and the average number of hovers was 3.69. This means that upon checking three to four menus, the task was answered within four clicks. The average time to complete a task was 113.19 seconds (00:01:53.19) – just under two minutes. Given that the website domain was unfamiliar, this timing was acceptable and considered adequate for the website.

Visual Appeal, Perceived Usability, and Objective Usability Correlation

The perceived usability and visual appeal ratings were first compiled. The data were analyzed for outliers (±3SD); none were found. The data were normally distributed and variance was homogenous (see Appendix C8 for details). Table 5.1 shows the outcomes of Pearson Product Moment Correlations between the average ratings for visual appeal (Vis.), perceived usability (Usab.), the number of clicks (Click) and hovers (Hover), the correctness of answers (Ans.), the number of hints (Hint), time (Time), and the success rates per task and participant (Pass). Perceived visual appeal was positively and significantly correlated with perceived usability post-use, as seen in Table 5.1.

In addition, perceived usability was strongly and positively correlated with the number of passed tasks, per participant. The average number of hovers per task and per participant was found to be strongly and positively correlated to visual appeal. Likewise, the number of hints and time per task was also significant. This was expected as hints were given at regular time intervals.

Table 5.1. Correlations between all measured variables in Phase 3.

Metric	Vis.	Click	Hover	Ans.	Hint	Time	Pass
Usab.	.776**	.017	.281	-.182	-.562	-.149	.733**
Vis.		-.157	.719*	-.346	-.372	.212	.394
Click			.036	-.020	.087	-.003	.393
Hover				-.111	.117	.426	.296
Ans.					.028	-.609	.167
Hint						.677*	-.210
Time							-.144

*Correlation significant at .05 (two-tailed).
**Correlation significant at .01 (two-tailed).

Discussion

The results of Phase 3 confirmed the selection of the city council website as the HuHv stimuli, given that the average score for the Gold Coast city council website for perceived usability was 3.92/5, and for visual appeal rating it was 5.775/7. In addition, 13.6 out of 15 tasks were completed, giving a success rate of 90%. Participants found answers within four clicks and hovers, and within two minutes per task. Given that the website domain was unfamiliar, this timing was acceptable and considered adequate for the website.

Limitations and Future Studies

Threats to construct validity. The metrics used in this study were all validated and recognized measures. They are widely used in the field. There were thus no threats to construct validity in this section.

Threats to statistical validity. Ten participants were used to test the usability of the website. As previously mentioned in Chapter 4, five participants find approximately 75% of all interface errors and 15 participants find 100% of them (Nielsen & Landauer, 1993). Therefore, ten participants were deemed to be enough to find nearly all usability issues.

Threats to internal and external validity. Since English proficiency was not found to be a barrier in this study, no other threats to internal validity are evident. The results of this usability test are as generalizable as any other user-based usability test. Thus, there were also no threats to external validity.

Conclusion

Altogether, these results suggest that the website usability level was indeed high, as found by the heuristic evaluation in Phase 2. Moreover, the visual appeal was high as deemed by the results of Phase 1. Therefore, no adjustments were made to the Gold

Coast city council website. The original was made the easy and pretty (HuHv) website condition. The data collected in this study (perceived and objective usability and visual appeal results) were used as a basis for comparison in Phase 4 to verify the success of the HuLv, LuHv, and LuLv website manipulations.

Phase 4: Manipulation and Verification via Usability Test

Phase 2 and 3, which involved the heuristic evaluation and the user-based usability test of the original Gold Coast city council website, supported the selection of the website to be used as the HuHv version of the website for the main study. Future studies need HuLv, LuHv, and LuLv versions of the website. Therefore, there were two purposes of this phase: to manipulate the original website in order to create LuHv, HuLv, and LuLv versions of the website, and to test and verify the manipulations with users. The details to the website manipulations are in this phase‘s Method section, below. Thus, there were three research hypotheses: (1) visual appeal was successfully lowered in the HuLv and LuLv versions, (2) perceived usability was successfully manipulated to be worse in the LuHv and LuLv versions, and (3) objective usability measures in the LuHv and LuLv versions are indeed lower than the HuHv version. The first and second research hypotheses only have one corresponding statistical hypothesis each which can be seen in Table 5.2.

Table 5.2. Visual appeal and perceived usability statistical hypotheses and tests used.

Statistical Hypotheses	Test
H_1: The visual appeal ratings of the manipulated versions differ from the original HuHv version. 1a. The visual appeal of HuLv would be lower than HuHv. 1b. The visual appeal of LuLv would be lower compared to HuHv.	1. ANOVA, F test Main effects & Simple main effects analysis. 1a/b. Tukey post-hoc multiple comparisons.
H_2: The perceived usability of the manipulated versions differs from the original version. 2a. The perceived usability of LuHv would be lower than HuHv. 2b. The perceived usability of LuLv would be lower compared to HuHv.	2. Independent-Samples Kruskal-Wallis Test, 2 a/b. Kruskal-Wallis multiple comparrison tests.

The third research hypothesis has six statistical hypotheses as there were multiple measures for objective usability, and can be seen in Table 5.3. Thus, to test all three research hypotheses, a total of eight statistical hypotheses were tested. Each website version was compared to the HuHv version for statistical significance. For the first research hypothesis, if the manipulation was done correctly, then the perceived visual appeal level from the VisAWI-S of the HuLv and LuLv websites should be lower than it is in the HuHv website. Similarity, for the second research hypothesis, the statistical hypothesis is that the perceived usability level from the SUS scale of the LuHv and LuLv website versions should be lower compared to the HuHv version. For the third research hypothesis, the six objective usability measures in the LuHv and LuLv website versions should vary significantly from the HuHv version.

Table 5.3. Objective usability statistical hypotheses and tests used.

Statistical Hypotheses	Test
H_3: The average number of clicks per task of the high and low usability website versions would differ.	3. Same as 2.
3a. Clicks of LuHv would be higher than HuHv.	
3b. Clicks of LuLv would be higher than HuHv.	
H_4: The average number of hovers per task would differ between the high and low usability website versions.	4. Same as 2.
4a. Number of hovers of LuHv would be higher than HuHv.	
4b. Number of hovers of LuLv would be higher than HuHv.	
H_5: The average number of correctly answered tasks per participant would be different between the low and high usability conditions.	5. Same as 2.
5a. The average number of correctly answered tasks per participant would be lower in LuHv than in HuHv.	
5b. The average number of correctly answered tasks per participant of LuLv would be lower compared to HuHv.	
H_6: The average number of hints per task would differ between the high and low usability website.	6. Same as 2.
6a. The average number of hints offered per task would be higher in LuHv than in HuHv.	
6b. The average number of hints offered per task of LuLv would be higher compared to HuHv.	
H_7: The average task completion time would differ between the low and high usability versions.	7. Same as 1.
7a. Time to complete each task in LuHv would be higher compared to HuHv.	
7b. Time to complete each task of LuLv would be higher compared to HuHv.	
H_8: The average number of tasks passed per participant would differ between the high and low usability website versions.	8. Same as 2.
8a. Tasks passed of LuHv would be lower than HuHv.	
8b. Tasks passed of LuLv would be lower than HuHv.	

Method

Participants

There were a total of 30 participants (22 male, 8 female). Out of these, 22 were aged 18-30, and eight were aged 31 or over. Thirteen were born in an English-speaking country, the remaining 17 were not. Twenty-six participants out of the 30 new ones tested in this phase indicated that they used the internet regularly for banking, 22 for shopping, all 30 for entertainment, 29 used it for study purposes, 26 for news, 26 used it for social, and 11 used the internet for travel purposes. When asked about their familiarity with the purposes of city councils, 14 indicated that they were not very familiar, 13 were somewhat, and only 3 were very familiar.

Please note that the statistics shown in the results section contain 40 participants. The extra 10 participants were added from the previous phase with the HuHv website, for the purpose of comparison. Therefore, there are a total of 40 participants in the

analysis below. Each participant did the usability test individually, and took approximately one hour.

Materials, Apparatus, and Location

All materials used for this study that differ from previous studies can be found in Appendix B. A demographic questionnaire was administered to determine the participants‘ background information. The SUS and VisAWI-S scales (also used in the previous phases) were given to participants to measure perceived values of usability and visual appeal. The presentation of these scales was counterbalanced. Also as in the previous phase, Morae was used to record participant interaction with the website. The same set of 15 randomly ordered tasks used in the previous phase was also given to the participants here. SPSS was used to calculate the statistics in the results. The same laboratory was used as in the previous phases.

The HuHv version of the website was manipulated in five ways, two manipulations for usability (Appendix C9) and three for visual appeal. To lower usability, (1) the titles in the top menu bar, in the form of tabs, were changed. This was done by randomizing the items in the top menu bar in order to alter the consistency and simplicity of the menu. With each click on the website, all of the titles in the menu bar would change. The titles were changed to synonyms of the original title, where some synonyms were not as intuitive as others in their application to a council website. Specifically, the title of the ―Council" tab would randomly change to one of: Board, Assembly, Committee, Congress, Politics, Government, Law, or Jury. With each click, the menu called ―The Gold Coast" would change to one of: The City, Streets, Miscellaneous, About, or Life. The ―Community" menu would change to one of: People, Public, Us, Neighbourhood, Open, or Civic. The menu title for ―Planning and Building" was randomized to any of: Development, Infrastructure, and Brick by Brick. The ―Environment" menu was changed to one of: Parks and Beaches, Nature, Flora and Fauna, Setting, Surroundings, Atmosphere, or Biosphere. The menu title of ―Business" was randomized to: Job, Stocks, Money, Corporate, Professional, and Commerce. The last menu, called ―Council and Online Services" was changed to: Law, Help, Rules, and Services. To further lower the usability level, (2) item contrast in the dropdown menu that appeared when a user hovered over the top menu bar was lowered. The dropdown menus were originally long and had multiple categories of submenus to choose from. These dropdown menus were altered by removing the contrast between the titles and the other menu options. This was done by un-bolding and un-underlining the titles, thus removing the titles and creating uncategorized lists under the menus.

The visual appeal was altered by (3) changing the main background colour from the original grey-beige to lilac and (4) changing the texts‘ background from a light off-white to evergreen. We left the textual background white so that contrast with text would be the same and so that usability would not be affected. Just the colours in the exterior background were changed, so that only visual appeal was affected (in response to Bartuskova & Krejcar‘s (2013) criticism of Tuch et al., 2012). The combination of

lilac and evergreen was chosen since it was the colour scheme of the Toowoomba tourism website, rated worst in the Phase 1. In addition, (5) the colours of all the images in the website were inverted to be negatives. Please see Figure 5.4-5.7 in order to see the different versions of the website.

Figure 5.4. The original HuHv website, used for comparison of manipulations.

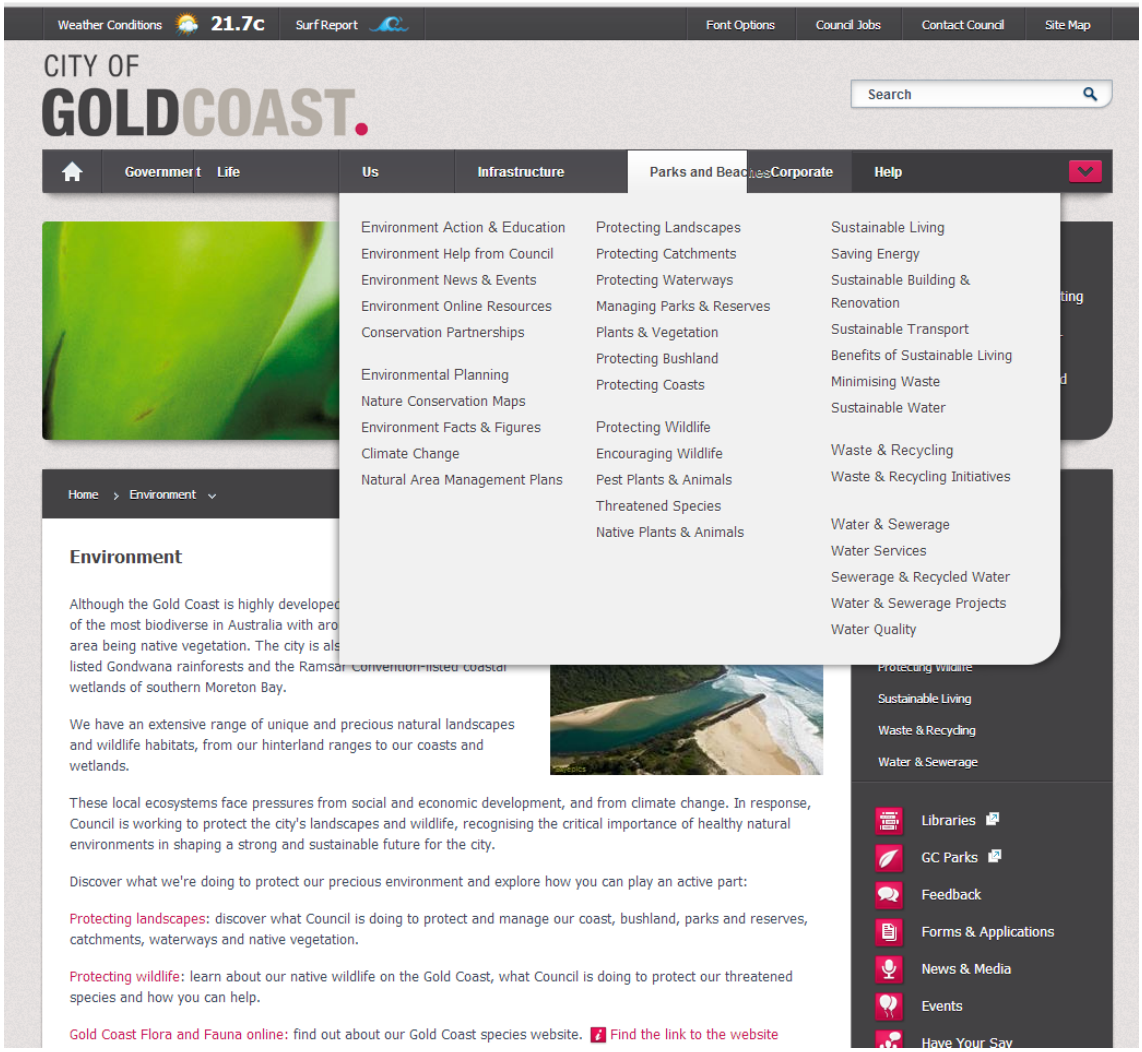


Figure 5.5. The LuHv version with an example of the new menu.

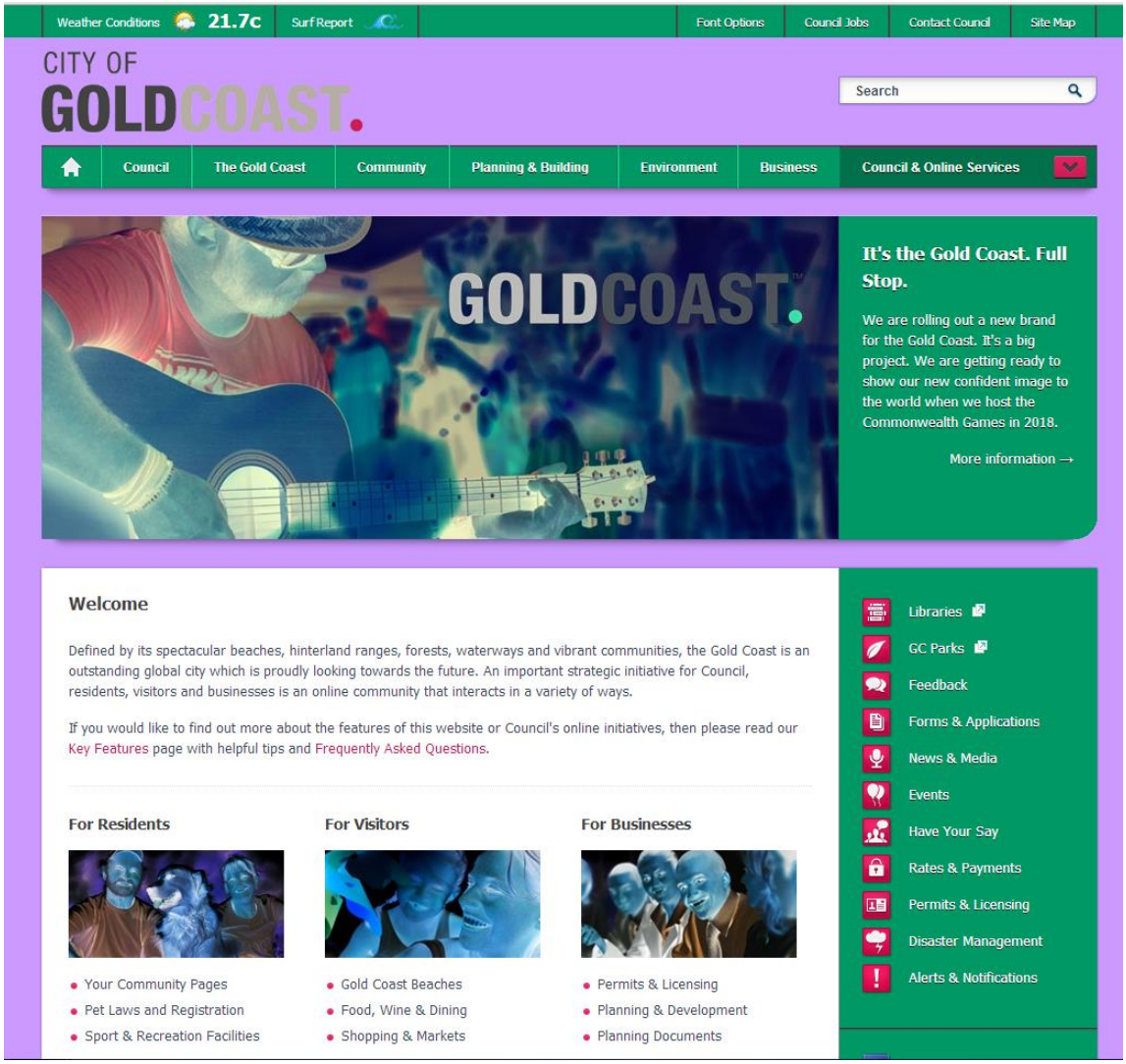


Figure 5.6. The HuLv website.

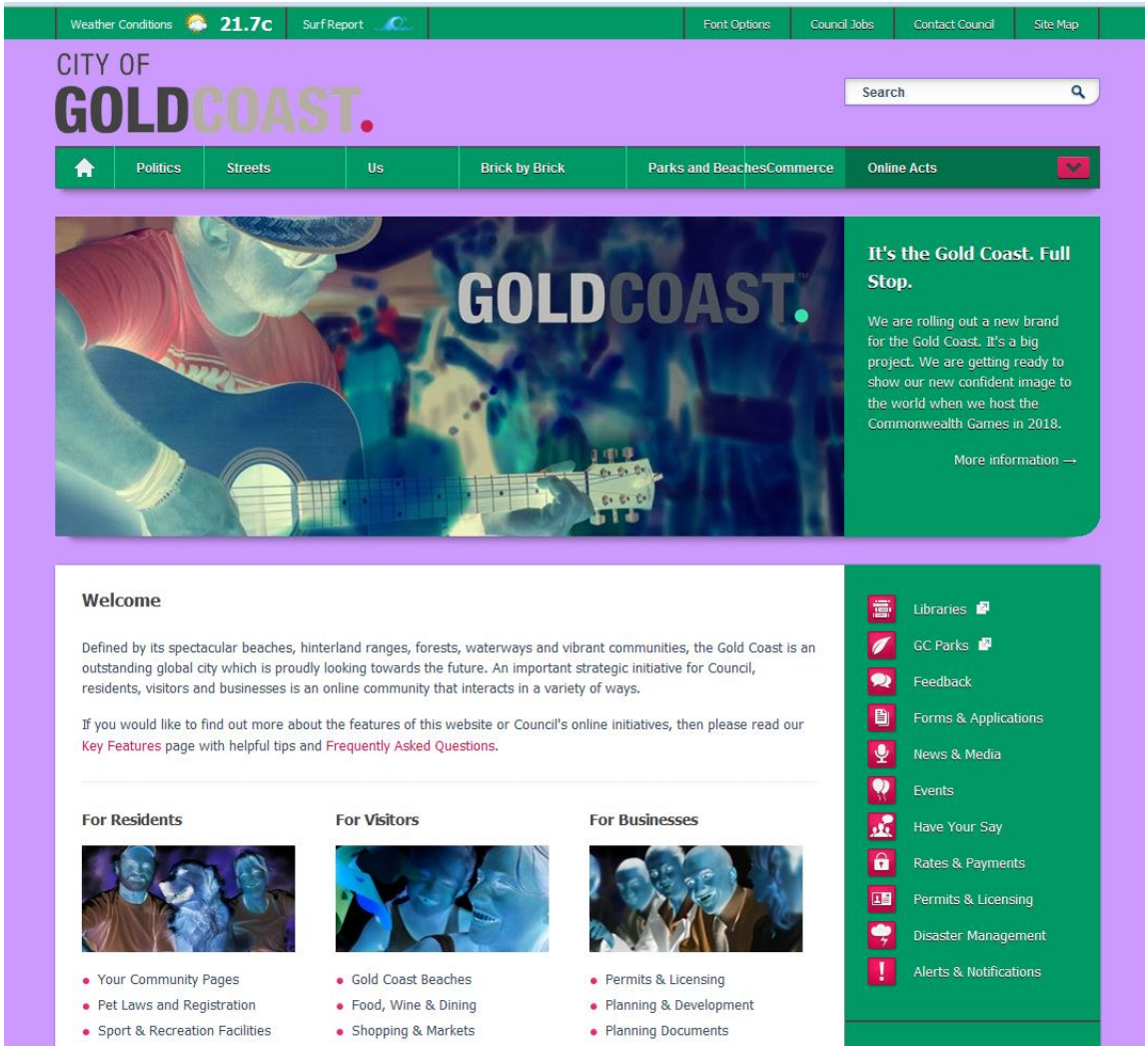


Figure 5.7. The LuLv website.

Procedure

Similarly as in Phase 1 and 3, participants were verbally briefed on the purpose and the procedure of the session. This briefing used a documented training script to ensure consistency. They were then asked to read and sign the consent form and fill in their demographic information. Once their details were filled in, participants started the test, by commencing on the tasks. The order of the 15 tasks was randomized and each task was to be started from the homepage. Task completion speed was not emphasized and participants were told that there were no right or wrong answers. Once the participants finished all of the tasks, they were asked to fill in the SUS and VIsAWI-S scales. At the end, participants were thanked for their participation again, given a gift card, and excused.

Results

The results are presented as follows. The demographic responses are summarized first. These are followed by the usability results. The results section concludes with the visual appeal results. The full statistical results for this section are in Appendix C8.

Verification of Statistical Assumptions

Normality. To test to see if the data are normally distributed, visual and numerical data must be examined (Lofgren, 2000). Numerically, there are the z-values from the skewness and kurtosis tests and the p-values from the Shapiro-Wilk test. Visually, histograms, the normal Q-Q plots, and box plots need to be examined. Checking these values at each level of the independent variables gives whether or not the data approximate the normal distribution. To get the z-value from the skewness and kurtosis tests, the statistic is divided by the standard error, the result of which should be less than $|1.96|$ (Shapiro & Wilk, 1965; Lofgren, 2000; Razali & Wah, 2011).

The z-values revealed that perceived usability in the LuHv website, hovers in the LuHv version, and hints in the HuLv website were larger than $|1.96|$. In addition, the p-values from the Shapiro-Wilk test, for perceived usability and visual appeal, revealed that the perceived usability at the HuHv website ($p<0.05$) was not normally distributed. In objective usability, the all four website versions in the average number of correct answers per participant ($p<0.05$), the HuLv website in the average number of hints per participant ($p<0.01$), and HuLv website version in the average number of passes per participant ($p<0.05$) were all not normally distributed. All other measures were normally distributed and the histograms, normal Q-Q plots, and box plots confirmed this.

Homogeneity of Variances. Given that visual appeal and time per task were normally distributed, the parametric Levene‘s tests were performed to test the equal variance assumption (Martin & Bridgmon, 2012; Lofgren, 2000). These tests found that both had equal variances. Given that the rest of the variables did not approximate the noramlity curve, the nonparametric Levene‘s test was conducted (Nordstokke & Zumbo, 2010; Nordstokke, Zumbo, Cairns, & Saklofske, 2011; Lofgren, 2000). These revelaed that all of the variances were homogenous.

Hypothesis Testing

Hypothesis 1: Visual appeal of the manipulated versions differs from the original version. The means for visual appeal were 5.775 for HuHv (from the previous phase), 4.85 for LuHv, 4.22 for HuLv, and 3.775 for LuLv. These can be seen in Figure 5.8. Since visual appeal was both normally distributed and had equal variance, an ANOVA was conducted to test research Hypothesis 1. Main effects were found ($p<0.01$). Given this finding, the remainder of Hypothesis 1 can be tested.

Hypothesis 1a: The visual appeal of HuLv would be lower than HuHv. To examine this, Tukey post-hoc tests were performed. The difference was significant ($p<0.05$) between HuLv and HuHv, rejecting the null hypothesis. *Hypothesis 1b: The visual appeal of LuLv would be lower compared to HuHv*. The Tuckey test revealed that the difference was also significant ($p<0.01$), also rejecting the null hypothesis.

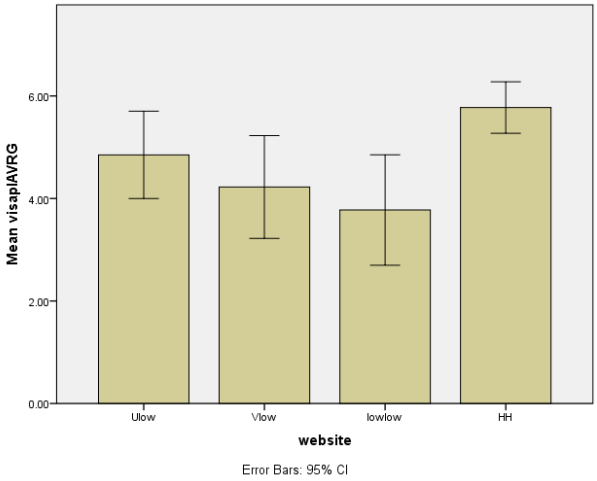


Figure 5.8. The mean visual appeal rating per website version in Phase 4.

Hypothesis 2: The perceived usability of the manipulated versions differs from the original version. The means for perceived usability were 3.84 for HuHv (from the previous phase), 3.26 for HuLv, 2.67 for LuHv, and 2.42 for LuLv. These can be seen in Figure 5.9. Since perceived usability was not normally distributed but had equal variance, Kruskal-Wallis, a non-parametric test, was conducted to test Hypothesis 2. The Kruskal-Wallis test is a multiple-comparison test that finds differences between group means or medians. A main effect was found ($p<0.01$) by the Independent-Samples Kruskal-Wallis Test.

Hypothesis 2a: The perceived usability of LuHv would be lower than HuHv. The difference was significant ($p<0.05$) between HuLv and HuHv, rejecting the null hypothesis. *Hypothesis 2b: The perceived usability of LuLv would be lower compared to HuHv.* The Kruskal-Wallis test revealed that the difference was also significant ($p<0.01$), also rejecting the null hypothesis.

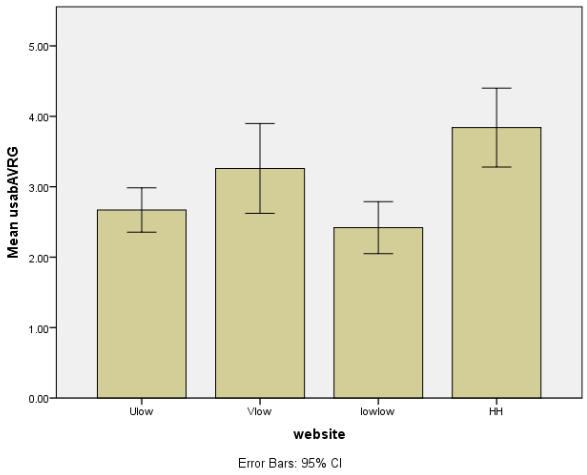


Figure 5.9. The mean usability result per website version.

Hypothesis 3: The average number of clicks per task the high and low usability website versions will differ. No main effects were found. *Hypothesis 4: The average number of hovers per task will differ between the high and low usability website versions*. No main effects were found.

Hypothesis 5: The average number of correctly answered tasks per participant will be different between the low and high usability conditions. The Independent-Samples Kruskal-Wallis Test rejected the null hypothesis ($p<0.05$) for main effects. *Hypothesis 5a: The average number of correctly answered tasks per participant will be lower in LuHv than in HuHv.* No difference was found. *Hypothesis 5b: The average number of correctly answered tasks per participant of LuLv will be lower compared to HuHv.* No difference was found.

Hypothesis 6: The average number of hints per task the high and low usability website versions will differ. Independent-Samples Kruskal-Wallis Test ($p<.01$) rejected the null. *Hypothesis 6a: The average number of hints per task will be higher in LuHv than in HuHv.* Pairwise comparisons found that they did differ ($p<0.05$). *Hypothesis 6b: The average number of hint per task of LuLv will be higher compared to HuHv*. No difference was found. *Hypothesis 7: The average time per task will be higher in the low usability versions than in the high*. No main effect was found, thus no differences can be concluded. *Hypothesis 8: The average number of tasks passed per participant will differ between the high and low usability website versions*. No main effects were found.

Therefore, the perceived usability and visual appeal values were statisticially different between the high and low website versions. However, the majority of the objecitve usability measures were did not differ significanly between website versions, as can be seen in Table 5.4.

Table 5.4. Hypothesis testing summary.

Hypothesis	Result
1. The visual appeal ratings of the manipulated versions differ from the original HuHv version. 1a. The visual appeal of HuLv would be lower than HuHv. 1b. The visual appeal of LuLv would be lower compared to HuHv.	1. Null rejected ($p<0.01$) 1a. Null rejected ($p<0.05$) 1b. Null rejected $p<0.01$)
2. The perceived usability of the manipulated versions differs from the original version. 2a. The perceived usability of LuHv would be lower than HuHv. 2b. The perceived usability of LuLv would be lower compared to HuHv.	2. Null rejected ($p<0.01$) 2a. Null rejected ($p<0.05$) 2b. Null rejected $p<0.01$)
3. The average number of clicks per task the high and low usability website versions would differ.	3. Null not rejected
4. The average number of hovers per task would differ between the high and low usability website versions.	4. Null not rejected
5. The average number of correctly answered tasks per participant would be different between the low and high usability conditions. 5a. The average number of correctly answered tasks per participant would be lower in LuHv than in HuHv. 5b. The average number of correctly answered tasks per participant of LuLv would be lower compared to HuHv.	5. Null rejected ($p<0.05$) 5a. Null rejected ($p<0.05$) 5b. Null rejected $p<0.01$)
6. The average number of hints per task the high and low usability website versions would differ. 6a. The average number of hints per task would be higher in LuHv than in HuHv. 6b. The average number of hint per task of LuLv would be higher compared to HuHv.	6. Null rejected ($p<.01$) 6a. Null rejected ($p<0.05$) 6b. Null not rejected
7. The average time per task would be higher in the low usability versions than in the high.	7. Null not rejected
8. The average number of tasks passed per participant would differ between the high and low usability website versions.	8. Null not rejected

Discussion

Perceived visual appeal ratings differed between the high visual appeal website version and the low. This suggests that visual appeal was successfully manipulated. Perceived usability was also rated as different between the high and low levels, suggesting that it was also successfully manipulated. However, when examining the objective usability measures, the results are not as convincing. The majority of the objective usability measures were not statistically different between the high and low usability websites, which indicates that participants did not actually struggle more when completing the usability test. Given that the goal of this thesis is to examine the impact of expectations, controlling for usability and visual appeal, it is imperative to have the objective as well as the subjective usability levels differ statistically.

Therefore, the next section involves the re-manipulation and re-testing of the low website version. Limitations and Future Studies

Threats to construct validity. The scales and measures used are widely used and accepted in the HCI literature so there should be no validity issues with regards to constructs.

Threats to statistical validity. This study had 10 participants per condition. This may pose a threat to statistical validity throughout the entire thesis. However, in each case, statistical results have been significant suggesting that 10 were enough to determine differences between conditions. In future studies, should time and money allow it, more participants should be recruited.

Threats to internal validity. The majority of the objective usability measures were not statistically different between the high and low usability websites, which indicates that participants did not actually struggle more when completing the usability test. This suggests that the original manipulations to lower usability were not strong enough to hinder use. The next study will re-manipulate and re-test the usability aspect of the Gold Coast city council website.

Threats to external validity. Given the purpose and results of this preliminary study, the results could not readily be generalized.

Conclusion

While this study showed that visual appeal was successfully manipulated to be worse (i.e. uglier) than the original Gold Coast city council website, objective usability was not. Therefore, the next preliminary study includes the description of the re-manipulation of usability and the re-testing of the low usability version of the website in order to ensure that usability was indeed worse between the high and low usability website versions.

Phase 5: Re-manipulate and Re-test

The purpose of this phase was to re-manipulate and re-test the low usability versions of the website, to make the objective usability measures significantly less usable than they are in the HuHv version. While the usability manipulations were done on both LuHv and LuLv versions, only the LuHv version of the website was tested with participants. Tuch and colleagues (2012) altered the usability of their website by changing the labels on three levels (the main menu and two submenus) but not on the final, actual answer page. The same was done in this phase, with the addition of randomizing the labels so that the labels along the completion paths for each task changed with every click on the website. In addition, the location of each of the menu items that were needed in order to complete the tasks were randomly scattered in the menu system so that the menu had no real categories. For example, for the link to the ―Pet Registration" was put under ―Business" rather than the ―Community" tab. Synonyms were used for the titles as well. For example, ―Council Rates" was changed to ―Assembly Taxes" and ―Jury Fees," and ―Beaches & Foreshores" was randomized to ―Sand and Cliffs" and ―Seawater & Fjords". For a complete list of these changes, please see Appendix C10. The hypotheses are can be seen in Table 5.5.

Table 5.5. Hypothesis number, predictions, and tests used.

Hypothesis	Test Used
1. The average number of clicks per task of LuHv would differ from HuHv.	1. ANOVA, F-test.
2. The average number of hovers per task would differ between LuHv and HuHv.	2. Kruskal-Wallis Independent Samples Test.
3. The average number of correctly answered tasks per participant would differ between LuHv and HuHv.	3. Kruskal-Wallis Independent Samples Test.
4. The average number of hints per task w would ill differ between LuHv and HuHv.	4. Kruskal-Wallis Independent Samples Test.
5. The average time per task would be lower in LuHv than HuHv.	5. ANOVA, F-test.
6. The average number of tasks passed per participant would differ between LuHv and HuHv.	6. Kruskal-Wallis Independent Samples Test

Method

Participants

Ten participants (6 male, 4 female) volunteered for this phase. Out of these, six were aged 18-30, and four were aged 31 or over. All were born in an English-speaking country. The statistics shown in the results section contain 20 participants. As was the case in Phase 4, in this phase, the extra 10 participants were added from Phase 3 with

the HuHv website, for the purpose of comparison. Therefore, there are a total of 20 participants in the analysis below. Each participant did the usability test individually, and took roughly one hour.

Materials, Apparatus, and Procedure

The materials used in this phase were the same as the ones in Phase 4, with one exception. The LuHv website was re-manipulated, as mentioned in the introduction. Given that the purpose of this phase was to make the website usability worse, the apparatus and procedure were identical to ones used in Phase 4.

Results

Demographics

Five out of the ten participants tested in this phase indicated that they used the internet regularly for banking, five for shopping, all five for entertainment, eight used it for study purposes, seven for news, five used it for social, and three used the internet for travel purposes. When asked about their familiarity with the purposes of city councils, five indicated that they were not very familiar, four were somewhat, and only one was very familiar.

Verification of Assumptions

Normality. The z-values were larger than $|1.96|$ for hovers in the LuHv website, in both skewness and kurtosis. Similarly, hints had z-values that were larger but only for skewness. In addition, the p-values from the Shapiro-Wilk test, for hovers ($p<0.01$), correctness of the answer in the HuHv version ($p<0.01$), and hints LuHv ($p<0.05$) were not normally distributed. All other measures were normally distributed and the histograms, normal Q-Q plots, and box plots confirmed this.

Homogeneity of Variance. Given that time per task, clicks, and passes were normally distributed, then parametric Levene‘s tests were performed to test the equal variance assumption. These tests found that both had equal variances. Given that the rest of the variables did not approximate the noramlity curve, the nonparametric Levene‘s test was conducted. These revelaed that all of the variances were homogenous.

Hypothesis Testing

Hypothesis 1: The average number of clicks per task of LuHv will differ from HuHv. The null was rejected ($p<0.05$), meaning that the number of clicks differs between the two websites.

Hypothesis 2: The average number of hovers per task will differ between LuHv and HuHv. Hovers did differ ($p<0.05$) between the two website versions.

Hypothesis 3: The average number of correctly answered tasks per participant will differ between LuHv and HuHv. The number of correct answers differs, the null was rejected ($p<0.001$).

Hypothesis 4: The average number of hints per task will differ between LuHv and HuHv. The average number of hints does differ ($p<0.001$) between the two website versions.

Hypothesis 5: The average time per task will be lower in LuHv than HuHv. The null was rejected ($p<0.05$), meaning that the average times do differ between the high and low conditions.

Hypothesis 6: The average number of tasks passed per participant will differ between LuHv and HuHv. The number of passed tasks was significantly different ($p<0.001$) between the websites.

No correlations were found between perceived usability and visual appeal, which was expected given that only the usability was manipulated to be lower and visual appeal was left to be high in this website version. This suggests that the website manipulations of visual appeal and usability were done independently.

Discussion

In this phase, objective usability differed significantly between the LuHv and HuHv website versions. These results, taken together with the results of Phase 4, strongly suggest that the usability manipulation was successful and that the result of Chapter 5 was four significantly different versions of the same website. These website versions would be used in the next three studies, to examine if expectations influence visual appeal and/or usability.

Limitations and Future Studies

Threats to construct validity. In this chapter, participants in the user-based usability tests worked on completing 15 tasks with the given website version. In the main studies, this task list is reduced to 10 questions. Please see Appendix D1 in order to see the new task list. Tasks were removed to shorten the tests in order for them to fit into an hour, given the addition of pre- and post- ratings, and given the introduction of expectations. Tasks were removed statistically, keeping the tasks that kept the statistical difference between website conditions. The process of removal was done manually, using a backward elimination process. Starting with all of the tasks, a task was removed, one at a time, and tested for significance against the original test with all tasks. Each task that did not impact the significant results in the differences between website versions (i.e. when taken out, the remaining set of tasks still maintained the significance of the original 15 tasks), was removed. This was done and all of the tasks that could be

removed, were. In the end, five tasks were removed. This task reduction also ensures that participants are not bored or tired towards the end of the study.

While qualitative data was not recorded for the preliminary studies, the researcher noted that some participants commented on their mood (i.e. mainly frustration) in Phase 5. Therefore, mood was measured pre- and post-use of the website but not manipulated in this study. It was measured using the Self-Assessment Manikin (SAM; Bradley & Lang, 1994). This scale was chosen as it is a short questionnaire with only three image-based scales asking participants to circle how they currently feel.

Threats to statistical validity. As was the case with the previous study, this study only had 10 participants per condition. This is a relatively small number of participants. However, even with 10 participants, we managed to acquire significantly different results between the website versions. Therefore, future studies in this thesis also had 10 participants per condition.

Threats to internal validity. Mood may have been an unaccounted, possibly confounding variable in this study. The next study examined its impact on visual appeal and usability.

Threats to external validity. This study‘s results are not easily generalizable because it was testing the difference between website versions in order to ensure that there was a difference. Usually, studies examine usability in order to determine what the usability level is, whereas we needed to verify that in indeed was as hard as we needed it to be. Therefore, this study‘s results could be generalizable to other studies of the same nature.

Summary

In summary, throughout the preliminary studies, we have established the Gold Coast website as the data sample. It was manipulated into three other versions to a grand total of foud different website versions. These versions were tested and statistically differed in visual appeal and perceived and objective usability. These website versions were used in the first experimental study, in the next chapter. Main Study 1 and 2 were two separate experiments which examined the congruent visual appeal and usability conditions, with textual, and then verbally and textually implemented expectations.

Chapter 6. Congruent Visual Appeal and Usability Levels

Main Study 1 Introduction

The preliminary studies in Chapter 5 resulted in the acquisition of a relatively unfamiliar website genre and website: The Gold Coast city council website. The website's usability level was tested using both users and experts and was deemed to be visually appealing (based on the first preliminary study) and highly usable (based on the second and third preliminary studies). The website was then manipulated into three more versions that varied in visual appeal and usability levels. Thus, in total, there were four website versions: easy/ugly (HuHv), easy/ugly (HuLv), hard/pretty (LuHv), and hard/ugly (LuLv). The studies in this chapter used the HuHv and LuLv versions of the website, where visual appeal and usability levels were congruent. The purpose of the first and second main studies was to see if expectations influenced the visual appeal, perceived and objective usability of a website, when usability and visual appeal levels are congruent. In particular, this study strived to answer the first five research questions (from the original six in the Introduction):

1. Do expectations influence visual appeal?
2. Are perceived and objective usability influenced by expectations?
3. What effect do textual expectations have on visual appeal and usability?
4. What effect do verbal expectations have on usability and visual appeal?
5. What happens when visual appeal and usability levels are congruent (i.e. are both either high or both are low)?

To test these, the HuHv and LuLv website versions were subjected to three expectation conditions: high expectation of visual appeal and usability (He), low in both (Le), and no expectations (Ne) which was the control condition. Thus, there were six conditions in this phase: (1) HuHv website with HuHv expectations, HuHvHe, (2) HuHv website with LuLv expectations, HuHvLe, (3) HuHv website with no expectations, HuHvNe, (4) LuLv website with HuHv expectations, LuLvHe, (5) LuLv website with LuLv expectations, LuLvLe, and (6) LuLv website with no expectations, LuLvNe.

There were six research hypotheses tested by the experiments in this chapter, based on the theory of cognitive dissonance. As mentioned earlier, cognitive dissonance is a disagreement of information which may cause stress (Festinger, 1957). When dissonance occurs, an individual strives to achieve consonance by reducing the inconsistency. The cognition that is most resistant to change is typically the one associated with the most recent behaviour/event (Harmon-Jones et al., 2009). In the case of this thesis, the most recent behaviour would be the act of reading a task description with a set expectation which would vary depending on the experimental condition. Therefore, if expectations influence visual appeal and usability, then participants should agree to the expectation given, and the perceived variables should be reported as either higher or lower than the control condition, in accordance with the expectation level. To

examine these hypotheses, causational and correlational statistics were computed, along with brief qualitative analysis.

The *first hypothesis* states that when expectations of visual appeal and usability are set to be high, then participants will perceive and rate these two variables to be higher than the control and low-expectation conditions. This hypothesis applies to the HuHvHe condition when it is compared to HuHvLe and HuHvNe, and similarly with LuLvHe when it is compared to LuLvLe and LuLvNe. Higher ratings are expected because participants will be, perhaps subconsciously, swayed to increase their ratings to reduce the inconsistency in order achieve consonance.

Similarly, lower ratings are expected because participants decrease their ratings to reduce the cognitive dissonance and agree with the expectations. Thus, the *second research hypothesis* states that when expectations are set to be low, then participants will perceive and rate them to be lower than the control and high-expectation conditions. This hypothesis applies to HuHvLe condition when it is compared to HuHvHe and HuHvNe, and similarly with LuLvLe when it is compared to LuLvHe and LuLvNe.

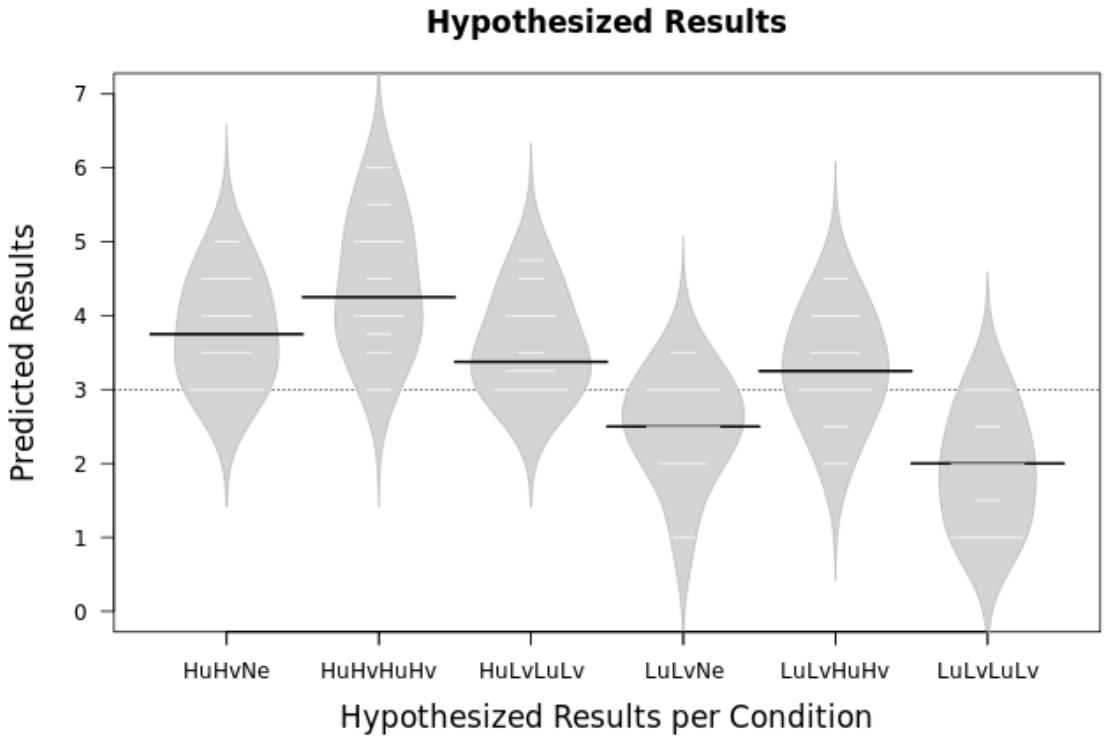


Figure 6.1. Beanplot of the hypothesized results.

The hypotheses are visualized in Figure 6.1, using dummy data. In Figure 6.1, the first three beans (to the left) correspond to the high usability, high visual appeal website, and the last three beans are the low usability and low visual appeal website conditions. The first bean is the control condition for the easy/pretty website. The second bean is higher since it represents the easy/pretty expectation which was predicted to be rated as better than the control and low expectation conditions. Then, the third bean is the lowest

for the easy/pretty website since it pertains to the low expectations. The fourth bean represents the hard/ugly website control condition. The fifth bean is the hard/ugly website with easy/pretty expectations; since the expectation is high then it was predicted to be perceived as easier to use and prettier than the control and low expectations conditions. The sixth and final bean is the hard/ugly website with hard/ugly expectations. It was expected that it would be the lowest given its low actual levels which would were perpetuated with the low expectations. These trends were predicted both pre- and post-use.

To test these first two research hypotheses, qualitative and quantitative data were analysed for evidence of the influence of expectations on usability and visual appeal. Qualitative data, in the form of comments made during the website interaction and post-task interview responses, were examined to see if participants were consciously aware of the influence of the expectations. The researcher went through the participant feedback to distill out what they thought of: the usability, the visual appeal, if the expectation in the task description was noticed, and if they agreed with it. Participants were not asked about the expectation effect directly to avoid raising their awareness of them (possibly altering their answers). Asking if participants were influenced by the expectations may be like asking someone if they were biased. Answering that question conscientiously and truthfully is difficult.

While the qualitative data was used to gain an understanding of what participant sentiment was in each of the conditions, statistical inferences were needed to fully understand if expectations were altering perceptions between conditions. Thus, the SUS and VisAWI-S scales were analyzed statistically to examine the significance of differences between groups. There were two statistical hypotheses, one for each variable. The two statistical hypotheses can be seen in Table 6.1. The statistics for visual appeal and perceived usability were tested separately, but they were not separated for the correlation analysis. Each of the two hypotheses has four sub-hypotheses which were tested for main effects. There were four sub-hypotheses because there were two websites (HuHv and LuLv) and two points of use (pre and post), creating a 2X2 matrix of tests. If a main effect was found, then pairwise comparisons were calculated in order to determine which conditions differed from the others.

The rationale for the statistical tests can be found below, in the Assumptions Testing section. In summary, assumptions testing revealed that the data was not normally distributed and variance was heterogeneous. Thus, non-parametric tests needed to be applied. In addition, the sample size per condition was small (n=10). Therefore, Independent-Samples Kruskal-Wallis Tests (two-tailed to examine differences) were done. If they showed a difference, then pairwise Mann-Whitney tests on specific groups was done to determine which pairs are significantly different, and they can be one-tailed to determine directionality. The test results were automatically generated by SPSS.

Table 6.1. Visual appeal and perceived usability statistical hypotheses and tests used.

Hypotheses	Tests
H_1: Visual appeal differs within each of the two websites (i.e. HuHv and LuLv), pre- and post-use. H_{1a}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe pre-use. H_{1b}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe post-use. H_{1c}: Visual appeal differs between LuLvHe, LuLvLe, and LuLvNe pre-use. H_{1d}: Visual appeal differs between LuLvHe, LuLvLe, and LuLvNe post-use.	Independent-Samples Kruskal-Wallis Test, Kruskal-Wallis multiple comparison tests, i.e. Wilcoxon-Mann-Whitney
H_2: Perceived usability differs within each of the two websites pre- and post-use. H_{2a}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe pre-use. H_{2b}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe post-use. H_{2c}: The perceived usability differs between LuLvHe, LuLvLe, and LuLvNe pre-use. H_{2d}: The perceived usability differs between LuLvHe, LuLvLe, and LuLvNe post-use.	Same as H_1.

As the second research hypothesis states that expectations influence perceived usability, the remaining research hypotheses states that participant performance (in the form of objective usability measures) is also affected by expectations. In particular, participants will perform better (i.e. with ease) when the expectations are set to be higher and will perform worse (i.e. struggle more) when the expectations are set to be lower. This is hypothesised because participants who perceive the website to be either easier or harder to use may reflect their perceptions in how they use the website as a confirmation bias or as a result of the conformance to the expectation. Thus, the last four hypotheses correspond to one of each of the objective usability measures (i.e. the performance measures: clicks, hovers, time, and passes), seen in Table 6.2. Basically, they state that participants would struggle more given the lower expectations and will use the website with greater ease in the higher expectation conditions.

Following this logic, the *third hypothesis* states that the average number of clicks per task per participant, henceforth referred to as ‗clicks' would differ between the expectation conditions within the same website. This could occur because higher expectations would suggest that it would be easier to interact with the website, making fewer clicks than with lower expectations.

The *fourth hypothesis* states that the average number of hovers per participant per task, which is referred to as ‗hovers' from now on, would differ between conditions in a website version. This would happen because, even though the website was the same for the high, low, and no expectation conditions, participants would still struggle more with

the low expectation condition. This may translate into a higher occurrence of hovers (i.e. examining more menu options).

The average time taken to complete a task was the third objective usability measure and will be referred to as ‗time'. Thus, the *fifth hypothesis* was that average task completion times would differ between the groups. This could occur because participants struggling to find the answers would require more time in the low expectation conditions.

The last objective usability variable was success, which was a binary variable in terms of pass/fail. A pass occurred when an answer to a task was correct and found within three minutes. Any form of deviation from that definition (e.g. took longer than three minutes) and the task was considered a fail. The definition of a success in this study varies from the previous preliminary studies because ‗hints' were taken out of this study. Therefore, the *sixth hypothesis* was that there would be a difference between conditions within a website version when it came to the success rates. In other words, participants who were in the low expectation condition would complete fewer tasks than those participants with high expectations. This was hypothesized because under the assumption that participants would internalize their expectations, they would portray explicit symptoms of the internalization, thereby reducing speed and accuracy of completing their information retrieval tasks.

Hints were not given for the main studies. These objective usability measures were chosen in order to get an understanding of the effect of expectations on participants during website use. Each of the four hypotheses has two sub-hypotheses that were tested, one for each website.

Table 6.2. Objective usability statistical hypotheses and tests used.

Hypotheses	Tests
H_3: The average number of clicks per task differs within each of the two websites. H_{3a}: Clicks differ between HuHvHe, HuHvNe, and HuHvLe. H_{3b}: Clicks differ between LuLvHe, LuLvLe, and LuLvNe.	Same as H_1.
H_4: The average number of hovers per task differs within each of the two websites. H_{4a}: Hovers differ between HuHvHe, HuHvNe, and HuHvLe. H_{4a}: Hovers differ between LuLvHe, LuLvLe, and LuLvNe.	Same as H_1.
H_5: The average time to complete each task differs within each of the two websites. H_{5a}: Completion time differs between HuHvHe, HuHvNe, and HuHvLe. H_{5a}: Completion time differs between LuLvHe, LuLvLe, and LuLvNe.	Same as H_1.
H_6: The average number of passed tasks per participant differs within each of the two websites. H_{6a}: Tasks passed differ between HuHvHe, HuHvNe, and HuHvLe. H_{6a}: Tasks passed differ between LuLvHe, LuLvLe, and LuLvNe.	Same as H_1 and Fishers Exact test.

Correlation analysis was also calculated, in search of support for the hypotheses. This work was exploratory, as with the qualitative results, to supplement the statistical results. As previously mentioned, the HCI literature has mixed findings with respect to correlations between visual appeal and usability, with Katz (2010) finding significant correlations between visual appeal and perceived usability before system use, but not after, while others found that users' judgments of a system are mainly guided by the visual appeal before use (Lee & Koubek, 2010), still Ilmberger et al. (2008) found that correlations increased rather than decreased after use, indicating that usability influences aesthetics, and not the other way around as found by Tractinsky et al. (2012).

Therefore, correlations will be examined between visual appeal and usability pre- and post-use in this study as well. In conditions where the expectation is congruent with the website level (i.e. HuHvHe and LuLvLe), the prediction was that correlations would be positive and strong between visual appeal and perceived usability pre- and post-use. This should occur because both values are congruent with the objective levels and viewing and using the website should just confirm the expectation. In the case where expectations are incongruent with website usability and visual appeal levels (i.e. HuHvLe and LuLvHe), the prediction is that the correlations will be positive but weaker between visual appeal and perceived usability pre- and post-use (H_1 and H_2). This should occur because both values are set to the exact opposite of the objective levels but according to cognitive dissonance, they will tend to agree with the expectation to reduce the dissonance. Thus, viewing and using the website should confirm the expectation but may waver a little because of the contrast in what they see and what was said (i.e. the initial dissonance).

As previously mentioned, participants commented on their mood (i.e. mainly frustration) in Phase 5. Therefore, mood was measured pre- and post-use of the website but not manipulated in this study. It was measured using the Self-Assessment Manikin (SAM; Bradley & Lang, 1994). This scale was chosen as it is a short questionnaire with only three image-based scales asking participants to circle how they currently feel.

Method

Participants

As previously mentioned in Chapter 4, random university students were used because they are a representative sample of the general population and do not pose a threat to external validity (e.g. Svahnberg, Aurun, & Wohlin, 2008; Druckman & Kam, 2009). Thus, a sample of 60 (39 males, 21 females; 48 aged 18-30 years, 12 aged 31+) Swinburne University student volunteers participated, all with 20/20 or corrected to 20/20 vision, and screened for colour blindness. All participants were technology-savvy regular Internet users. Twenty-eight were born in an English speaking country and 40 spoke it frequently at home. Thirty-five out of the 60 were undergraduate students, 21 masters, and four PhD students. Out of the 60, 47 were studying computer science, three design, 2 each for games development, arts, psychology, and one each engineering,

business, biomedical engineering, and astrophysics and supercomputing. Participants were randomly assigned and individually tested, approximately one hour per session, ten participants per condition.

Apparatus and Location

Participants were tested using a Hewlett Packard desktop computer, running Intel® Core™2 Duo CPU with 3GB of RAM, and a screen resolution of 1290 X 720. Microsoft Excel, SPSS, R and RStudio, and an online calculator for the Fisher's Exact Test (http://quantpsy.org/fisher/fisher.htm) were used to analyze the data statistically and to produce the figures in this chapter. The study took place in the usability laboratory at Swinburne University of Technology, which had the participant and observer in two separate rooms with a one-way mirror between them. The Morae software was used to connect the participant's computer to the observer's computer and to record participant interaction with the website. Participants' audio and video were not recorded.

Materials

All documentation pertaining to Main Study 1 can be found in Appendices B and D. The same informed consent, project information, demographics form, usability and visual appeal scales were used as in the previous studies in this thesis. As mentioned earlier, two versions of the website were used: HuHv and LuLv. Three different task descriptions were prepared, a paragraph long each, setting expectations high for visual appeal and usability, low for both, or neither (a control/neutral paragraph without any expectations). For example, the high visual appeal, high usability task description was (the other task descriptions can be seen in Appendix D3):

> —Welcome to Gold Coast, Australia's greatest travel destination! Your boss was delighted with your work and decided to promote you to senior manager of the company in Gold Coast. You are bound to love it there and the job's pay is great. Before you start packing and head off, you're going to check the city's city council website out, to get some information which will help you get ready for the move. Recent surveys have found that the website is as beautiful as the gorgeous city. People are finding it incredibly easy to use, and they all recommended it to their family and friends. The developers created a professional masterpiece and the website won an award for best city council website in Australia in 2013."

Similarly to what was done in Phases 4 and 5 from the previous chapter, the SUS scale was used as a subjective usability measure pre- and post-use, and objective usability in the form of performance measures were acquired per task. These were: the number of clicks, the number of hovers, task completion time in seconds with a

maximum of 180sec (i.e. three minutes), and success (pass/fail; pass if the answer is correct and within time limit). For more information and the definitions of these measures, please see Chapter 5, Phases 3 and 4. Participants were not given time or the opportunity to surf or look idly at the website. The higher the number of clicks, hovers, and time per task, the more participants had to explore the website in order to find the answers to the tasks, suggesting that higher values for these variables indicate lower usability levels. Inversely for success, if the success rates were higher or closer to 1, then participants were more likely to finish a task correctly and higher values for the average number of passed tasks indicates a higher usability level for the website.

The SAM mood scale was also given to participants. Two five-slide PowerPoint presentations (one for each of the two versions of the website) were prepared with the same format as Figure 5.1, with the first slide being the instructions, the second a _+' focus slide, and the last three screenshots of the interface. Ten out of the 15 information retrieval tasks from Preliminary Studies 4 and 5 were given to participants, in random order. For more information about the tasks, please see Appendix D1 as well as the Phase 3 Discussion in Chapter 5. An example of a task is: –How many beaches are located in the Gold Coast?" All tasks were on the same page.

Each participant was asked for their feedback at the end of their session, in addition to any comments that they may have had during website use. Specifically, they were asked four questions: what they thought of (1) the usability, (2) the visual appeal, (3) if they remembered the task description, and (4) if the task description gave them an expectation.

Design

This study adopted a two-by-three (two websites, three expectations) between-group design. The website was shown in two parts, the first was the slideshow needed for pre-use data, and the second was the functioning website needed for post-use data. Each participant was scheduled and did the experiment separately.

Procedure

Each participant did the experiment separately (separate days and times), in one-on-one sessions with the researchers. Once the given participant was briefed on the purpose and procedure, s/he signed the consent forms and filled in the pre-use mood questionnaire. Then, the participant was given written expectations in the form of task descriptions according to the condition that they were randomly assigned to. At this point, they were ready to start the first part of the study and were given instructions accordingly. The participant was told that the instructions would be repeated on the computer screen in front of them and that they would be able to read them at their own pace. They then viewed the slideshow of the website and rated it on usability and visual appeal. When the ratings were complete, the researcher turned the slideshow off, opened the website, and gave instructions for the second part of the study. All

participants were instructed to start each task from the homepage, told that the search bar would not work, to avoid using other websites or prior knowledge to answer the tasks, and asked to persist with a task until they got an answer or were told to move to the next one. The researcher then left the participant in the observation room and went to the control room. As soon as the researcher and participant were both ready in their separate rooms, the second part of the usability test began. The participant and researcher were connected via a phone on speaker (hands free). In the second part, participants attempted to complete ten tasks using the website. At the end of the last question (i.e. post-use), participants filled out the visual appeal, usability, and mood questionnaires again. The researcher then returned to the participant room and asked the participant for feedback on usability, visual appeal, and if they recalled the task description before giving them the gift card and thanking them for their help. Notes were taken regarding comments made and body language during participant responses, since no participant audio and video was recorded.

Data Analysis

Normality and homogeneity of variance were tested. Then, the averages were calculated per condition pre- and post-use for visual appeal, perceived usability, and mood. The average results for the objective usability measures were calculated across tasks, per participant. Non-parametric tests were applied, chiefly Kruskal-Wallis for main effects, Fisher‘s Exact Test and Wilcoxon Mann-Whitney for pairwise comparisons. Spearman‘s Correlation Coefficients were used to examine other relationships that may exist between variables.

Results

The Results section is structured as follows. The first section discusses the qualitative findings, in the form of participant feedback during and after their test sessions. The second section describes the statistical assumptions testing which was necessary in order to determine which statistical tests to further apply to the data. The statistical hypotheses testing results then follow. The results section concludes with discussing the correlations discovered between usability and visual appeal and is followed by a discussion section.

Participant Feedback

The researcher then went through the participant feedback to distill out the following: what they thought of (1) the usability, (2) the visual appeal, (3) if the expectation in the task description was noticed (4) if they agreed with the task description, and (5) any last remarks.

Perceived Ease of Use. When asked about the usability level, 11 of the 30 participants in the HuHv website conditions (two in the He, four in Le, and five in Ne) said that it was easy to use. They stated that ~~it~~'s easy to access general information", ~~I~~ found most of the answers to the questions", ~~it~~ did what it needed to do", and ~~it~~ was user-friendly." Interestingly, the smallest number of participants that stated it was easy to use came from the group that was given the expectation that it would be easy to use. Perhaps their expectations were confirmed and did not see the need to comment on it as it was not a _new' finding.

No one in the LuLvHe and LuLvLe conditions thought it was easy to use, and only two in LuLvNe thought that only the ~~h~~omepage was good, the rest not so much." Therefore, while the participants in the HuHv conditions were not all convinced that it was an easy website to use, more of them did think so than the participants in the LuLv conditions, which was expected.

More participants expressed that there were problems with the website's usability. The website was deemed ~~h~~ard to use" by 14 of the 30 in the HuHv conditions (seven in He, five in Le, and two in Ne), and 24 out of 30 in the LuLv conditions (eight in each of the three conditions). Again, in the HuHv website, the majority of participants in the He condition complained about usability. This suggests that the high expectations may have been too high, and participants thus believed that it would be easier than it was. However, the main complaints were that they were not allowed to use the search bar since the website was scrapped from the internet and that automatically deactivated the search bar. Manually going through the website using the menus at the top and right-hand side of the screen were out-dated search methods according to the participants across all conditions. One participant even said that he felt like an old man searching through that website because no one looked for information without the search bar anymore. In fact, participants across conditions commented that the website needed to be further developed because this feature was unavailable. The menu on the right-hand side of the page also confused many participants, as they were not used to looking in this location, where ads are normally placed on other websites. However, the difference between the two websites (HuHv and LuLv) was that the website was only hard to use in the beginning for the HuHv website, while it was hard to use the entire time for the LuLv website. The comments from participants in the LuLv website conditions were far more negative, with the most negative coming from participants in the LuLvLe condition. These participants said that the website was unorganized because the menus and links were scattered throughout the website (menus were on the top, right, and bottom). This meant that participants had too many places to go to and that they had to jump back and forth between menus and locations in the website. For example, one participant noted that water rates could not be found from the water link, where all the other water information was. This sentiment was shared with three in LuLvHe, six in LuLvLe and four in LuLvNe condition, where each condition had ten people.

Seven participants people in HuHv also mentioned that there was too much text on the website. Specifically, three participants in HuHvHe and four in HuHvLe mentioned there being too much text versus two from LuLvLe and one from LuLvNe.

One participant in HuHv mentioned that it was ―bad for elders" because there was too much information and too much to scroll through. This may have occurred because the problems were more severe in the LuLv website versions and participants could not get to the text in order to read or assess it. In addition, some participants encountered terminology issues, such as lack of domain knowledge leading to not understanding some terms (e.g. what a ‗division' was). Another terminology issue that arose came from participants who were non-native English speakers (e.g. not knowing what a ‗vaccination' was and if it differed from a ‗vacancy'). Luckily, terminology issues were only reported by one participant from HuHvHe, two in HuHvLe, three from both HuHvNe and LuLvNe.

Summary of Perceived Usability Comments. Positive usability comments were given by two in the He, four in Le, and five in Ne from the HuHv website conditions, whereas only two participants in the control condition of the LuLv website found it easy to use. Therefore, while the participants in the HuHv conditions were not all convinced that it was an easy website to use, more of them did think so than the participants in the LuLv conditions, which was expected. This also suggests that there was some influence of the textual expectation, given that the neutral conditions were the most positive. In fact, in the HuHv website, the He condition was the least surprised about the usability level. In the HuHvLe condition, four out of ten thought that it indeed was better than they thought it would be.

Negative feedback was also received for the website's usability. The LuLv (eight from Le, eight in He, and eight from Ne) website conditions had more instances of people giving up or not being able to find the answer (HuHv had seven in He, five in Le, and two in Ne). This supports the previous findings that the bad website was indeed harder to use. Moreover, it further provides evidence, in the HuHv website, that expectations may indeed be influencing users. For example, the He condition has the highest number of users complain about the usability suggesting that they actually had higher expectations of the website. The users in the Le condition may have also struggled more with the website, or perceived it to be just as bad as they expected. The control condition had the least complaints, as they were neutral going in to the experiment.

Visual Appeal. When asked about the visual appeal level, over half of all of the HuHv website participants (five participants from He, six from Le, and seven from Ne) liked the overall appearance of the website while only four from LuLvNe said that it was pretty. This was expected in the HuHv conditions, since their opinions correspond to the website's actual visual appeal level. Four people from the control condition for the LuLv website said that the layout was ―alright" and the negative images were a novel idea. The opposite effect occurred for those who did not like the appearance of the website. Particularly, eight participants in the HuHv conditions (two in each of He and Le, four in Ne), and 19 participants from the LuLv conditions (seven in each of He and Ne, and five in Le) said that the website was ―really ugly". One participant from

LuLvHe said that it would be rated as ~~–2~~/10" and that it was the ugliest website he had ever seen in his life. These reactions were expected, given that they correspond to the website's actual visual appeal levels and allow for individual differences.

Only one participant in the HuHvLe condition stated that they did not like the colours because they appeared a bit washed out. However, 17 participants in the LuLv website conditions (eight from He, four from Le, and five Ne), were far more vocal about the colours used. For example, one person from Le said that the website was ~~–m~~ade by someone colour-blind." Participants in the Le condition stated that the purple-green combination was terrible and that ~~–r~~ed buttons are a disaster." Again, this reaction was expected since the website was created to be lower in visual appeal. However, the most negative (in quality) comments came from the Le group.

Seven participants in the HuHv website conditions (three in He, three from Le, and one from Ne) stated that they did not like the images as they appeared out-dated. Participants stated that city views and landscape images were missing and that the site would be better if they were there. This reaction was mirrored in the LuLv conditions, where five participants (two from He and three from Ne) stated that the negative images were ugly, distracted and disturbed the participants, and were not eye-catching. The relatively small number of people to comment on the images in the LuLv conditions may have occurred because overall, the negative images had purple and green in them, which would have gone with the background. However, images of people, such as councilors, would have been unidentifiable. In addition, the website was overall ugly and there were too many elements that bothered the participants so not many would be able to identify the images as a source of the low visual appeal.

Summary of visual appeal feedback results. In general, the results pertaining to the HuHv and LuLv conditions showed that there was more positive feedback from the HuHv website conditions. This result was expected, given that it reflects the websites' actual usability and visual appeal levels. Spreading the feedback across the conditions shows that most of the positive feedback came from the two control conditions and that the meanest comments came from the Le groups.

Feedback on Expectations. Only three participants in HuHv (two in He and one in Le) and five in LuLvLe said that they believed the task description expectation from the start of the study. However, this was a question that required subjective recall and whether or not they believed it would be seen in the results. Participants were then asked if they agreed with the initial task description's portrayal of the website. Two participants from HuHvHe, three from HuHvLe, and seven from LuLvLe said that they agreed with the descriptions, while one from HuHvHe, four from HuHvLe, and six from LuLvHe said that they disagreed. These results correspond with the condition in which they were in, as the expectation was incongruent with the website's usability and visual appeal levels.

Although, 24 participants stated that they had no expectations (three in HuHvLe, six in HuHvNe, three in LuLvHe, six from LuLvLe, and six from LuLvNe). Thus,

participants either really did not have expectations or were not consciously aware of them. While this makes sense in the control conditions, where there was no expectation in the task description, it should not have happened in the experimental conditions. Some participants (one from LuLvHe and four from LuLvLe) stated that they did not read or had forgotten what was written in the task description. Participants went on to state that they were suspicious of the expectations (two in LuLvHe), or had their own/different expectations (two in each of HuHvNe, LuLvLe, and LuLvNe). For example, one participant said that they had a given set of expectations of any modern website while another participant said that they thought all government websites were poorly made so he did not expect any better for the city council website of Gold Coast. These could present confound variables, and will be further discussed in the discussion section.

Beanplot Results

Beanplots were created to gain a general understanding of the VisAWI-S and SUS scale data, with pre- and post-use visual appeal in Figures 6.2 and 6.3, and pre- and post-use perceived usability in Figures 6.4 and 6.5, respectively. In all of the figures, the grey beanplots on the top are the HuHv website measures and the purple ones on the bottom are the LuLv measures. The first columns on the left represents the Ne conditions, the middle columns are the He conditions, and the ones on the far right are the Le conditions. The red lines indicate each condition‘s mean. The dotted lines indicate the website‘s overall mean, across the three conditions (Ne, He, and Le). The influence of expectations on visual appeal and usability cannot be readily compared using the graphs since the SUS is a five-point scale while visual appeal is a seven-point scale.

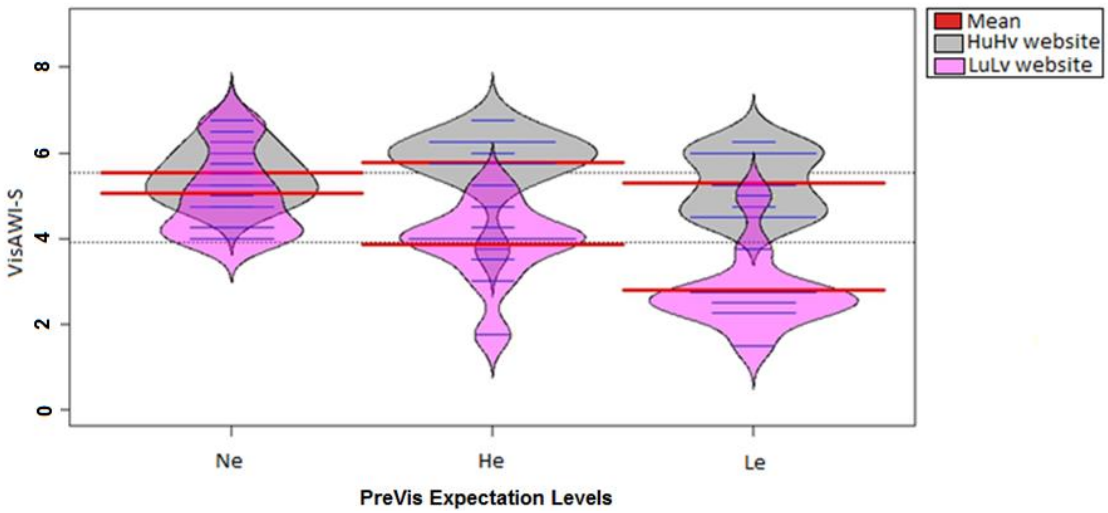


Figure 6.2. Beanplot of the pre-use visual appeal results.

In Figure 6.2, in the HuHv website, there are very slight differences between the three expectation conditions, evidenced by the proximity of the red lines. The second

column which corresponds to the He condition is the highest, followed closely by Ne and then Le. The LuLv website conditions depict a slightly different story. In LuLv, Ne was considered to be the prettiest, with He in the middle, and Le the lowest. Specifically, participants in the Le condition perceived it to be on average one point (on a seven point scale) uglier than participants in the He condition, and two points uglier than in the Ne condition. Also in LuLv, the distribution of data points in the He condition is similar to Le, with the highest and lowest scores being the same which suggests that the visual appeal measures may not be significantly different. However, the statistical results discuss this in the next section below.

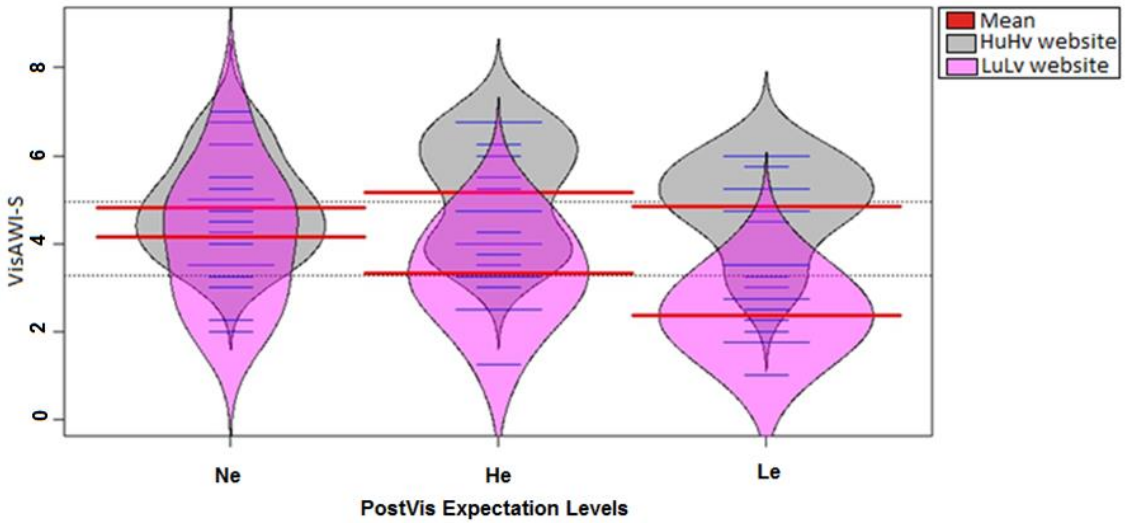


Figure 6.3. Beanplot of the post-use visual appeal results.

In Figure 6.3, across the HuHv website conditions, there are some visually discernable differences pre- and post-use; altogether dropping roughly a point in visual appeal after use. Participants in He appear to be polarized, either rating it slightly higher or slightly lower than the group‘s mean. However, the mean is the highest amongst the expectation conditions. The difference between ratings in the post-use visual appeal of LuLv website shows a greater disparity between conditions. The Ne condition is rated highest, followed by He and then Le was rated as the worst, by far. In addition, LuLvNe has a larger spread suggesting that some participants thought it was prettier and some uglier than pre-use, but the majority was still centered around the mean. Altogether, some evidence exists to support the first hypothesis that visual appeal differs between conditions in both HuHv and LuLv websites, pre- and post-use, with larger differences evident in the LuLv conditions. Statistical analysis in the next section determines the significance of these differences.

In Figure 6.4, the difference between the pre-use perceived usability is again visible across all conditions for pre-use perceived usability. The control conditions seem to be rated as the easiest to use, followed by He, and then by Le which is rated as the hardest, across both websites. However, the means to these groups are within a one-point range which can be considered a minor difference. Yet, given the small sample

size per condition (n=10), the visible differences in the graph suggests that expectations did impact pre-use perceived usability as well.

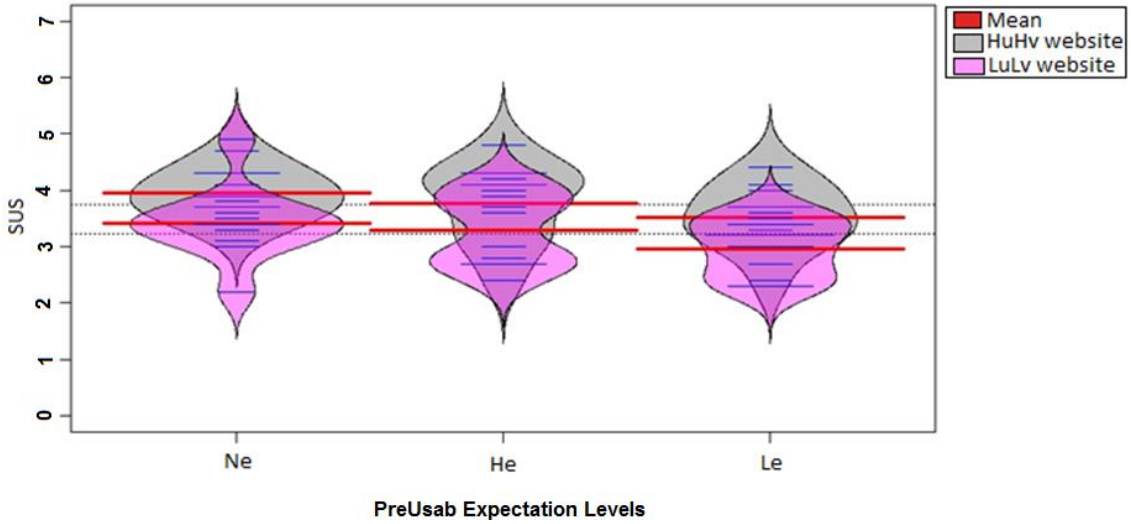


Figure 6.4. Beanplot of the pre-use perceived usability results.

Post-use perceived usability, in Figure 6.5, in both HuHv and LuLv, He was rated as the easiest but by a fraction. The Le conditions were again rated as worst. The differences are minute between the conditions in both websites, but are slightly more prominent in the LuLv website. Again, the difference between the red bars, representing the means, need to be verified for significance using statistical analysis, in the next section.

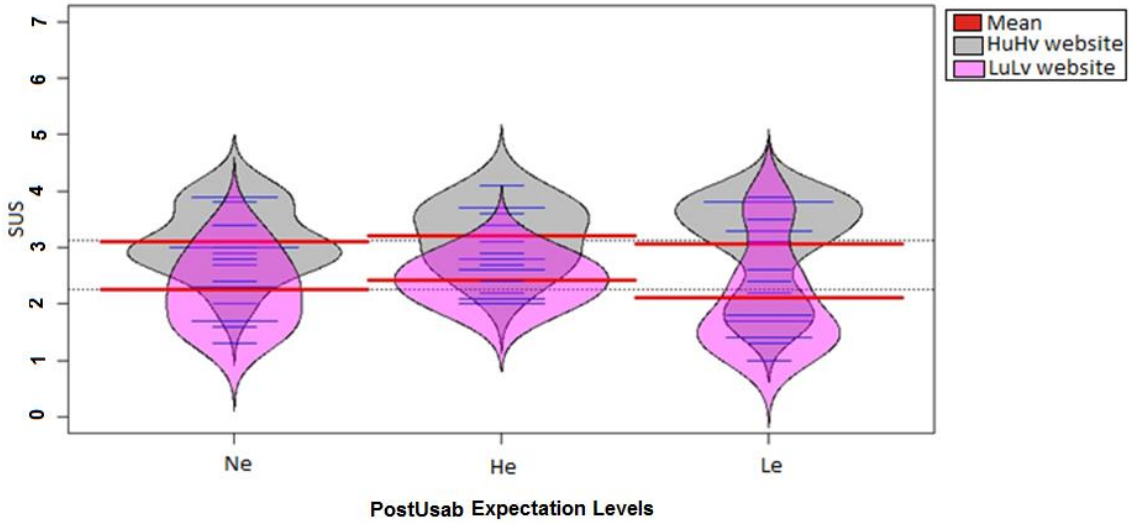


Figure 6.5. Beanplot of the post-use perceived usability results.

Beanplot result summary. He is rated highest and Le lowest amongst the pre-use visual appeal HuHv conditions. Across the LuLv conditions, the control condition is rated as prettiest, with Le rated the lowest. The same trend occurs with post-use visual appeal. Pre-use perceived usability also rated with the control condition being rated as

the highest, followed by He and Le, in both websites. Post-use, the trend in perceived usability changes slightly, with He being rated as the easiest and Le rated as the hardest.

The evidence in the beanplots furthers the finding in the participant feedback and supports the first two hypotheses: visual appeal and perceived usability appear to differ between conditions in the two websites, both pre- and post-use. This is evident especially in the LuLv website conditions, where He is rated as highest and Le is rated as lowest. Thus, participants seem to be internalizing the textual expectations and this seems to be influencing their judgment of visual appeal and usability.

The problem is that beanplots are used as a mechanism to visualise complex datasets and provide visual indications of differences between datasets. Since they provide indications, further analysis of the data is needed in order to have conclusive results. To statistically verify the significance of the findings in the beanplots, the next two sections deal with statistical assumptions and hypothesis testing.

Assumptions Testing

Statistical assumptions for normality and homogeneity were tested in order to determine which statistical tests were appropriate to apply to the visual appeal, subjective and objective usability data. The assumptions for normality and homogeneity of variance were checked for each variable across all conditions and were not unilaterally met. All of these values were computed by SPSS from the data.

Shapiro-Wilk tests showed that post-use usability at HuHvLe ($p=.027$), pre-use visual appeal for HuHvHe ($p=.025$), clicks at LuLvNe ($p=.021$), and passes HuHvHe ($p=.030$) were not normally distributed. In addition, while the skewness of pre-use usability for LuLvLe was 0.610, (SE=0.687), the kurtosis measure of 2.904 (SE=1.334) revealed that it may not be normally distributed. Pre-use visual appeal at HuHvHe had a skewness of -1.809 (SE=0.687) and a kurtosis of 4.312 (SE=1.334), and at LuLvLe had a skewness of 1.418 (SE=0.687) and a kurtosis of 2.794 (SE=1.334). Performance measures also had some non-normal values. Specifically, the average number of clicks may not be normal at LuLvNe because skewness was 1.634 (SE=0.687) but kurtosis was 2.351 (SE=1.334). Time also appears to violate the normality assumption with a skewness of 1.604 (SE=0.687) and a kurtosis of 3.510 (SE=1.334) at LuLvHe. Passes had a skewness of 0.569 (SE=0.687) which may be normal however with a kurtosis of 2.922 (SE=1.334) at HuHvHe, it violated the normality assumption as well. The rest of the factors appeared to be normal. Therefore, the data was largely not normally distributed. Since the data was not normally distributed, non-parametric data needed to be used.

The non-parametric Levene‘s test revealed that the homogeneity of variance assumption was violated (i.e. heterogeneous) for pre-use usability ($p<.05$). The test of homogeneity of variances from SPSS also found that clicks ($p <.01$) and hovers ($p<.05$) violated the homogeneity assumption. Therefore, the homogeneity of variance assumption was not unilaterally met, either.

Given that assumptions for constant variance and normality were not met, that some variables were binary (passes), some were discrete (clicks and hovers) and others continuous (time), and that sample size per condition was relatively small (n=10), ANOVAs could not be applied to the data. As mentioned in Tables 6.1 and 6.2, Kruskal-Wallis and Fishers Exact tests were applied where appropriate.

Statistical Hypothesis Testing

Kruskal-Wallis tests were done on the pre- and post-mood data. No significant findings were obtained. Thus, no main effects were found for mood, pre- or post- test. This means that mood was not impacted nor did it impact any of the other measured variables in this study. Due to the lack of significant results and since mood was not an integral part of this thesis, non-parametric analysis for differences in measures of centrality will not be displayed for mood. The SAM scale is discussed in the construct validity section of the discussion, below.

Table 6.3. Visual appeal and perceived usability statistical hypotheses and results.

Hypotheses	Results
H_1: Visual appeal differs within each of the two websites (i.e. HuHv and LuLv), pre- and post-use.	
H_{1a}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe pre-use.	H_{1a}: Null not rejected.
H_{1b}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe post-use.	H_{1b}: Null not rejected.
H_{1c}: Visual appeal differs between LuLvHe, LuLvLe, and LuLvNe pre-use.	H_{1c}: Null rejected ($p<.01$).
H_{1d}: Visual appeal differs between LuLvHe, LuLvLe, and LuLvNe post-use.	H_{1d}: Null rejected ($p<.05$).
H_2: Perceived usability differs within each of the two websites pre- and post-use.	
H_{2a}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe pre-use.	H_{2a}: Null not rejected.
H_{2b}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe post-use.	H_{2b}: Null not rejected.
H_{2c}: The perceived usability differs between LuLvHe, LuLvLe, and LuLvNe pre-use.	H_{2c}: Null not rejected.
H_{2d}: The perceived usability differs between LuLvHe, LuLvLe, and LuLvNe post-use.	H_{2d}: Null not rejected.

For the remainder of the results, the main effects testing for visual appeal and perceived usability (i.e. H_1 and H_2) can be seen in Table 6.3 below (full results are in Appendix D4). Out of the eight statistical sub-hypotheses tested for visual appeal and perceived usability, two were significant. Both pre- ($p<.01$) and post-use ($p<.05$) visual appeal ratings were significantly different within the LuLv website (i.e. main effects

were found in the LuLv website conditions). In other words, visual appeal differed between LuLvHe, LuLvLe, and LuLvNe both pre- and post-use. Paired comparisons showed that LuLvLe and LuLvNe differed in visual appeal both pre- ($p<.001$) and post-use ($p<.05$). Therefore, there is statistical evidence to support the first hypothesis: visual appeal does appear to be influenced by textual expectations. Perceived usability may also be influenced by textual expectations, but a larger sample size may have to be used in order to exhibit significant results.

One out of the eight sub-hypotheses for objective usability measures was significant, seen in Table 6.4. Full results are in Appendix D4. The average number of clicks per task significantly differed ($p<.05$) within the HuHv website (i.e. one main effect found with clicks). This means that there was a difference between the HuHvHe, HuHvNe, and HuHvLe conditions with respect to the number of clicks participants made per task. Pairwise comparisons showed that the number of clicks were different ($p<.05$) between the HuHvHe ($x=3.98$) and HuHvLe ($x= 3.07$) conditions. This suggests that participants interacted more so with the website that had the positive expectation than with the low. This is contrary to what was expected, suggesting that while participants had, statistically, the same success rates across conditions in a website, they interacted with the website less when they were told it was going to be hard to use. One possibility for this finding was that the low expectations may have discouraged the participants from exploring the website. There is insufficient statistical evidence to conclude that the third research hypothesis is true.

Table 6.4. Objective usability statistical hypotheses and their results.

Hypotheses	Results
H_3: The average number of clicks per task differs within each of the two website conditions	H_{3a}: Null rejected ($p<.05$). H_{3b}: Null not rejected.
H_{3a}: Clicks differ between HuHvHe, HuHvNe, and HuHvLe.	
H_{3b}: Clicks differ between LuLvHe, LuLvLe, and LuLvNe.	
H_4: The average number of hovers per task differs within each of the two website conditions.	H_{4a}: Null not rejected.
H_{4a}: Hovers differ between HuHvHe, HuHvNe, and HuHvLe.	H_{4b}: Null not rejected.
H_{4b}: Hovers differ between LuLvHe, LuLvLe, and LuLvNe.	
H_5: The average time to complete each task differs within each of the two website conditions.	
H_{5a}: Time differs between HuHvHe, HuHvNe, and HuHvLe.	H_{5a}: Null not rejected.
H_{5b}: Ttime differs between LuLvHe, LuLvLe, and LuLvNe.	H_{5b}: Null not rejected.
H_6: The average number of passed tasks per participant differs within each of the two website conditions.	
H_{6a}: Passes differ between HuHvHe, HuHvNe, and HuHvLe.	H_{6a}: Null not rejected.
H_{6b}: Passes differ between LuLvHe, LuLvLe, and LuLvNe.	
	H_{6b}: Null not rejected.

As mentioned earlier, the third research hypothesis stated that participant performance is affected by expectations, in that participants would perform better when the expectations are high and will perform worse when the textual expectations are low. Partial evidence exists to support this since one of eight main effects were found, where the number of clicks differed between the HuHvHe and HuHvLe conditions.

General Correlations

Spearman correlations were examined between the variables, without taking the conditions into account. Full test results can be found in Appendix D4. The majority of variables were significantly correlated. If there were no conditions, then visual appeal and perceived usability pre-use were positively and strongly correlated ($r=.657$, $p<.001$). Post-use, the correlation between visual appeal and perceived usability was slightly weaker but still positive and significant ($r=.585$, $p<.001$).

Correlations per Website

In the HuHv website conditions, there were nine significant correlations. Pre-use perceived usability was correlated with post-use perceived usability ($r=.535$, $p<.01$), pre- ($r=.498$, $p<.01$) and post-use visual appeal ($r=.484$, $p<.05$), the average number of hovers ($r=-.449$, $p<.05$), and time ($r=-.397$, $p<.05$). Post-use perceived usability was correlated with post-use visual appeal ($r=.563$, $p<.01$). Pre-use visual appeal was correlated with post-use visual appeal ($r=.405$, $p<.05$). Post-use usability was also correlated with post-use mood ($r=.387$, $p<.05$). Time and passes were almost correlated, with a p-value of .055.

In the LuLv website conditions, there were 11 significant correlations. Pre- and post-use perceived usability were correlated ($r=.485$, $p<.01$). Pre-use perceived usability was also correlated with pre- ($r=.615$, $p<.001$) and post-use ($r=.498$, $p<.01$) visual appeal. Post-use perceived usability was correlated with post-use visual appeal ($r=.411$, $p>.05$) and hovers with ($r=-.437$, $p<.05$). Pre- and post-use visual appeal were correlated ($r=.634$, $p<.001$). Post-use visual appeal was correlated with post-use mood ratings ($r=.452$, $p<.05$). Hovers were further correlated with time ($r=.377$, $p<.05$) and passes ($r=.423$, $p<.05$). Time was correlated with passes ($r=-.729$, $p<.01$) as well. In addition, pre- and post-use mood ratings were ($r=.364$, $p<.05$) correlated.

Correlations in the Control Conditions

First, we examine the Spearman correlations between the two control conditions (HuHvNe and LuLvNe), in Table 6.5. **In the HuHvNe condition, both pre- and post-use usability ($r=0.780$, $p<0.01$) and pre- and post-use visual appeal ($r=0.635$, $p<0.05$) were highly and positively correlated.**

Perceived usability and visual appeal measures were not correlated with each other in this condition. This would suggest that while both values were previously

measured as high in Chapter 5, they were not perceived to be equally high in the HuHvNe condition.

Table 6.5. Correlations between visual appeal and perceived usability for HuHvNe and LuLvNe, respectively.

HuHvNe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.780**	.415	.422
PostUsab	-	.009	.334
PreVis		-	.635*

LuLvNe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.202	.716*	.924**
PostUsab	-	-.193	.280
PreVis		-	.665*

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

In the LuLvNe condition, **pre-use usability was highly and significantly correlated** both with pre- (r=.716, p<.05) and post-use (r=.924, p<.01) visual appeal. Pre- and post-use visual appeal were also highly and significantly correlated with each other (r=.665, p<.05).

No correlations were found between pre- and post-use perceived usability, suggesting that use of the website changed their perceptions of it but without pattern.

Correlations when Expectations and Website levels are Congruent

All Spearman correlations between visual appeal and perceived usability within conditions where expectation levels were congruent with the actual website levels can be seen in Table 6.6.

In the HuHvHe condition, pre- and post-use perceived usability were highly and positively correlated (r=.686, p<.01). Pre-use perceived usability was also correlated highly and positively with both pre- (r=.658, p< .05) and post-use visual appeal (r=.782, p< .01). In addition, post-use perceived usability was correlated with post-use visual appeal (r=.715, p<.05). In the LuLvLe condition, only pre- and post-use perceived usability (r=.797, p<.01) was highly and positively correlated.

Table 6.6. Correlations between visual appeal and perceived usability for HuHvHe and LuLvLe, respectively.

HuHvHe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.686*	.658*	.782**
PostUsab	-	.119	.715*
PreVis		-	.250

LuLvLe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.797**	.187	.204
PostUsab	-	.208	.585
PreVis		-	-.053

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

Therefore, in the conditions where the expectation levels of visual appeal and usability were congruent with the website's visual appeal and usability levels, the hypothesis was that correlations would be positive and strong between visual appeal and perceived usability pre- and post-use. **This was indeed the case in the HuHvHe website.** However, when the usability and visual appeal were worse in the LuLvLe condition, **it seems that participants attributed the poor website to usability and did not agree all on visual appeal**. This was evident in the mixed participant feedback as well.

Correlations when Expectations and Website levels are Incongruent

All Spearman correlations between visual appeal and perceived usability in conditions where expectations of these were incongruent with the actual website levels can be seen in Table 6.7.

Nothing was correlated in the HuHvLe and LuLvLe. Therefore, in the case where expectations are incongruent with website usability and visual appeal levels, the research hypothesis was that correlations would be positive but weaker between visual appeal and perceived usability pre- and post-use.

With no significant correlations, this hypothesis could not be supported.

Table 6.7. Correlations between visual appeal and perceived usability for HuHvLe and LuLvHe, respectively.

HuHvLe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.163	.440	.093
PostUsab	-	-.533	.555
PreVis		-	.053

LuLvHe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.454	.577	.221
PostUsab	-	.470	.164
PreVis		-	.470

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

Correlations summary. Looking at the general correlations (n=60) across all the data (Appendix D4), **usability and visual appeal were strongly and positively correlated with each other pre- and post-use**. In other words, the two concepts were perceived similarly and were in some way related. When correlations were examined within a website (n=30), **usability and visual appeal were correlated pre- and post-use, with each other and themselves,** in both the HuHv and LuLv websites. Thus, participants did not drastically change their minds from the beginning of the study to the end on what they thought of visual appeal and usability, and once again, visual appeal and usability were perceived as similar. This makes sense since the usability and visual appeal levels were manipulated to be the same (either high or low) in both website versions.

When correlations were examined per condition (n=10), then a slightly different story emerged. The HuHvHe condition had the most correlations, **having usability and**

visual appeal correlated pre- and post-use, along with pre- and post-use usability correlated. The LuLvLe condition only had one significant correlation, where pre- and post-use usability were correlated. The HuHv control condition (Ne) showed no relationship between visual appeal and usability, only pre- and post-use for visual appeal, and pre- and post-use usability (i.e. each post-use rating was related to the pre-use but not with the other variable). **The LuLvNe condition had significant correlations between pre-use visual appeal and usability, and with pre- and post-use visual appeal.** Interestingly, both the HuHvHe and LuLvNe conditions showed that pre-use usability was correlated with post-use visual appeal. This would suggest that the impression pre-use usability gave participants managed to influence their post-use ratings of visual appeal, but no causality can be established with correlations. **No correlations were found in conditions where the expectation was incongruent with the website's actual visual appeal and usability levels.** However, the lack of significant results here could have been due to the small sample size.

While causality cannot be inferred from correlations, there does seem to be some evidence to support the idea that expectations do impact visual appeal and usability. Mainly, the control conditions seem to be behaving the same way as many studies in the literature. The conditions in which expectations and website visual appeal and usability levels are congruent also obtain similar results to the literature. However, when the expectations and website levels are incongruent (i.e. when there is dissonance), the correlations disappear. This may have occurred if the expectations were internalized and acted upon differently. The lack of significant findings is further examined in the next section.

Discussion

A summary of the results is presented first. This is followed by a limitations and future research section. The conclusion section will be presented last, to end Main Study 1. Main Study 2 follows immediately after the conclusion.

Results Summary

The participant feedback varied, as most opinions are. Apart from the control conditions being perceived as better (it received the most positive feedback for both usability and visual appeal) than the rest, the results pertaining to the HuHv and LuLv conditions were expected, given that they reflect the websites' actual usability and visual appeal levels. However, the most negative/aggressive comments came from the low expectation conditions.

The beanplots showed variations in the means across usability and visual appeal, both pre- and post-use. The LuLv conditions showed the greatest differences between the experimental conditions: the control condition was rated highest pre- and post-use visual appeal and pre-use perceived usability, with Le rated the lowest. Post-use, the trend in perceived usability changed slightly, with He rated as easiest and Le rated as

hardest. However, the differences between the means were marginal. Therefore, a small trend did emerge post-use, supporting the first two hypotheses.

Out of the eight hypotheses tested for visual appeal and perceived usability, two were significant. Specifically, LuLvLe pre- and post-use visual appeal ratings were rated as lower than the control pre- and post-use visual appeal ratings, respectively. This would suggest that participants in the LuLvLe condition perceived the website to be uglier than participants in the control group, irrespective of use. This was also found in the beanplots. Furthermore, one out of the eight statistical hypotheses in objective usability measures was significant. The number of clicks per task varied significantly between the HuHvHe (M=3.98) and HuHvLe (M=3.07) conditions. This suggests that participants used the website less when they were told that it was hard to use, seemingly uninterested or slightly deterred from using it. Therefore, evidence suggests that expectations have an impact on both the perceptions and actions of people using the websites.

Overall, correlations between visual appeal and perceived usability were significant and positive pre- and post-use, agreeing with Tractinsky et al.'s (2012) results. This was the case in the HuHvHe condition where the expectation was congruent with the website level. However, when the usability and visual appeal were worse in the LuLvLe condition, participants attributed the poor website to usability and did not agree on visual appeal. In the case where expectations are incongruent with website usability and visual appeal levels, results were insignificant. The lack of significance across many of the statistical tests suggests that in addition to there being a low sample size (ten per condition), there may have been an issue with the implementation of expectations, explained in the next section. As mentioned earlier, no main effects were found for mood, and mood will therefore not examined anymore in upcoming studies, but will be mentioned in the threats to construct validity section below. Future work should examine mood as a factor in the relationship, however it was out of scope for this thesis.

Thus, there is some support for the hypotheses, but the bigger picture is more complex. For example, while some people were influenced positively by positive expectations, others became more critical. In any case, written expectations influenced the participants. These results further the knowledge relating to the factors that affect the acceptance of website usage in particular with consideration of visual appeal and usability. To gain a deeper understanding of the effect expectations on the visual appeal and usability, future studies should examine the impact of verbal expectations, using confederates.

Implications for Theory

These results support the cognitive dissonance theory, in that participants agreed with most recent information they have. Particularly, in the Le conditions, low expectations at the beginning of the study influenced participants' perceptions and they rated the website uglier and harder to use than participants in other conditions. In fact,

the low expectation also discouraged participants from interacting with the website, on average clicking less through the interface per task. The lack of statistically significant results for the HuHv website suggests that there might have been a problem with the implementation of expectations and Main Study 2, below, re-examines this website before any further theoretical implications are made.

Limitations and Future Studies

Threats to construct validity. There was a problem with instrumentation. The scales used to measure visual appeal and usability produced consistent results but the mood scale (SAMS) was ambiguous. Participants did not understand the drawings in the scale, often asking the researcher for advice on what they meant. The arbitrary and highly subjective interpretation of the images in the SAMS scale, in addition to the lack of main effects with mood and other variables, strongly suggests that the scale be removed from future studies in this thesis.

Threats to statistical validity. While the non-parametric statistics applied to the data in this study are appropriate, there is one limiting factor to the analysis. A relatively small sample size (ten participants per condition) may have not been sufficient to capture some significant differences that a larger sample would have shown. However, due to time and monetary constraints, the sample size will not be increased in future studies.

Threats to internal validity. Another limitation in this study was the assumption that expectations can be formed in a matter of seconds, by reading a short task description, in an unfamiliar physical environment (i.e. ecological validity for the formation of expectations). Unfamiliarity of the location and experimenter could have also influenced trustworthiness of the expectation, lowering its value. Moreover, city council websites were chosen because it seemed that participants did not have clear expectations or experience with such a domain. Therefore, controlling for expectations would, in theory, be easier. Although, domain inexperience may have been a limitation because, when asked at the end of the study, the majority of participants expressed that they expected a website with more images and with less information (i.e. a tourism website).

Furthermore, expectations were set by participants reading a task description that outwardly stated what they were intended to believe. This was then followed by a set of instructions for the two parts of this usability test. They may not have believed the set expectation, evidenced by a participant asking ~~"I~~s this true?" after reading the task description. Additionally, given that the instructions particular to the two parts of the study (the slideshow for pre-use measures, and the website for use and post-use measures) were given after the expectations were set, it may be that participants were overloaded with information and forgot about the expectations. It may also have also been the case that they did not read the task descriptions, since Rettig (1991) found that

participants do not always thoroughly read hard-copy or online documents. Another possible limitation could be a conflicting learning style (Felder, 1993) – if participants in this study were predominantly verbal learners, then the textual (i.e. visual) expectation may not have been effective. Without prolonged exposure, text may not have been enough to fully understand the impact of expectations on visual appeal and usability.

It may have been more effective to give the expectation subliminally or at least less overtly, but this will be subject to similar issues as written expectations (e.g. they may not be understood, etc.). For example, implementing expectations using a confederate who would act as a participant just finishing the usability test, and who would either praise or complain about the website may strengthen the implementation. It is also likely that expectations in fact do not influence or do not significantly influence ratings of visual appeal and usability. These issues were addressed in Main Study 2 by the addition of a confederate who verbally reinforced the textual expectations.

Threats to external validity. There were no pretests to indicate the possibility of a reaction or interaction effect during testing. Each participant was given one treatment, so multiple treatment interference was not a concern in this thesis. Participant recruitment and selection was random and participants were only screened for eye-sight and colour-blindness, given that the colours and images in the website are important in order to ascertain the appropriate visual appeal level. Having met the 20/20 vision requirement, participant assignment to conditions was randomly chosen to eliminate the possibility of selection biases.

However, the participants predominantly studied computer science (47/60), with varying degrees of English fluency, and of different cultural backgrounds. These factors were not controlled for and may have skewed or randomized the results as they influence the perception of visual appeal and could have added some unaccounted difficulties in usability. They also may have contributed to the lack of normality and constancy of variance, and consequently the applicability of some statistical tools. Although, HCI studies are predominantly held in the information technology departments of universities and their participant demographics are similar to this study‘s. Also, the participants were randomly selected and randomly assigned to conditions. In addition, university students were used because they do not threaten external validity (e.g. Svahnberg, Aurun, & Wohlin, 2008; Druckman & Kam, 2009). Thus, the results of the condition outcomes in Main Study 1 are comparable and as generalizable as any other study.

Conclusion

Overall, expectations did seem to have an effect on visual appeal and usability in some circumstances. Yet the effect of expectations on visual appeal and usability is smaller than anticipated. To gain a deeper understanding of the effect expectations on

the relationship between visual appeal and usability, the next study will re-examine the HuHv website version, with only the HuHvHe and HuHvLe conditions. In Main Study 2, a confederate acting like a participant verbally reinforced the task description (written implementation of the expectations) by telling participant their ‗opinion' after having completing the study themselves.

Main Study 2 Introduction

Expectations did seem to have an effect on visual appeal and usability in some circumstances, as seen in Main Study 1. However, the effect of expectations on visual appeal and usability was smaller than anticipated. Therefore, the purpose of Main Study 2 was to reinforce expectations by implementing them verbally as well as in text. Main Study 2 only used the HuHv website, with the HuHvHe and HuHvLe conditions to re-examine the effect of expectations on visual appeal and usability, to see if more significant results could be obtained. Expectations were reinforced by a confederate who acted like a participant finishing the study and gave the real participant their ‗opinion' (i.e. verbal expectation) after having completed the study themselves. This opinion was in fact a similar speech to the task description (i.e. written expectation), reinforcing the expectations. The same research hypotheses from Main Study 1 were used here, seen in Table 6.8, but only for the applicable conditions. The control condition from Main Study 1 was added to the analysis here for purposes of comparison.

Table 6.8. All statistical hypotheses and tests.

Hypotheses	Tests
H_1: Visual appeal differs within each of the two website conditions (i.e. HuHv and LuLv), pre- and post-use. H_{1a}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe pre-use. H_{1b}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe post-use.	Independent-Samples Kruskal-Wallis Test, Kruskal-Wallis multiple comparrison tests, i.e. Wilcoxon-Mann-Whitney
H_2: The perceived usability differs within each of the two website conditions pre- and post-use. H_{2a}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe pre-use. H_{2b}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe post-use.	Same as H_1.
H_3: The average number of clicks per task differs between HuHvHe, HuHvNe, and HuHvLe.	Same as H_1.
H_4: The average number of hovers per task differs between HuHvHe, HuHvNe, and HuHvLe.	Same as H_1.
H_5: The average time to complete each task differs between HuHvHe, HuHvNe, and HuHvLe.	Same as H_1.
H_6: The average number of passed tasks differs between HuHvHe, HuHvNe, and HuHvLe.	Same as H_1 and Fishers Exact test.

Method

Participants

A sample of 20 (16 males, 4 females; 16 aged 18-30 years, 4 aged 31+) different participants were recruited in Main Study 2. Participants were Swinburne University student volunteers, all with 20/20 or corrected to 20/20 vision, and screened for colour blindness. Eleven out of the 20 were born in an English speaking country and 16 spoke English at home frequently. All participants were technology-savvy regular Internet users. Thirteen were undergraduate students, seven masters, divided between courses: 16 in computer science, two engineers, one in business, and one was studying physics. All 20 participants used the internet for banking, 18 for study, 17 for entertainment, 15 for shopping, 14 got travel and news, and 10 for social purposes. Thirteen were not very familiar with the purpose of city councils, seven were only somewhat familiar, and the rest were not familiar. Participants were randomly assigned and individually tested, approximately one hour per session, ten participants per condition. In the analysis below, there are 30 participants which is the result of the addition of the control condition data from Main Study 1.

Apparatus, Materials, and Location

All apparatus and materials pertaining to this study are the same as in Main Study 1. Only the mood scale was excluded because no main effects were found with mood in the previous study. The confederate used a standard script found in Appendix D5. The same usability lab used in the previous study was also used here, with the same computer screen and software.

Design and the Confederate

This study adopted a one-by-two (one website, two expectations) between-group design. Expectations were reinforced verbally by a confederate. One confederate was used in this study. She was a native English speaker and a PhD student from Swinburne University of Technology. Her role was to act like a participant finishing the study and then tell her opinion (i.e. verbal expectation) of the website to the real participant. This opinion was in fact a similar speech to the task description (i.e. written expectation), reinforcing the expectations. The confederate speeches are in Appendix D5. Therefore, a confederate was added at the very beginning.

Identically to the design of Main Study 1, the website was shown in two parts, the first was the slideshow from needed for pre-use data, and the second was the functioning website needed for post-use data.

Procedure and Data Analysis

The procedure is identical to Main Study 1, with the exception of the beginning in which the confederate was added. A confederate would be in the experiment room, picking up their things and getting ready to leave as the participant entered the room. The experimenter would ask the confederate if they were all done and the confederate would respond that they were just leaving. The experimenter would thank them and tell the participant to go ahead in and wait a minute while the experimenter left to set up the computers. The confederate then told them the usability and visual appeal expectations in the form of their experience with the website and left. The experimenter came back into the room and then started with the brief and rest of the procedure from Main Study 1. The data was analysed in the same way as it was in Main Study 1.

Results

This section is structured as follows. The first section outlines the participant feedback findings from the re-test of the HuHvHe and HuHvLe conditions with a confederate. This is followed by preliminary results. The assumptions testing and statistical hypotheses testing results are presented next. The results section concludes with the correlations between usability and visual appeal. The discussion section follows.

Participant Feedback

Similarly as in Main Study 1, at the end of the testing sessions, participants were asked four questions: what they thought of (1) the usability, (2) the visual appeal, (3) if they believed the confederate, and (4) if they agreed with the confederate.

Previously, all of the HuHv website conditions had participants who mentioned some positive aspects about the website‘s usability. This was no longer the case, as only the He condition had four participants mention that it was easy to use, consistent, and a well-structured layout overall (i.e. none of the Le participants said anything positive about the usability). Both groups (seven in He and six in Le) had participants mention that the website was hard to use. While there seemed to be ever so slightly more complaints from the He group, the He group‘s three main complaints were relatively minor: that they could not use the search bar, that the right-hand side of the menu took time to get used to, and there were too many menus/options to take in. Out of the seven, four participants from He had some terminology issues, including English as a second language and city councils being an unknown genre so not understanding some of the jargon. However, everyone in the Le group was far more hostile about the website, having the same complaints as the He group but being more vocal about finding the website hard to use. For example, one said that he would fire the developer, while another said during the test, ~~–e~~an I just put _no‘ [*as the answer*] without looking for it in the website? It‘s so bad.”

A similar finding was seen in with visual appeal. Seven participants from He said that the website looked great, had great colours, and that it looked ―easy on the eyes". Only four participants from Le said that it looked good but were unsure of their opinion. For example, one said that the website ―looks modern, I guess." One participant in each of the two conditions said that they thought that the website looked bad, with the participant in Le adding that the colour did not stand out.

When asked if they believed the confederate, three people from He said that they did and four in Le said that they did. As previously mentioned, in Main Study 1, two participants in He and one in Le said that they believed the task description expectation. Thus, the number of people that believed the confederate tripled in the He condition and quadrupled in the Le condition. In Main Study 2, one participant in He mentioned that this time, he did have high expectations when usually he did not have any of city council websites. One participant in Le said that he ―tried not to be biased but subconsciously, I was."

Therefore, it might be the case that people heard the confederate and tried to be neutral about the website usability test, but were in fact influenced by the expectation.
After using the website, one participant in He agreed with the confederate saying that it was a great website, while three agreed with the confederate's description of the _bad' website in the Le condition. One participant in the He condition disagreed with the confederate, while no one disagreed in Le condition, saying that they ―had really low expectations <of the website> before-hand."

Therefore, there was less disagreement with the confederate in this study than in the previous one, suggesting that the addition of the confederate influenced participants more so than just the written task description. Lastly, and also similarly as it was in Main Study 1, some participants in this study also had no expectations. In the He condition, two participants said that ―<the confederate> did not say anything about the website," and ―sorry, I'm really tired, I have no idea what she said but I had no biases. I've never seen a city council website." In the Le condition, three participants said that they had no expectations. Their reasons were that they never interacted with a city council website and did not know what to expect. One of the three participants said that the confederate had no influence on them because they ignored what they were saying at the beginning.

Summary of participant feedback results. The majority of participants in the He condition thought that the website was pretty, whereas the majority of participants in the Le condition would not agree with that. Similarly, only the participants in the He condition had positive things to say about the website's usability. In addition, participants in the Le condition were much more critical and more things seemed to bother them about the website's usability level. This strongly suggests that the first two research hypothesis are true, and that the low expectation influenced their perception of both variables, in accordance with the cognitive dissonance theory.

Beanplot Results

To gain a general feel for the data, the VisAWI-S and SUS scale results were graphed into beanplots, as in Main Study 1. Pre- and post-use visual appeal (Figure 6.6) and pre- and post-use perceived usability (Figure 6.7). In both of these figures, the first column on the left represents the HuHvNe condition, the middle one is the HuHvHe condition, and the one on the far right is the HuHvLe condition. The grey beanplots are the pre-use measures and the purple ones are the post-use measures. As mentioned earlier, the Ne condition was not re-done in Main Study 2, but the data was taken from Main Study 1 for the purposes of comparison.

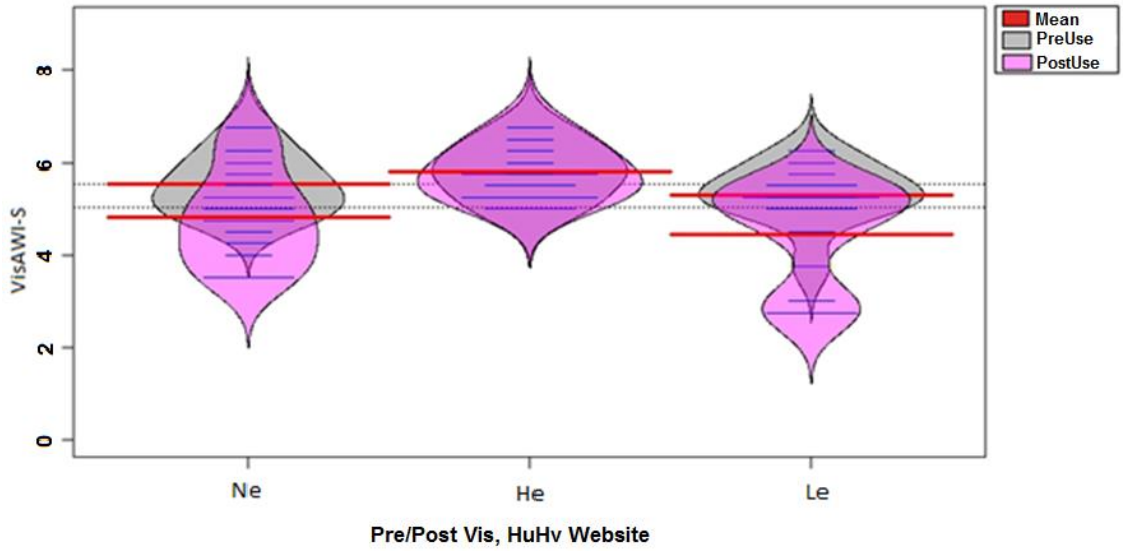


Figure 6.6. Beanplot of the pre- and post-use visual appeal results.

The beanplots and their means are slightly higher (about one point) for the pre-use visual appeal ratings than the post-use ratings in only the Ne and Le conditions. Pre-use and between conditions, the He condition was perceived to be slightly prettier than the control condition, which was perceived to be slightly prettier than the Le condition. Post-use, the participants in the Ne condition seem to have lowered their ratings of visual appeal while the ratings stayed identical to what they were pre-use in the He condition. Post-use for the Le condition, some participants seem to have stayed close to their opinions pre-use, while others thought it was quite a bit uglier after they interacted with it. Thus, the verbal implementation of expectations via confederate seems to affect participants‘ perceptions of visual appeal more drastically post-use.

Altogether, there seems to be support for the first hypothesis: visual appeal does differ between conditions in the same website, pre- and post-use.

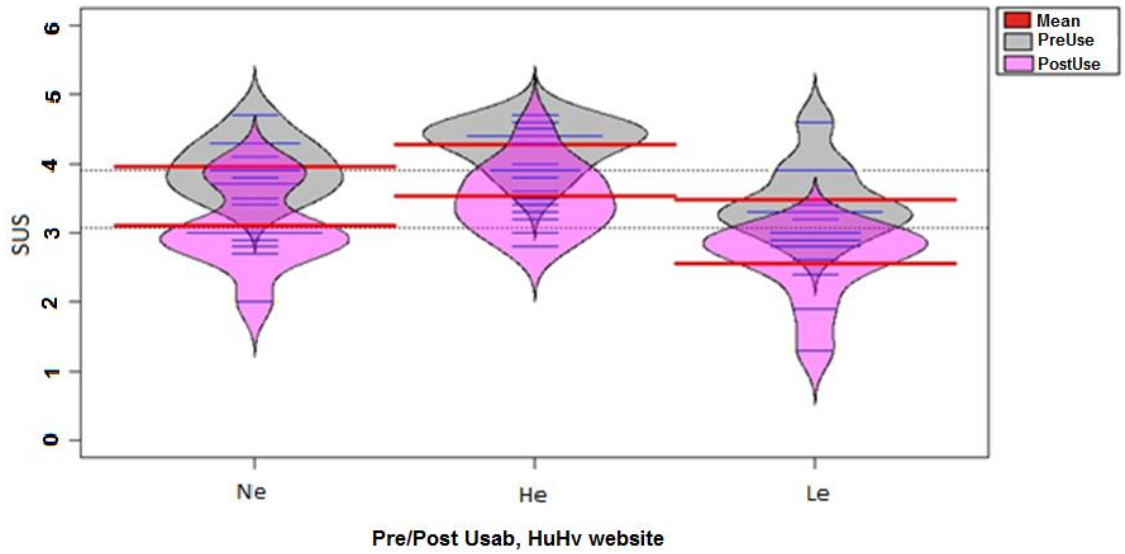


Figure 6.7. Beanplot of the pre- and post-use perceived usability results.

The effect of use seems to be more pronounced with perceived usability than with visual appeal given that use does not affect visual appeal in the He condition (see Figures 6.6 and 6.7). Pre-use perceived usability ratings are all slightly higher than the post-use ratings (i.e. the red lines representing the means are one point lower pre- than post-use). This suggests that use itself influenced the ratings of the website. Between conditions, it is evident that the He condition was perceived to be the easiest to use and that the Le condition was perceived to be the hardest, given the height of the distributions and the red lines which depict their means. The verbal implementation of expectations via confederate does seem to affect usability perceptions, and the impact of that expectation lasts post-use as well.

Therefore, hypothesis two seems to be supported: perceived usability differs between conditions in the HuHv website, both pre- and post-use. To statistically verify the significance of the findings in the beanplots, the next two sections deal with statistical assumptions and hypothesis testing.

Assumptions Testing

The assumptions for normality and homogeneity of variance were checked for each variable across all conditions and were not unilaterally met. Shapiro-Wilk tests showed that clicks ($p<.05$), post-use visual appeal ($p<.05$), and post-use usability ($p<.05$), each at HuHvLe, were not normally distributed. Post-use perceived usability at HuHvLe had a skewness of -1.627 (SE=0.687) and a kurtosis of 2.091 (SE=1.334), suggesting that it was not normally distributed. The rest of the factors appeared to be normal. The non-parametric Levene‘s test revealed that the homogeneity of variance assumption was not violated. Given that assumptions for normality were not met, that a variable was binary (success), some were discrete (clicks and hovers) and others continuous (time), and that sample size per condition was relatively small (n=10),

ANOVAs could not be applied to the data. As was done in Main Study 1, Kruskal-Wallis and Fishers Exact tests were applied where appropriate.

Statistical Hypothesis Testing

Three out of four statistical sub-hypotheses for the main effects for visual appeal and perceived usability, from Table 6.9, were statistically significant. Specifically, main effects were found in pre-use usability ($p<0.01$), post-use visual appeal ($p=.01$), and post-use usability ($p<.01$). This means that a statistical difference was found amongst the HuHvHe, HuHvLe, and HuHvNe. Paired comparisons showed that HuHvHe and HuHvLe differed in pre-use usability ($p<.01$), post-use visual appeal ($p<0.05$), and post-use usability ($p<.01$). Full results are in Appendix D6.

Table 6.9. Visual appeal and usability statistical hypotheses and results.

Hypotheses	Results
H_1: Visual appeal differs within each of the two website conditions (i.e. HuHv and LuLv), pre- and post-use.	
H_{1a}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe pre-use.	H_{1a}: Null not rejected.
H_{1b}: Visual appeal differs between HuHvHe, HuHvNe, and HuHvLe post-use.	H_{1b}: Null rejected ($p<.01$).
H_2: The perceived usability differs within each of the two website conditions pre- and post-use.	
H_{2a}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe pre-use.	H_{2a}: Null rejected ($p<.01$).
H_{2b}: The perceived usability differs between HuHvHe, HuHvNe, and HuHvLe post-use.	H_{2b}: Null rejected ($p<.01$).

Similarly, three out of four statistical sub-hypotheses, from Table 6.9 (full results are in Appendix D6), for the main effects for objective usability were significantly different. The results summary can be seen in Table 6.10 below. The average number of clicks per task ($p<.05$) differed within the HuHv website conditions. Pairwise comparisons showed that the number of clicks were different ($p<.01$) between the HuHvHe and HuHvLe conditions. Specifically, participants in the Le condition, on average, clicked more often per task (3.82 clicks) than those in the He condition (2.8 clicks). Main effects were also found for task completion time ($p<0.01$) and the average number of passed tasks ($p<.05$). Pairwise comparisons found that the difference in time ($p<.01$) and passes ($p<.05$) was between the HuHvHe and HuHvLe conditions. Participants took over half a minute longer to complete a task in the Le condition (i.e. 108.13sec in Le versus 70.67sec in He). The significance of the comparison of the average of passed tasks was confirmed ($p<.01$, one-tailed) with a Fishers Exact test, in which HuHvHe had a larger success rate (0.83) than HuHvLe (0.58).

Table 6.10. Objective usability statistical hypotheses and their results.

Statistical Hypotheses	Results
H_3: The average number of clicks per task differs between HuHvHe, HuHvNe, and HuHvLe.	H_3: Null rejected ($p<.05$).
H_4: The average number of hovers per task differs between HuHvHe, HuHvNe, and HuHvLe.	H_4: Null not rejected.
H_5: The average time to complete each task differs between HuHvHe, HuHvNe, and HuHvLe.	H_5: Null rejected ($p<.01$).
H_6: The average number of passed tasks differs between HuHvHe, HuHvNe, and HuHvLe.	H_6: Null rejected ($p<.01$).

Summary of statistical results. HuHvHe and HuHvLe differed in pre-use usability, post-use visual appeal, and post-use usability. In addition, the average number of clicks per task, average task completion time, and proportion of passes (success rate) differed between the HuHvHe and HuHvLe conditions.

Therefore, there is substantial evidence to support all three research hypotheses. Participants rated the same website as prettier and easier to use when they were told that it was going to be well made, pretty, and usable. Moreover, they struggled more with the website when completing the information retrieval tasks when told that the website was hard to use. Therefore, expectations influence both how participants viewed and interacted with the website.

HuHv Website Correlations

To get an understanding of the data as a whole, Spearman correlations were first examined between visual appeal and perceived usability, pre- and post-use without taking into consideration the different conditions. They can be seen in Table 6.11, where the columns and rows are both the measured variables: pre-use perceived usability (PreUsab), post-use perceived usability (PostUsab), pre-use visual appeal (PreVis), and post-use visual appeal (PostVis).

Table 6.11. Spearman Correlations for the HuHv website conditions.

	PostUsab	PreVis	PostVis
PreUsab	.669**	.419*	.564**
PostUsab	-	.233	.543**
PreVis		-	.524**

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

The majority of variables were significantly correlated, as seen in Table 6.11. Across the HuHv website, pre-use perceived usability was positively and significantly

correlated to both post-use usability (r=.669, p<.001) and pre-use visual appeal (r=.419, p<.05). Similarly, post-use visual appeal was highly and positively correlated with both pre- (r=.564, p<.01) and post-use (r=.543, p<.01) perceived usability. In addition, pre- and post-use visual appeal (r=.524, p<.01) were all also positively and moderately correlated. **As with the findings of the HuHv correlations in Main Study 1, these correlations also agree with the literature and show a relationship between usability and visual appeal pre-and post-use.** To examine the impact of the expectations, the next section describes the correlations per condition in the HuHv website.

Correlations per Condition

Upon separating the data between the HuHvHe and HuHvLe, the Spearman Correlations can be seen in Table 6.12. Out of 12 correlations, only one correlation was significant. Pre-use visual appeal was highly and positively correlated (r=.725, p<.01) with post-use visual appeal, only in the HuHvHe condition.

Contrary to the results in Main Study 1 of HuHvHe, where most of the variables were correlated, here the correlations disappear for the most part. Participants seem to evaluate the usability and visual appeal separately, when indeed they are the same on the HuHv website. In fact, given that the correlation between pre- and post-usability does not exist here, one might conclude that the high expectation set the bar perhaps too high, and that using the website impacted/brought down the rating of usability substantially, whereas visual appeal stayed highly rated. **The correlations for HuHvLe on the other hand are equally as insignificant as they were in Main Study 1**. This would suggest that participants initially had hope for the website but later agreed with the expectation, and lowered their ratings for it even more so than initially (given the discrepancy in means between He and Le).

Table 6.12. Spearman correlations between visual appeal and perceived usability for HuHvHe and HuHvLe, respectively.

HuHvHe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.098	.611	.349
PostUsab	-	.330	.519
PreVis		-	.725**

HuHvLe			
	PostUsab	**PreVis**	**PostVis**
PreUsab	.444	-.263	.268
PostUsab	-	.034	.409
PreVis		-	.125

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

These results strongly suggest that expectations do indeed influence ratings of visual appeal and usability, and that participants largely agree with these expectations. This also supports the cognitive dissonance theory, as participants reduced the dissonance by agreeing with the expectation.

Discussion

The discussion section is outlined as follows. The first section will summarize the results. This will be followed by a limitations and future research section. The conclusion section will be presented last in this chapter.

Results Summary

While the website presented to both groups of participants was the same, the majority of participants in the He condition thought that the website was pretty, whereas the majority of participants in the Le condition did not agree with that. No one in the Le condition mentioned that the website was easy to use. Instead, they were highly critical of the website‘s usability and visual appeal levels. Yet, eight participants in He did say that they liked the usability and visual appeal of the website.

For visual appeal, pre-use and between conditions, the He condition was perceived to be slightly prettier than the control condition, which was perceived to be slightly prettier than the Le condition. Post-use, the participants in the Ne condition seem to have lowered their ratings of visual appeal while the ratings stayed identical to what they were pre-use in the He condition. For usability, the He condition was rated as the easiest to use and the Le condition rated as the hardest.

Statistically, HuHvHe and HuHvLe differed in pre-use usability, post-use visual appeal, and post-use usability. This means that the same website was differently rated, depending on what the confederate told them before the experiment. Specifically, participants rated the website better when they were told it was going to be easy and pretty, and they rated it as worse when they were told the opposite. In addition, the average number of clicks, completion time, and the success rates differed between the HuHvHe and HuHvLe conditions. More precisely, the Le condition made more clicks, took nearly double the time, and had a lower success rate than the He condition, doing the same tasks and using the same website.

Across all three conditions, pre- and post-use correlations existed between visual appeal and perceived usability. These results are congruent with the general results of Main Study 1, as well as with current literature. However, when correlations were examined within each condition separately, only one pre-use visual appeal was highly and positively correlated with post-use visual appeal, in the HuHvHe condition. Nothing else was correlated. These findings differ from the literature, which normally does find a correlation between visual appeal and usability. Assuming that the sample size is large enough to indicate a correlation, this leads us to the conclusion that expectations did influence ratings of visual appeal and usability, and that participants largely agreed with these expectations. This also supports the cognitive dissonance theory, as participants reduced the dissonance by agreeing with the expectation.

Implications for Research Hypotheses

Evidence from participant feedback, the general graph behaviour, statistical and correlational calculations all point to the conclusion that all three research hypotheses are supported. Visual appeal and perceived usability differed between conditions in the same website, pre- and post-use, based on a pre-set expectation. Furthermore, the verbal implementation of expectations via confederate also impacted how participants interacted with the website, struggling more with it when they were told it was going to be hard. Participants agreed with the expectation given to them and were convinced that the same website was either great or terrible, depending on the condition. These results suggest that expectations do impact the perception and use of a website, but further experiences on different web sites with different populations is needed.

Implications for Website Design

For website design, unfortunately this means that how well a website is made is not the only factor that influences what people think about it. As demonstrated in this research through the use of a confederate, a bad reputation can turn people against your website, even if the reputation is not true. To overcome this, one should invest in marketing to give a website a more positive reputation right from the beginning. It will influence people before they use it and, according to the results of this study, last throughout use to influence their opinions after having used the website. In this study, participants were forced to use it, whereas in real life there are thousands of websites to choose from and competition can be fierce. If you advertise, people will (1) know about it, (2) know something *good* about it, (3) be willing to check your website out, and (4) like it a bit more after they use it.

Limitations and Future Research

Threats to construct validity. There were no threats to construct validity as the scales and measures used in this study were all widely used and accepted. There were no problems with them in previous studies in this thesis. However, in the SUS scale, one of the items asked participants if the website was _cumbersome' which is a term some students did not know. The researcher gave the definition and some synonyms (difficult, hard, etc.) and participants were able to complete the SUS scale successfully.

Threats to statistical validity. One obvious limitation is the small sample size, as was the case in Main Study 1. However, given that each participant took about an hour to test and that the testing had to be done in person, acquiring a larger data sample was nearly impossible to do. It was difficult to schedule and re-schedule participants and the confederate, who was also a full-time PhD student. Participants often late or did not show up, making it hard to finish data collection on time. This also influenced the confederate's schedule who tried to be as flexible as possible. Funding also restricted

this decision as each participant was thanked for their time with a $20 gift certificate, and the confederate was paid for their time as well.

Also, it would have been nice to view and compare pre-use perceived usability with pre-use visual appeal on the same graph. This was not possible because those two variables are differently scaled (1-5 and 1-7) and the beanplot would have been skewed. In future studies, it would be better if the two measures were scaled in the same way so as to allow for graphing and to make comparison more intuitive.

Threats to internal validity. This study was done using a confederate who acted like a participant just finishing the usability test, and either praised or complained about the website. The confederate was added to hopefully strengthen the implementation of expectations. However, this may not be the best way to do so given the unfamiliarity, untrustworthiness, and minimal exposure to the confederate and the expectation. Yet, in this study, the results showed that expectations did influence usability and visual appeal more so than in Main Study 1. Therefore, a confederate will be used in the next as well. However, only one confederate was used in this study. To balance the possible impact of gender, the next study will use one female and one male confederate.

Threats to external validity. Similarly as in Main Study 1, there were no pretests in this study to indicate the possibility of an interaction effect during testing. As previously mentioned, random university students were used because they are a representative sample of the general population and do not pose a threat to external validity (e.g. Svahnberg, Aurun, & Wohlin, 2008; Druckman & Kam, 2009). Also, the use of a confederate is a widely acceptable method in experimental studies. Thus, the results of this study are generalizable.

Summary

All three research hypotheses were supported to varying degrees. Visual appeal and perceived usability were rated as higher when the expectation was set to be high, and lower when the expectation was set to be low. In addition, participant performance was also affected by expectations. Comparing the results of Main Study 1and Main Study 2, the results suggest that verbally enforced expectations do impact the perception and use of a website, more so than just written task descriptions on their own. Based on this, we can conclude that using a confederate to verbally reinforce expectations was successful and a confederate will therefore be used in the next study as well. Given that expectations influence visual appeal and usability when both are either high or low, the next study will examine what happens with this relationship when the usability and visual appeal are incongruent with each other. In other words, the next study will examine the HuLv and LuHv website versions, along with HuLv and LuHv expectations.

Chapter 7. Incongruent Visual Appeal and Usability Levels

Main Study 3 Introduction

Main Studies 1 and 2 from Chapter 6 examined congruent cases of visual appeal and usability. In other words, they examined the easy/pretty (HuHv) and hard/ugly (LuLv) websites. The purpose of this third study was to gain a deeper understanding of what effect expectations had on usability and visual appeal.

There were three options. The first was to use the easy/pretty and hard/ugly website versions but to use partially congruent expectations: easy/ugly and hard/pretty. This would have given a deeper understanding of what happens when only one variable (between usability and visual appeal) is incongruent with the actual website levels. So, for example, with the easy/pretty website, an easy/hard expectation would have meant that the usability expectation was congruent with the website‘s actual level while the visual appeal expectation was incongruent. However, this would have also examined the impact of the congruent variable as well as the expectation on the incongruent variable.

The second option was to use the two other website versions: easy/ugly and hard/pretty, with completely congruent or completely incongruent expectations, as was the case with Main Study 1 and 2. The third option was to use the same website but with one congruent and one incongruent expectation, as was the case with the first option. However, the third option would have also had the interference of the influence of the congruent variable. While this would have provided more information on the relationship between usability and visual appeal, this would have given unclear information on the impact of expectation. The second option gave the highest potential to gain a deeper understanding of the purest impact of expectations on visual appeal and usability.

Thus, the purpose of Main Study 3 was to examine the influence of expectations on visual appeal and usability when they are incongruent with each other. Specifically, the easy/ugly (HuLv) and hard/pretty (LuHv) website versions and expectations were examined in this chapter. These website and expectation levels were chosen in order to gain a better understanding of the impact of expectations on visual appeal and usability.

Therefore, this study examined the first, second, fourth, and sixth research questions:

1. Do expectations influence visual appeal?
2. Are perceived and objective usability influenced by expectations?
4. What effect do verbal expectations have on usability and visual appeal?
6. What happens when visual appeal and usability levels are incongruent (i.e. one is high and the other is low)?

To answer these, the easy and ugly, HuLv, and hard but pretty, LuHv, versions of the Gold Coast city council website were used as the test case. Each website version was subjected to three expectation conditions all of which had incongruent visual appeal and usability levels. These expectations were: high usability and low visual appeal

(HuLv), low usability and high visual appeal (LuHv), and no expectations (Ne) which was the control condition. This way, the expectations for usability and visual appeal were either both congruent or both were incongruent with the actual website levels. Given the two website versions and three levels of expectations, there were six conditions in this study: (1) easy but ugly website with congruent expectations, HuLvHuLv, (2) easy but ugly website with hard but pretty expectations, HuLvLuHv, (3) easy but ugly website with no expectations, HuLvNe, (4) hard but pretty website with easy but ugly expectations, LuHvHuLv, (5) hard but pretty website with hard but pretty expectations, LuHvLuHv, and (6) LuHv website with no expectations, LuHvNe.

To explore the research questions, there were three hypotheses, similar to the ones in Main Studies 1 and 2. If expectations influence visual appeal and usability, then according to cognitive dissonance, participants should agree to the expectation given, and the perceived variables should be reported as either higher or lower than the control condition, in accordance with the variable‘s expectation level.

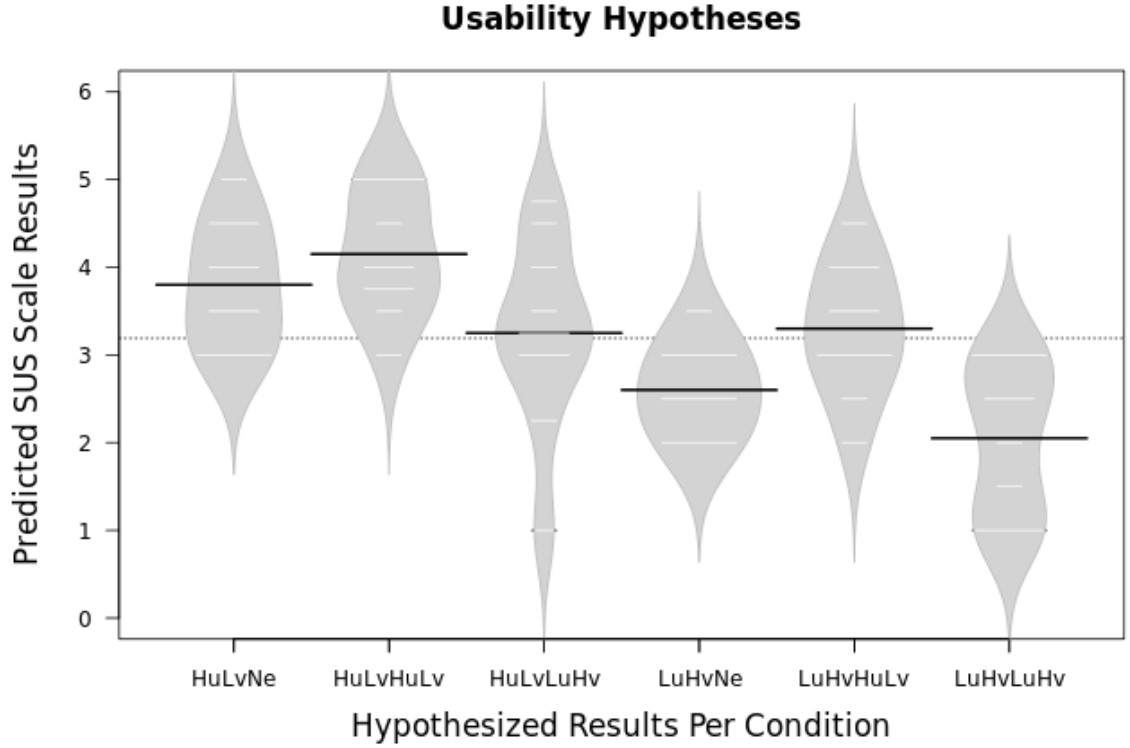


Figure 7.1. Beanplot of the hypothesized usability results.

Thus, the *first research hypothesis* states that when expectations of usability are set to be high and expectations of visual appeal are set to be low, then participants will rate them accordingly: they should perceive and rate usability as easier and visual appeal as prettier. Higher usability ratings and lower visual appeal ratings are expected because participants should be swayed to agree with the most recently learned information, being the expectation. Subconsciously accepting the expectation and adjusting the perception of the website will help them achieve consonance, according to the cognitive dissonance theory.

The hypotheses can be visualized in Figures 7.1 and 7.2, for usability and visual appeal respectively. The figures were created using dummy data. In Figure 7.1, the first three beans (to the left) correspond to the high usability, low visual appeal website, and the last three beans are the low usability and high visual appeal website conditions. The first bean is the control condition for the easy/ugly website, this is followed by a slightly higher bean for the easy/ugly expectation, since usability is predicted to be perceived as easier with the higher expectation. Then, the third bean is the lowest for the easy/ugly website since it pertains to the expectation in which usability is lowest for that website. The fourth bean pertains to the hard/pretty website control condition. The fifth bean is the hard/pretty website with easy/ugly expectations; since the expectation for this condition is that the website is easier to use then it is predicted that it will be easier to use – hence its higher mean and distribution compared to the control condition. The sixth and final bean is the hard/pretty website with congruent expectations. Thus, it is expected that it will be perceived to be the lowest given its low actual and expected usability levels. These trends are predicted both pre- and post-use.

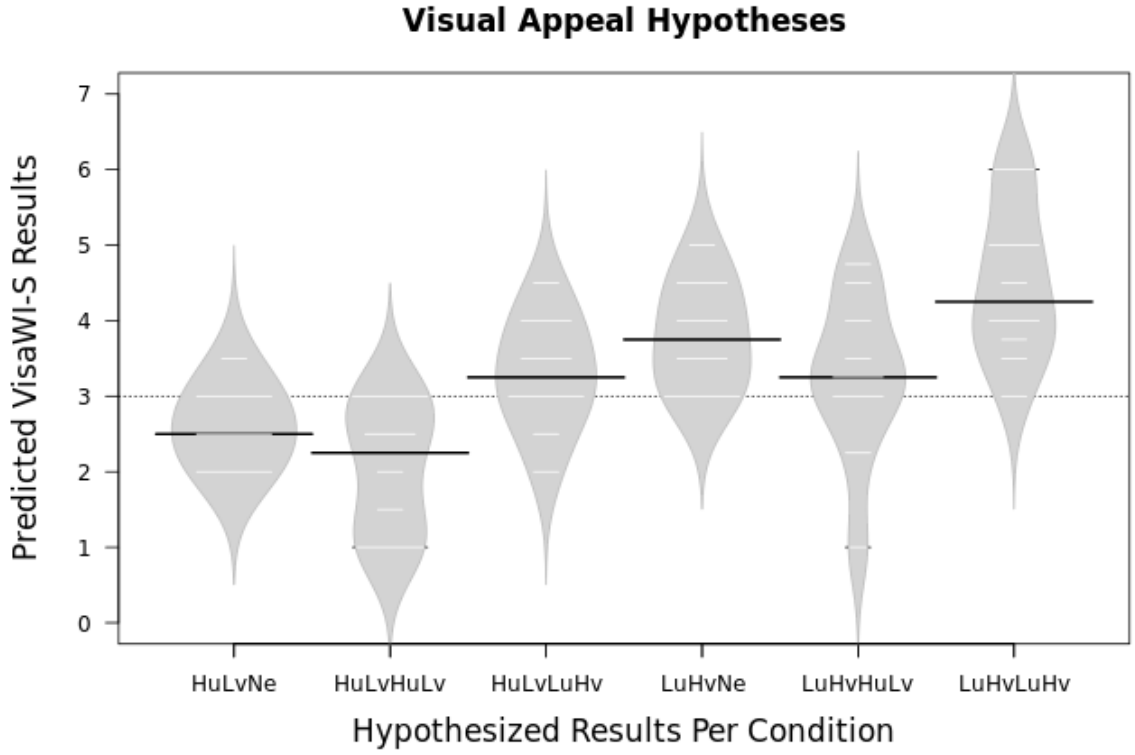


Figure 7.2. Beanplot of the hypothesized visual appeal results.

In Figure 7.2, the beanplots represent the hypothesized results for visual appeal. Starting from the left, the first three beans correspond to the high usability and low visual appeal website. The last three beans correspond to the low usability and high visual appeal website conditions. As it was in Figure 7.1, here, in Figure 7.2, the first bean is the control condition for the easy/ugly website. The second beanplot is a lower bean for the easy/ugly expectation, because visual appeal is predicted to be perceived as uglier with the lower expectation. Then, the third bean is the highest for the easy/ugly

website since it pertains to the expectation in which visual appeal is highest for that website. The fourth bean is the predicted hard/pretty website control condition which is higher than the easy/ugly control condition. The fifth bean is the hard/pretty website with incongruent expectation, which makes it lower than the control because the expectations are worse. The sixth bean is the hard/pretty website with congruent expectations, which is rated as prettiest because of its high actual and expected visual appeal levels. These trends are predicted both pre- and post-use.

Using the same reasoning as in the first hypothesis, the *second hypothesis states* that lower ratings are expected for usability when expectations are set to be low, and higher ratings are expected for visual appeal when the expectations are high. These conditions can be seen in Figures 7.1 and 7.2, in the third and sixth beans.

To examine these hypotheses, causational and correlational statistics were computed, along with brief qualitative analysis. For the first two research hypotheses, there were four statistical hypotheses, one for each variable, seen in Table 7.1. Each of the two hypotheses has four sub-hypotheses which were tested for main effects. There were four sub-hypotheses because there were two websites (HuLv and LuHv) and two points of use (pre and post), creating a 2X2 matrix of tests. If a main effect was found, then pairwise comparisons were calculated in order to determine which conditions differed from the others.

Table 7.1. Visual appeal and perceived usability statistical hypotheses and tests used.

Hypotheses	Tests
H_1: Visual appeal differs within each of the two websites (HuLv and LuHv), pre- and post-use.	Main Effects: Independent-Samples Kruskal-Wallis Test
H_{1a}: Visual appeal differs among HuLvHuLv, HuLvLuHv, and HuLvNe pre-use.	Paired comparisons: Kruskal-Wallis multiple comparison tests, i.e. Wilcoxon-Mann-Whitney
H_{1b}: Visual appeal differs among HuLvHuLv, HuLvLuHv, and HuLvNe post-use.	
H_{1c}: Visual appeal differs between LuHvHuLv, LuHvLuHv, and LuHvNe pre-use.	
H_{1d}: Visual appeal differs between LuHvHuLv, LuHvLuHv, and LuHvNe post-use.	
H_2: Perceived usability differs within each of the two website pre- and post-use.	Same as H_1.
H_{2a}: The perceived usability differs between HuLvHuLv, HuLvLuHv, and HuLvNe pre-use.	
H_{2b}: The perceived usability differs between HuLvHuLv, HuLvLuHv, and HuLvNe post-use.	
H_{2c}: The perceived usability differs between LuHvHuLv, LuHvLuHv, and LuHvNe pre-use.	
H_{2d}: The perceived usability differs between LuHvHuLv, LuHvLuHv, and LuHvNe post-use.	

We also wanted to find out if expectations affect participant performance. Consequently, the third research hypothesis states that expectations affect participant performance (in the form of the classical objective usability measures). Specifically, in the easy but ugly website (HuLv), participants should find the easiest to use to be the HuLv expectations group (i.e. congruent expectations), followed by the control group, and then should struggle the most in the LuHv expectations conditions (i.e. incongruent expectations). In the hard but pretty website (LuHv), again, participants should find the HuLv website as the easiest, followed by the control group, and the hardest should be the LuHv expectations condition. With similar reasoning as in the previous studies, this is hypothesised because participants who perceive it to be either easier or harder to use may reflect their perceptions in how they use the website as a confirmation bias.

Table 7.2. Objective usability statistical hypotheses and tests used.

Hypotheses	Tests
H_3: The average number of clicks per task differs within each website. H_{3a}: Clicks differ between HuLv(HuLv, LuHv, Ne). H_{3b}: Clicks differ between LuHv(HuLv, LuHv, Ne).	Same as H_1.
H_4: The average number of hovers per task differs within each website. H_{4a}: Hovers differ between HuLv(HuLv, LuHv, Ne). H_{4a}: Hovers differ between LuHv(HuLv, LuHv, Ne).	Same as H_1.
H_5: The average time to complete each task differs within each website. H_{5a}: Completion time differs between HuLv(HuLv, LuHv, Ne). H_{5a}: Completion time differs between LuHv(HuLv, LuHv, Ne).	Same as H_1.
H_6: The success rate per participant differs within each website. H_{6a}: Tasks passed differ between HuLv(HuLv, LuHv, Ne). H_{6a}: Tasks passed differ between LuHv(HuLv, LuHv, Ne).	Same as H_1 and Fishers Exact test.

As in previous studies in this thesis, the third research hypothesis has four statistical hypotheses since there were four objective usability measures (i.e. hovers, clicks, time, and passes), seen in Table 7.2. The first objective usability measure was the average number of clicks per task per participant, henceforth referred to as ‗clicks'. If participants struggled more with the usability of the website, then there would be more clicks, and this also applies to hovers. The second objective usability measure was the average number of hovers per participant per task, which will be referred to as ‗hovers' from now on. The average time taken to complete a task was the third objective usability measure and will be referred to as ‗time'. If participants struggled with the usability of the website, then that would be reflected in longer time periods per task. The last objective usability variable was success, which was a binary variable in terms of pass/fail. A pass occurred when an answer to a task was correct and found within three minutes. Any form of deviation from that definition (e.g. took longer than three minutes) and the task was considered a fail. Success rates would be lower in the case

that the participants struggled more with the website. These objective usability measures were chosen in order to get an understanding of the effect of expectations on participants during website use. Each of the four hypotheses has two sub-hypotheses that were tested, one for each website.

As was done in Main Studies 1 and 1.2, correlation analysis was also calculated here, in search of support for the hypotheses. This work was exploratory, as with the qualitative results, to supplement the statistical results. Therefore, correlations will be examined between visual appeal and usability pre- and post-use in this study as well.

Method

Participants

A sample of 60 (38 males, 22 females; 49 aged 18-30 years, 11 aged 31+) Swinburne University student volunteers participated, all with 20/20 or corrected to 20/20 vision, and screened for colour blindness. All participants were technology-savvy regular Internet users. Thirty-five were born in an English speaking country and 47 spoke it frequently at home. Forty-two out of the 60 were undergraduate students, 14 masters, and four PhD students. Out of the 60, 38 studied computer science, 11 business, four design, three education, one each one each of arts, psychology, engineering, and law. Thirty participants were not at all familiar with the purposes of city councils, 24 were somewhat familiar, and six were very familiar. Participants were randomly assigned and individually tested, approximately one hour per session, ten participants per condition.

Apparatus, Materials, and Location

All apparatus and materials pertaining to this study are the same as in Main Study 2, and can be seen in Appendix B. The same usability lab used in the previous study was also used here, with the same computer screen and software. Participants‘ audio and video were not recorded.

All documentation pertaining to Main Study 3 can be found in Appendix E. The same informed consent, project information, demographics form, usability and visual appeal scales were used as in the previous studies in this thesis. As mentioned earlier, two versions of the website were used: HuLv and LuHv. Three different task descriptions and speeches for the confederate were prepared, a paragraph long each, setting expectations high for usability and low for visual appeal, low for usability and high for visual appeal, or neither (the same control paragraph used in Main Study 1). The confederate used a standard script found in Appendix E1. For example, the low usability, high visual appeal task description was:

> —Welcome to Gold Coast, a big city in Australia. You recently got a job there and will be moving quite soon. Before you start packing and head off, you‘re going to check the city council website out, to

> get some information which will help you get ready for the move. Recent surveys have found that the website is as beautiful as the gorgeous city. The colours are very professional and flattering. However, people are finding it incredibly hard to use, even for a government website. They said using it was harder than doing their taxes."

As was the case with previous studies in this thesis, the SUS scale was used as a subjective usability measure pre- and post-use, and objective usability in the form of performance measures were acquired per task. These were: the number of clicks, the number of hovers, task completion time in seconds with a maximum of 180sec (i.e. three minutes), and success (pass/fail; pass if the answer is correct and within time limit). For more information and the definitions of these measures, please see Chapter 5, Phases 3 and 4. The higher the number of clicks, hovers, and time per task, the more participants had to explore the website in order to find the answers to the tasks, suggesting that higher values for these variables indicate lower usability levels. Inversely for success, if the success rates were higher or closer to 1, then participants were more likely to finish a task correctly and higher values for the average number of passed tasks indicates a higher usability level for the website.

The same ten information retrieval tasks from Main Study 1 and 2 were given to participants, again in random order. An example of a task is: "Who is the councillor of Robina?" All tasks were on the same page.

Procedure

The procedure from Main Study 2 was repeated here, including the confederate. A confederate would be in the experiment room, picking up their things and getting ready to leave as the participant entered the room. The experimenter would ask the confederate if they were all done and the confederate would respond that they were just leaving. The experimenter would thank them and tell the participant to go ahead in and wait a minute while the experimenter left to set up the computers. The confederate then told them the usability and visual appeal expectations in the form of their experience with the website and left. The experimenter came back into the room and then started with the brief and rest of the procedure from Main Study 1. Each participant was asked for their feedback at the end of their session, in addition to any comments that they may have had during website use. See Appendices B and E for all instruments used.

Design

This was exactly the same as Main Study 1, this study adopted a two-by-three (two websites, three sets of expectations) between-group design. The website was shown in two parts, the first was the slideshow needed for pre-use data, and the second was the functioning website needed for post-use data.

Data Analysis

The data was analysed in the same way as it was in Main Studies 1 and 2. Normality and homogeneity of variance were tested. Then, the averages were calculated per condition pre- and post-use for visual appeal, perceived usability, and mood. The average results for the objective usability measures were calculated across tasks, per participant. Non-parametric tests were applied, chiefly Kruskal-Wallis for main effects, Fisher‘s Exact Test and Wilcoxon Mann-Whitney for pairwise comparisons. Spearman‘s Correlation Coefficients were used to examine other relationships that may exist between variables.

Results

The results section is structured as follows. The first section discusses the qualitative findings, in the form of participant feedback during and after their test sessions. The second section describes the statistical assumptions testing which was necessary in order to determine which statistical tests to further apply to the data. Naturally, the statistical hypotheses testing results follow. The results section concludes with the correlations between usability and visual appeal. The results section is followed by the discussion section.

Participant Feedback

At the end of their sessions, participants were asked four questions: what they thought of (1) the usability, (2) the visual appeal, (3) if they believed the task description, and (4) if they agreed with the task description. The feedback will be split between websites, with the HuLv website results presented first, and the LuHv participant feedback results presented second.

HuLv participant feedback results. For usability, the HuLv website had seven out of thirty participants who thought that the website was easy to use, (four from the HuLv expectation group, two from LuHv, and one from the control group). In addition to having the most participants commenting positively about usability, the comments from the HuLv expectations group were also much more optimistic. For example, participants from the HuLv expectations group said that it ~~–g~~ets easier with use and I learned pretty quickly – well designed” and another said that it was ~~–in~~formative, easy to use, and overall good.” The participants from the LuHv and control groups expectations group were slightly less positive, and said that the ~~–li~~nks on the homepage worked the best.” Usability issues were more readily discussed.

The HuLv website had 15 out of 30 participants who thought that the website was hard to use, (four from the HuLv expectation group, seven from LuHv, and four from the control group). The comments were all relatively the same, with the main concerns being that there was no functional search bar, the right-hand side menu took time to get

used to, and that there was too much text/info/options to go through. The LuHv expectation participants were by far the most negative in their feedback from the HuLv condition, saying that it was ~~"~~extremely terrible," ~~"~~not professional," that they would ~~"~~walk away from the website", and that it was ~~"~~super crazy." One participant admitted that they found it hard to use because of colours. The HuLv control group had three participants with minor terminology issues. For example, one participant asked what vaccines were. The HuLv expectations group had one participant who reported domain-specific terminology issues. These issues were also found in the previous studies with the non-native English speakers.

When asked about visual appeal, only three out of 30 participants from the HuLv website (one from the HuLv expectation group, two from LuHv, and no one thought it looked good from the control group) said that it looked alright. Participants in the HuLv expectations group were the most critical in their positive feedback, saying that ~~"~~the website should be ok for someone colour blind, even though it's gross." The high-visual appeal expectations group (LuHv) were relaxed in their comments, saying ~~"~~I don't mind the colours," and ~~"~~it doesn't look horrible, not disturbing, just adequate." One even said: ~~"~~I think its visually attractive - good colours." No one from the control group said anything positive about the website.

More people (rightfully) complained about the visual appeal of the HuLv website. In total, 17 out of 30 participants (six from the HuLv expectation group, eight from LuHv, and three from the control group) said that they thought it was not visually appealing. While participants in all three conditions agreed that the website had terrible visual appeal, participants from the HuLv condition group were the most negative in their feedback, with the control group having the most neutrally negative comments. For example, one participant from the HuLv condition said that they ~~"~~hated it since I first saw it," while another said ~~"~~someone had a stroke while choosing the colour palette." Interestingly, these groups of participants did mention that the usability was well done (e.g. ~~"~~it's well-structured but the colours are horrible"), which does suggest that they were able to differentiate between these two variables. They also mentioned that if the colours were better, the website would be better overall as well. The participants from the LuHv condition said that they wouldn't be able to recognize the councillors from the picture portraits because all the colours in the images were inverted (i.e. negatives). One participant said, ~~"~~I don't like the colours, developers had a visual impairment." However, participants were divided about the colours, with one saying that there were too many colours and another saying that it need more colour. In the control condition, only three complained about the visual appeal. Their main complaint was about the colour scheme as well, saying that it was poorly chosen.

When participants from the HuLv website were asked about their expectations of the website prior to the testing, one person in the HuLv reported that they did not expect it to be easy (i.e. disagreed with confederate). In the LuHv expectations condition, one person believed the confederate and said, ~~"~~I thought it would be bad but I'm biased." Also in the LuHv expectations condition, three people reported no expectations, saying that they did not remember what the confederate said, that the confederate did not say

anything, or that they did not take the confederate seriously. No one from the control condition reported any expectations or biases.

LuHv participant feedback results. No one from the LuHv website mentioned that it was easy to use or well designed. Yet, 20 out of 30 said that it was hard to use (seven from LuHv, seven from HuLv, and six from the control condition). Similarly as in the HuLv website, the main concern across the conditions was that there were too many options, too much information, and no search bar, which made it overwhelming. The control condition seemed to have the most docile responses, followed by the HuLv expectations condition, and then the most hostile opinions came from the LuHv condition. For example, one participant from the HuLv condition said, "change the usability, don't change the visual appeal." Meanwhile, participants in the LuHv condition said that "the menu bar is bad, the elderly would not be able to use it, needs to be much simpler, as if it were created for tourists." One participant went as far as saying that the website was a "pain in the [*behind*], super hard, looked functional when I first saw it. It's terrible," and "I'm a software engineer and whoever created this website should be fired or at least ashamed."

From the LuHv website, 16 out of 30 people said that it was visually appealing (six from the HuLv expectation group, eight from LuHv, and two from the control group). Participants in the HuLv condition said that it looked nice and that it was a standard city council website (i.e. "what you'd expect of a city council"). Participants in the LuHv expectations group agreed with his, but also added that the usability dragged the visual appeal down. For example, one said that "visual appeal isn't bad but usability completely overshadows it, completely hate it." Statements like these clearly indicate that these visual appeal and usability influence each other. The control group was the most positive, saying that it was "reasonably appealing, 7/10", and "fairly pretty, actually visually pretty good." Only seven people said that the website was not visually attractive (five from HuLv and two from the control group). Both groups mentioned that there was too much gray, which did not represent the Gold Coast.

When asked about their expectations, the LuHv website with HuLv expectations had one participant who said that they agreed with the confederate and that it was even harder than they said it would be. Three disagreed, with one saying "[the confederate] said it was ugly but easy but I didn't believe her, I'm not biased," while another said "she's got a really odd opinion of what's ugly. Don't listen to her, she's wrong." One other participant said that they had no expectations and another said that they were suspicious of the confederate because she had the same accent as the researcher (i.e. not an Australian one). The LuHv expectations group had one participant who agreed with the usability expectation, and two had no expectations because they either did not remember the confederate or said that she did not say anything. The control group reported no expectations.

Summary of participant feedback results. **Overall, participants were able to differentiate usability and visual appeal and state which one of the two needed to**

be improved. However, they did mention that the worse variable lowered the rating of the better variable. For example, poor usability brought down the rating for visual appeal as well. In addition, there seems to be an impact of expectations. Higher expectations lead to slightly more positive feedback towards the corresponding variable, and negative expectations allowed participants to be more hostile in their feedback but only towards the corresponding variable, leading towards support for the hypotheses. To examine the data visually, the next section will show the visual appeal and perceived usability results using beanplots.

Preliminary Beanplot Results

Beanplots were created to gain a general understanding of the data, with pre- and post-use visual appeal in Figures 7.3 and 7.4 respectively, and pre- and post-use perceived usability in Figures 7.5 and 7.6, respectively. In all of the figures, the gray beanplots are the HuLv website measures and the purple ones are the LuHv website measures. The first columns on the left represents the control (N for none) conditions, the middle columns are the HuLv (i.e. easy but ugly) expectation conditions, and the ones on the far right are the LuHv (hard but pretty) expectation conditions. The red lines indicate each condition‘s mean. The dotted lines indicate the website‘s overall mean, across the three expectation conditions (control, HuLv, and LuHv).

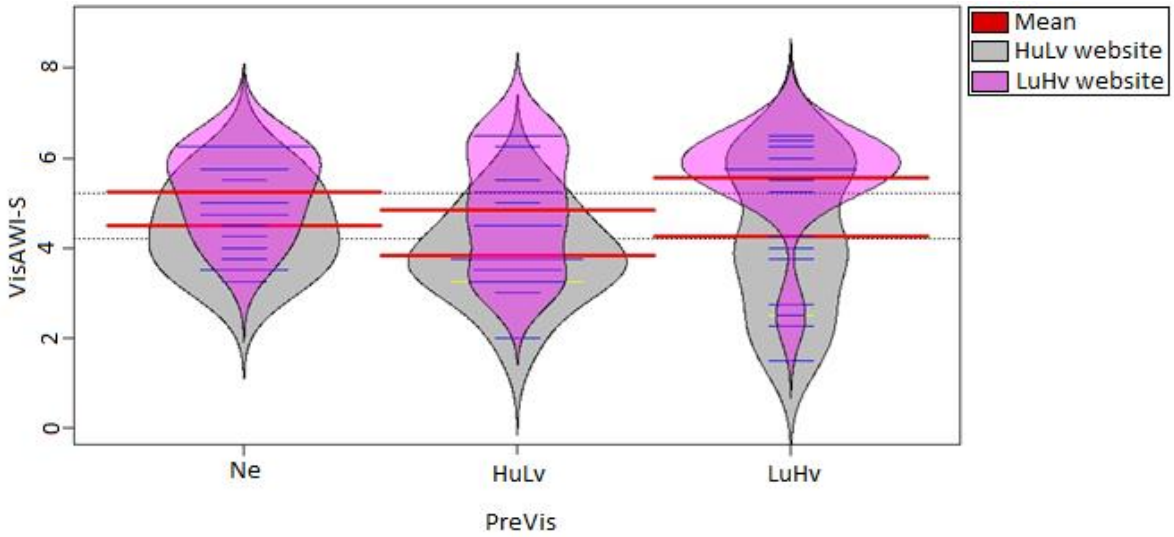


Figure 7.3. Beanplot of the pre-use visual appeal results.

Pre-use, the distributions in the beanblots in Figure 7.3 show that the visual appeal was generally rated higher (i.e. prettier) in the purple beans which were the LuHv website conditions. This result accurately reflects the website‘s actual visual appeal levels, with the prettiest condition being the LuHvLuHv. The distribution of the HuLv website control group (the first gray bean) appears to be normal and has a mean that is higher than the two experimental groups of the same website. The HuLv website has the lowest mean, suggesting that the low expectation did impact the visual appeal

rating to be lower than the other two conditions. **Also in the HuLv website, the LuHv condition has a higher mean than the HuLv expectation, suggesting that the high expectation did marginally increased and the low expectation marginally lowered the rating of visual appeal in the respective conditions, pre-use.**

Also can be seen from Figure 7.3, **in the LuHv website, where the website was pretty but hard to use, the visual appeal was rated highest in the high visual appeal expectation condition, and lowest in the low visual appeal expectation condition**. However, the LuHv condition (which corresponds to the actual website level) appears to be slightly bimodal, with a very small number of participants disagreeing with the high expectations of the visual appeal, and rate it as uglier than the low expectation category. For these couple of participants, the expectation worked inversely.

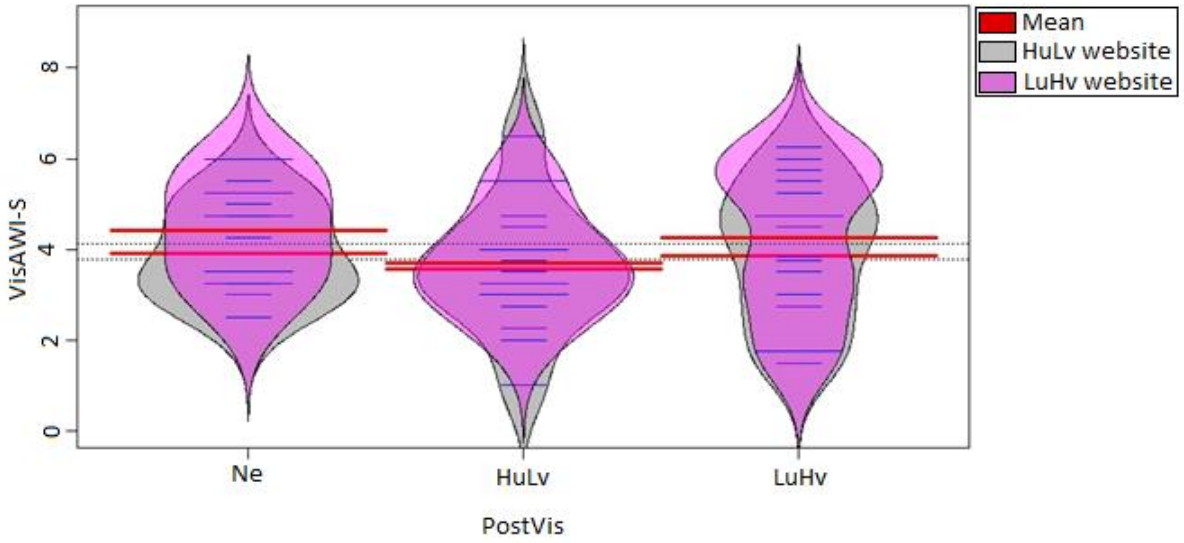


Figure 7.4. Beanplot of the post-use visual appeal results.

Post-use, visual appeal more or less equalizes throughout all six conditions, as seen in Figure 7.4. All of the ratings are lower post-use than they were pre-use. **This suggests that website use impacted the ratings of visual appeal.** Specifically, low usability may have lowered the visual appeal ratings in the LuHv website.

Out of the slight variations between the means within a website, the low visual appeal expectation group (HuLv column in Figure 7.4) have the lowest post-use visual appeal means, for both website versions. This does suggest that the low expectation impacted the perception of the website‘s appeal. The right- and left-most columns have similar means, with the high visual appeal expectations condition having a slightly higher distribution than the control group. The LuHv (purple beans) website is still rated as slightly prettier than the low visual appeal website (HuLv – gray beans) but the averages largely do not portray the actual difference in visual appeal between the two website versions. For the HuLv website, the ratings slightly dropped for visual appeal, between pre- and post-use. In addition, post-use, as seen in Figure 7.4, for the HuLv website, the ugliest rated was the low expectation condition, while the highest was the control condition but the high expectation condition for visual appeal was very close

second. A similar result occurred in the LuHv website. Statistical tests were done to determine if the difference was significant.

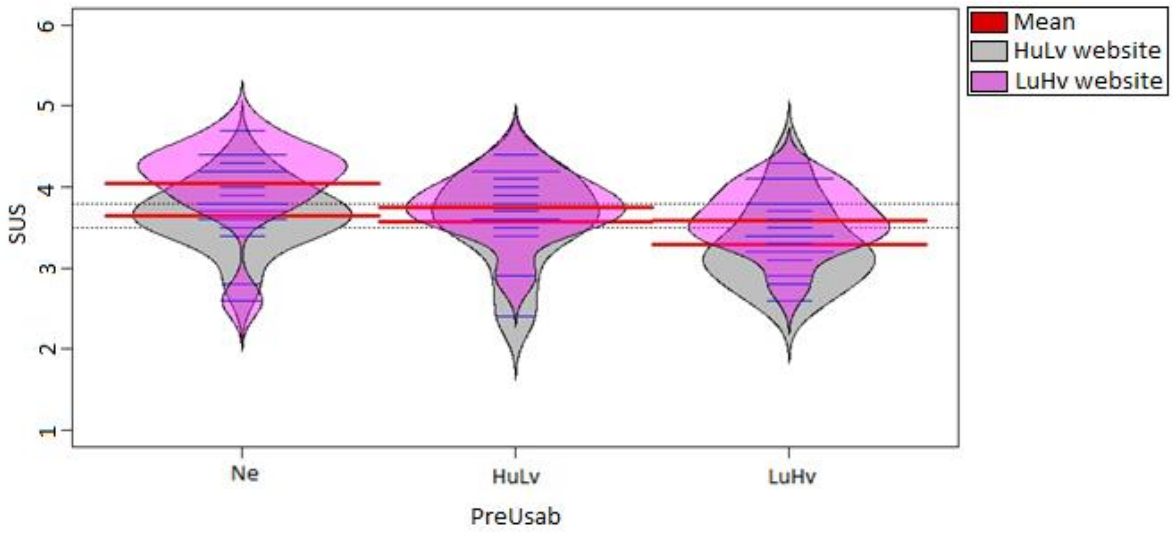


Figure 7.5. Beanplot of the pre-use perceived usability results.

Pre-use usability ratings across all six conditions can be seen in Figure 7.5. **The control condition for the hard but pretty website (LuHvNe, first column, purple bean) was rated as easiest to use,** thanks to the high visual appeal of the website and no expectations before-hand. However, this condition also appears to be bimodal, given the second hump at the bottom suggesting that one or two participants thought that it was not going to be easy to use. The high expectations condition is rated as second highest in usability pre-use, followed by the low expectations for usability condition, in the LuHv website. **The low expectation condition seems to show more variance from the control condition, suggesting the impact of low expectations was stronger than the high expectation condition.** This result is consistent with Main Study 2 pre-use usability ratings, in which judgments for pre-use usability were strongly based on pre-use visual appeal and on the expectation. For the HuLv website, the results follow the same trend as well.

The usability ratings completely change post-use (as seen in Figure 7.6 when compared to the pre-use ratings in Figure 7.5). Post-use ratings normalize across all conditions. Post-use usability ratings dropped for the LuHv website (purple beans), to better reflect the website‘s poor usability. Within the HuLv website, the highest rated usability levels came from the high usability expectations group, while the lowest are from the low expectations condition. Again, **showing evidence that low expectations impact user perception of usability.** However, the differences within each website seem to be marginal and statistics were done to determine the significance of these findings. That being said, there seems to be an impact of expectations, and of use, on the perception of visual appeal and usability.

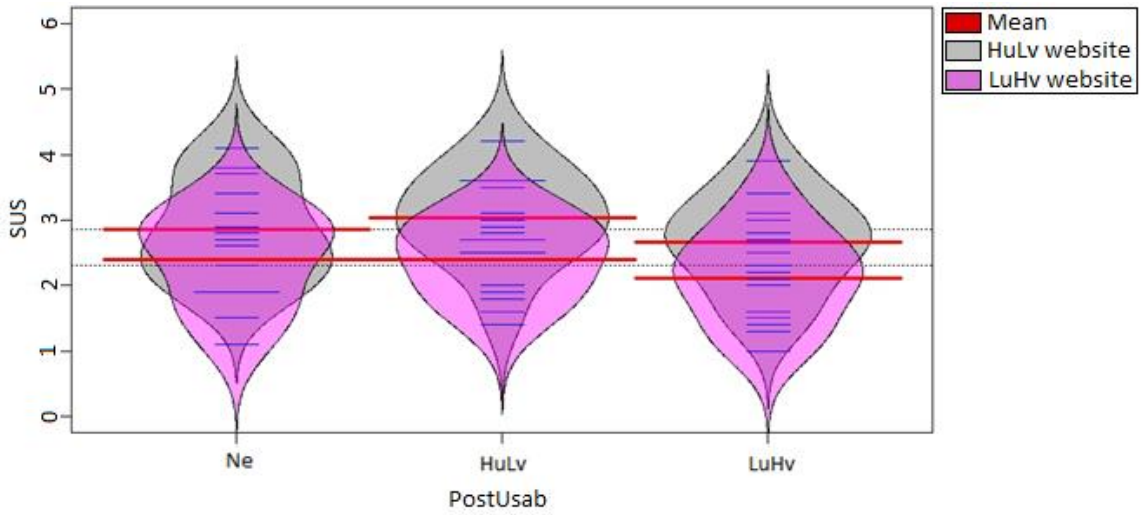


Figure 7.6. Beanplot of the post-use perceived usability results.

Beanplot result summary. Pre-use visual appeal was rated highest in the LuHvLuHv condition, and lowest in the HuLvHuLv condition (i.e. when expectations were congruent with the visual appeal levels). Post-use visual appeal was rated as highest in the LuHv control condition, with the LuHvLuHv condition appearing slightly bimodal, and the HuLvHuLv having the lowest dip. However, these were less apparent than pre-use. Pre-use perceived usability ratings showed that the LuHv website was altogether rated as easier to use, with the control group rated as easiest. The lowest mean came from the HuLvLuHv condition, in which both the visual appeal and expectation of usability were low. Although, the HuLvHuLv condition (where the website was ugly but easy, and the expectation was congruent) had a distribution that was somewhat lower than the HuLvLuHv condition, which had the lowest mean. That being said, the distributions for pre-use usability were not as easily differentiable as the pre-use visual appeal ratings. Post-use usability ratings showed a clear difference between the two websites, in which the LuHv website ratings of usability sank. Post-use usability means appear to be more or less the same within a website with the exemption of the low expectation condition, which was lower than the other two.

In general, it seems that the high visual appeal website and expectation levels increased ratings of both usability and visual appeal, especially pre-use. Specifically, the highest rated usability levels came from the conditions in which usability was said to be highest, and likewise for visual appeal. The lowest ratings of these two variables came from the conditions in which expectations were low. **Thus, expectations do seem to impact the perception of visual appeal and usability**, more vividly pre-use, supporting the hypotheses. However, some of the differences seem to be marginal and statistics are used below to determine the significance of these findings.

Assumptions Testing

As was done in previous studies, statistical assumptions for normality and homogeneity were tested in order to determine which statistical tests were appropriate to

apply to the data. The assumptions for normality and homogeneity of variance were checked for each variable across all conditions and were not unilaterally met.

Shapiro-Wilk tests showed that pre-use visual appeal at LuHvLuHv ($p=.001$) and pre-use usability for the LuHv control condition ($p<.05$) were not normally distributed. This was confirmed with the skewness and kurtosis measures. Specifically, while the skewness of pre-use visual appeal for LuHvLuHv was -2.498, (SE=0.687), the kurtosis measure of 7.038 (SE=1.334). Moreover, pre-use usability in the LuHv control condition had a skewness of -1.878 (SE=0.687) and a kurtosis of 4.546 (SE=1.334), revealing that it may not be normally distributed. The rest of the factors appeared to be normally distributed. The non-parametric Levene's test revealed that the homogeneity of variance assumption was not violated. Given that assumptions for constant variance and normality were not met, that some variables were ordinal (the likert scales used for visual appeal and usability), one was binary (passes), some were discrete (clicks and hovers), another was continuous (time), and that sample size per condition was relatively small (n=10), ANOVAs could not be applied to the data. As mentioned earlier, Kruskal-Wallis and Fishers Exact tests were applied where appropriate.

Statistical Hypothesis Testing

Of the eight statistical sub-hypotheses tested for visual appeal and perceived usability, only one was found to be significant. All results tables are in Appendix E3. Pre-use perceived usability ($p<.05$) was found to vary in the LuHv website conditions (i.e. a main effect was found). In other words, pre-use perceived usability differed amongst LuHvHvLu, LuHvLuHv, and LuHvNe. Paired comparisons showed that LuHvLuHv and LuHvNe differed in pre-use usability ($p<.05$). Therefore, partial statistical evidence exists for the first hypothesis: usability was rated lower when expectations were set to be low, especially between the control group and the low-expectation group, which rated the website as harder to use. This difference was only found pre-use, suggesting that the impact of the expectation was only strong enough to influence ratings before having been exposed to the website for roughly an hour. However, this difference was neither found in the HuLv website conditions, nor with visual appeal.

No significant results were found for objective usability. This finding, or lack thereof, suggests that low visual appeal created just as much difficulty as low usability in completing the tasks. Expectations did not seem to significantly impact use. Thus, there is insufficient statistical evidence to conclude that the third research hypothesis is true, since there is insufficient evidence to state that participants struggled more when expectations were low or that they did better when expectations were high.

General Correlations

Spearman correlations were first examined between all the variables, without taking the conditions into account (n=60). The majority of variables were significantly

correlated, as seen in Table 7.3 below. In this and other tables in this subsection, the columns and rows are both the measured variables: pre-use perceived usability (PreUsab), post-use perceived usability (PostUsab), pre-use visual appeal (PreVis), post-use visual appeal (PostVis), the average number of clicks, the average number of hovers, the average completion time per task, the proportion of passed tasks. If there were no conditions, then visual appeal and perceived usability pre-use were positively and moderately correlated (r=.479, p<.01). Post-use, the correlation between visual appeal and perceived usability was slightly weaker but still positive and significant (r=.426, p<.01).

Table 7.3. Spearman correlations for all website conditions.

	PreUsab	PostVis	PostUsab	Hovers	Clicks	Time	Pass
PreVis	.479**	.608**	.003	.101	.114	.200	-.026
PreUsab	-	.426**	.261*	-.004	-.009	.055	.032
PostVis		-	.445**	-.211	-.020	-.013	.037
PostUsab			-	-.456**	-.336**	-.361**	.292*
Hovers				-	.592**	.699**	-.495**
Clicks					-	.618**	-.459**
Time						-	-.811**

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

Correlations per Website

Upon separating the data between the two website versions (n=30 in each), HuLv and LuHv, the Spearman Correlations can be seen in Tables 7.4 and 7.5 respectively. In the HuHv website conditions, there were eight significant correlations in total, as seen in Table 7.4. Pre-use perceived usability was correlated with post-use perceived usability (r=.533, p<.01) and pre-use visual appeal (r=.539, p<.01). This suggests that participants did not drastically change their opinions on usability after having used the website, and that usability judgements were largely based on the website's visual appeal prior to using it. Pre-use visual appeal was strongly and positively correlated with post-use visual appeal (r=.726, p<.05), which suggests that participants did not change their opinions on visual appeal after having used the website. Post-use perceived usability was correlated with post-use visual appeal (r=.528, p<.01).

This suggests that even though the website was created and empirically tested to be easy to use, *participants judged it as hard because it was ugly, even after having used it.*

In the LuLv website conditions, seen in Table 7.4, there were again eight significant correlations. Pre- and post-use perceived usability were correlated (r=.533, p<.01). Pre-use perceived usability was also correlated with pre- (r=.539, p<.01). Post-use perceived usability was correlated with post-use visual appeal (r=.528, p>.01). Pre- and post-use visual appeal were correlated (r=.726, p<.01).

Table 7.4. Spearman Correlations for the HuLv website conditions.

	PreUsab	PostVis	PostUsab	Hovers	Clicks	Time	Pass
PreVis	.539**	.762**	.254	-.205	-.275	.002	.205
PreUsab	-	.608**	.533**	-.253	-.509**	-.031	.155
PostVis		-	.528**	-.230	-.360	.005	.082
PostUsab			-	-.299	-.264	-.138	.273
Hovers				-	.351	.506**	-.260
Clicks					-	.269	-.157
Time							-.754**

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

These results again suggest that *ratings did not change significantly after use* and that the *rating of pre-use usability was reliant on the website's visual appeal*. Moreover, even though the website *visual appeal and usability levels were not congruent, they were similarly judged*.

Table 7.5. Spearman Correlations for the LuHv website conditions.

	PreUsab	PostVis	PostUsab	Hovers	Clicks	Time	Pass
PreVis	.311	.417*	.083	-.015	.037	-.017	.084
PreUsab	-	.203	.229	-.043	-.030	-.093	.119
PostVis		-	.552**	-.318	.120	-.166	.145
PostUsab			-	-.288	-.041	-.307	.156
Hovers				-	.292	.635**	-.498**
Clicks					-	.420*	-.417*
Time							-.785**

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

Correlations in the Control Conditions

In this section, the results for the Spearman correlations between visual appeal and perceived usability within each control condition (HuLvNe and LuHvNe; n=10 in each) and are presented can be seen in Table 7.6. In the HuLvNe condition (i.e. easy but ugly with no expectations), pre- and post-use usability ($r=.661$, $p<.05$) were highly and positively correlated.

This suggests that their *opinions on usability did not change much with use*. Visual appeal and usability were neither significantly correlated in HuLvNe, nor in LuHvNe. This suggests that *participants' options on visual appeal changed after use, and that they were separately judging usability and visual appeal*.

Table 7.6. Correlations between visual appeal and perceived usability for HuLvNe and LuHvNe, respectively.

HuLvNe			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.532	.443	.273
PreUsab	-	.577	.661*
PostVis		-	.545

LuHvNe			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.290	.466	.254
PreUsab	-	-.092	.202
PostVis		-	.541

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

In HuLvNe, none of the objective usability measures were correlated. However, in LuHvNe, hovers were correlated with time per task (r=.891, p<.01) and passes (r=-.920, p<.01). The success rate (passes) and time were also correlated (-.963, p<.01).

These results confirm the use of the objective usability measures as they all seem to be measuring the same construct (i.e. objective usability).

Correlations when Expectations and Website levels are Congruent

All Spearman correlations between visual appeal and perceived usability within conditions where expectation levels were congruent with the actual website levels can be seen in Table 7.7. In the HuLvHuLv condition, pre- and post-use perceived usability were highly and positively correlated (r=.657, p<.05). Pre-use perceived usability was also correlated highly and positively with pre-use visual appeal (r=.827, p< .01). In addition, post-use perceived usability was correlated with post-use visual appeal (r=.835, p<.01).

These results suggest that participants *did not drastically change their opinions on usability after having used the website, and that usability judgements were largely based on the website's visual appeal prior to using it.* Moreover, participants did not change their opinions on visual appeal after having used the website (i.e. experiencing the usability did not affect the perception of visual appeal). **Even though the website was created and empirically tested to be easy to use, *participants judged it as hard because it was ugly, even after having used it*.**

Table 7.7. Correlations between visual appeal and perceived usability for HuLvHuLv and LuHvLuHv, respectively.

HuLvHuLv			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.827**	.904**	.870**
PreUsab	-	.848**	.657*
PostVis		-	.835**

LuHvLuHv			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.421	.302	.325
PreUsab	-	-.040	.073
PostVis		-	.888**

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

In the LuHvLuHv condition, only post-use visual appeal and post-use perceived usability (r=.888, p<.01) were highly and positively correlated. Thus, while participants seemed to have graded the usability and visual appeal differently before use, they seemed to think that they were very similar after use. This could be due to having lower opinions of usability before use, and then having the frustration of using the website lower the visual appeal of the website after use. This explanation was based on participant feedback.

In addition, for HuLvHuLv, post-use usability was correlated with passes (r=.652, p<.05), passes was also correlated with clicks, (r=-.656, p<.05), and clicks was correlated with time (r=.669, p<.05). In LuHvLuHv, hovers and clicks were correlated (r=.731, p<.05). Again, these results show that the objective usability measures seem to be in agreement and that post-use usability strongly reflects the usability level of the website, especially in the easier to use website.

Therefore, in the conditions where the expectation levels of visual appeal and usability were congruent with the website's visual appeal and usability levels, having an ugly website seems to lower the usability rating as "the colours distract from use", and having a pretty website does not affect usability ratings before use but both ratings drop after having used a hard website.

Correlations when Expectations and Website levels are Incongruent

All Spearman correlations between visual appeal and perceived usability in conditions where expectations of these were incongruent with the actual website levels can be seen in Table 7.8. In HuLvLuHv, pre- and post-use visual appeal were positively and significantly correlated (r=.896, p<.05). This suggests that their first impressions of visual appeal did not change after use. This was not the case in LuHvHuLv, where only pre-use visual appeal and pre-use usability were moderately and positively correlated (r=.641, p<.05).

This shows that pre-use, participants judged these two similarly even though the expectation given was different for both. The absence of other correlations is an indication that the two variables were being perceived and graded differently from each other and that initial opinions often changed after use.

Table 7.8. Correlations between visual appeal and perceived usability for HuLvLuHv and LuHvHuLv, respectively.

HuLvLuHv			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.329	.896**	.067
PreUsab	-	.494	.511
PostVis		-	.438

LuHvHuLv			
	PreUsab	**PostVis**	**PostUsab**
PreVis	.641*	.446	-.071
PreUsab	-	.537	-.189
PostVis		-	.319

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

In HuLvLuHv, time and passes were correlated (r=-.916**, p<.01), evidence that there were fewer passed tasks with longer times (as per the definition of passes). In LuHvHuLv, pre visual appeal was correlated with passes (r=.719, p<.05), clicks was correlated with time (r=.697, p<.05), and time was correlated with passes (r=-.755*, p<.05).

Correlations summary. Overall (n=60), visual appeal and perceived usability were positively and moderately correlated pre- and post-use, with the post-use correlation being slightly weaker. In HuHv (n=30), participants did not significantly change their opinions on usability after having used the website, and usability judgements were based on the website's visual appeal prior to using it. In addition, participants did not change their opinions on visual appeal after having used the website. This suggests that even though the website was created and empirically tested to be easy to use, participants judged it as hard because it was ugly, even after having used it. Similarly in LuLv (n=30), ratings did not change significantly after use and the rating of pre-use usability was reliant on the website's visual appeal. **Moreover, even though the website visual appeal and usability levels were not congruent, they were similarly judged.**

In the control condition of HuLvNe (n=10, easy but ugly with no expectations), participants' opinions on usability did not significantly change with use but visual appeal ratings changed after use, and participants were separately judging usability and visual appeal. Visual appeal and usability were not correlated in the other control condition, HuLvNe (n=10, hard but pretty with no expectations), meaning that they were perceived to be different from each other, and that these ratings changed after use.

In the congruent condition of HuLvHuLv (n=10), participants did not change their opinions on usability or visual appeal after having used the website, and usability judgements were largely based on the website's visual appeal prior to using it. Even though the website was created and empirically tested to be easy to use, participants judged it as hard because it was ugly, even after having used it. In the other congruent condition, LuHvLuHv (n=10), participants rated usability and visual appeal differently before use, but seemed to think that they were more similar after use. This could be due to having lower opinions of usability before use, and then having the frustration of using the website lower the visual appeal of the website after use.

In the incongruent condition of HuLvLuHv (n=10), first impressions of visual appeal did not change after use. This was not the case in LuHvHuLv, where participants judged visual appeal and usability similarly even though the expectation and empirical and objective levels were different for both. However, post-use, these were no longer correlated.

Discussion

A summary of the results is presented first. This is followed by a discussion of implications for theory. Next, a limitations and future research section are presented. The conclusion section is presented last.

Results Summary

Based on the feedback, participants were able to differentiate scores for usability and visual appeal and state which one of the two needed to be improved. However, they did mention that the worse variable lowered the rating of the better variable. The most common example of this was poor usability bringing down the rating for visual appeal, after use. In addition, higher expectations lead to slightly more positive feedback towards the corresponding variable, and negative expectations lead participants to be more hostile in their feedback but only towards the corresponding variable. Participant feedback had perhaps the strongest evidence that expectations were impacting usability and visual appeal judgements, supporting the hypotheses.

Based on the beanplots, the high visual appeal website and expectation levels increased ratings of both usability and visual appeal, especially pre-use. The highest rated usability levels came from the conditions in which usability was said to be highest, and likewise for visual appeal. The lowest ratings of these two variables came from the conditions in which expectations were low. Thus, expectations did seem to impact the perception of visual appeal and usability, more vividly pre-use, supporting the hypotheses. However, some of the differences seem to be marginal and statistics will have to be done to determine the significance of these findings.

One out of the eight statistical sub-hypotheses tested for visual appeal and perceived usability was significant. LuHvLuHv (x=3.58) was rated lower than LuHvNe (x=4.04) in pre-use usability. Therefore, there no support for the first hypothesis, that visual appeal is influenced by expectations. However, there is partial evidence for the second research hypothesis that pre-use usability is. No significant results were found for objective usability. Thus, the third research hypothesis is not supported, expectations did not seem to significantly impact use when visual appeal and usability were incongruently leveled.

Based on the correlations, in the easy but ugly website (no expectations), usability ratings were not affected with use, but visual appeal was, and participants separately judged these two variables. In the hard but pretty website (also no expectations), visual appeal and usability were rated differently from each other, both pre- and post-use. In the easy but ugly website with easy but ugly expectations, use did not impact usability and visual appeal ratings, and usability judgements were largely based on the website‘s visual appeal prior to using it. In the hard but pretty website, usability and visual appeal differed before use, visual appeal dropped and these two variables were rated similarly after use. In the easy but ugly website with hard but pretty expectations (i.e. the opposite), first impressions of visual appeal did not change after use. This was not the

case in the hard but pretty website with easy but ugly expectations, where participants judged visual appeal and usability similarly, but only pre-use.

Result Implications

Given the participant feedback, beanplots, statistical and correlational analysis, the following implications can be made with respect to the research questions.

Do expectations influence visual appeal? Evidence in the beanplots suggests that, yes, expectations do impact visual appeal. Specifically, low expectations lowered the rating of it, and high expectations raised it. This was evident pre- and post-use. However, this trend was not significant enough to be picked up in the statistical analysis. This could be due to the small sample size.

Are perceived and objective usability influenced by expectations? From the boxplot, some evidence exists to state that expectations impacted perceived usability. Specifically, the low expectation condition was rated evidently worse than the control condition. Statistical examination found that LuHvLuHv differed from LuHvNe in pre-use usability ($p<.05$). This means that the ugly website was significantly rated worse with low expectations, as compared to the control group.

What effect do verbal expectations have on usability and visual appeal? The answer is not automatically clear. Again, verbal expectations on their own were not examined. Instead, they were implemented in conjunction with textual expectations. Moreover, while the results of the previous experiment (Main Study 2) showed that verbal and textual expectations do influence these variables, these results occurred when the message about the expectation was either fully positive or fully negative. In addition, the website was either usable and pretty, or hard to use and ugly. Congruency thus allowed for an easier transmission of information with little to no confusion. In the case when expectations and the website had incongruent usability and visual appeal levels, then it seems that only low verbal and textual expectations of usability influence the perception of usability, pre-use. It may be the case that more work is needed to further answer this question.

What happens when visual appeal and usability levels are incongruent? An ugly website seemed to lower the usability rating as ~~"th~~e colours distract from use." A pretty website did not always affect usability ratings before use but both ratings drop after having used a hard-to-use website. Thus, the frustration of using a hard website lowers the visual appeal of the website, after use. Thus, an ugly website is terrible from the beginning, but a hard website will initially have good ratings, eventually being too annoying for the visual appeal to make a difference.

Implications for Theory

The findings in this study may still support the cognitive dissonance theory, in that participants all internalized the expectations and reacted to them differently, according to the four options stated by the theory. Thus, the absence of statistically

significant results does not automatically eliminate the possibility that the effect is still there because of the different responses available to participants upon dissonance. As previously mentioned, the cognitive dissonance theory states that one may either (1) add or (2) increase the importance of the information causing dissonance, or can (3) take away or (4) reduce the importance of the information causing dissonance, in order to reduce the dissonance. Evidence for the use of the fourth mechanism (i.e. the reduction of importance of the information causing dissonance) was highlighted by the participant feedback. Thus, while some participants may have agreed with the expectations, others may have gone with the other approach and disagreed with them. There is evidence of this in the beanplots, where some plots were slightly bimodal (Figures 7.3 and 7.4). Given these four options, the randomness of the results may make sense, since people reacted to the dissonance differently. It may be that the lack of significantly different results was due to the opposing expectations (high levels of one variable and low levels of the other) could have been confusing; this is further discussed in the next section.

Limitations and Future Studies

Threats to construct validity. There were no threats to construct validity as the scales and measures used in this study are all widely used and accepted. There were no problems with them in previous studies in this thesis.

Threats to statistical validity. The small sample size may have been the largest factor in the lack of significance in the statistical testing, as was the case in Main Studies 1 and 1.2. Again, due to time and monetary constraints, larger sample sizes were not possible for this thesis. Future studies should strive to acquire more participants or perhaps automate the testing process so that participants could do the test online, individually, and at their own convenience.

Threats to internal validity. One possible issue with the implementation of expectations, as was found before, would be unfamiliarity of the location and experimenter, which could have influenced expectation trustworthiness, lowering its internal value. Future studies should strive to include peer pairing of the confederate to increase their influence over the participant.

Another possibility is that having two different expectations (i.e. one that is high and one that is low) was a bit confusing for participants. However, based on their feedback, they were able to differentiate between visual appeal and usability, suggesting that two pieces of information were not hard to follow or understand. Future studies should strengthen the wording from the confederate to see if that would have a greater impact on participants.

Threats to external validity. As was the case with the two previous studies in this thesis, there were no pretests to indicate the possibility of a reaction or interaction effect during testing. Each participant was given one treatment, so multiple treatment

interference was not a concern in this thesis. Participant recruitment and selection was random and participants were only screened for eye-sight and colour-blindness, given that the colours and images in the website are important in order to ascertain the appropriate visual appeal level. Having met the 20/20 vision requirement, participant assignment to conditions was randomly chosen to eliminate the possibility of selection biases. However, many participants studied computer science (38/60), with varying degrees of English fluency, and of different cultural backgrounds. These factors were not controlled for and may have skewed or randomized the results as they influence the perception of visual appeal and could have added some unaccounted difficulties in usability. Although, HCI studies are predominantly held in the information technology departments of universities and their participant demographics are similar to this study‘s. Thus, the results of the condition outcomes in Main Study 3 are comparable and as generalizable as any other study.

Summary

Participant feedback and the beanplots showed the strongest evidence that expectations were impacting usability and visual appeal judgements, supporting the hypotheses. Based on the feedback, higher expectations lead to slightly more positive feedback towards the corresponding variable, and negative expectations lead participants to be more hostile in their feedback but only towards the corresponding variable. Based on the beanplots, some of the highest rated visual appeal and usability ratings came from the corresponding high expectation conditions, and the lowest rating came from the corresponding low expectations. However, these trends were not captured in the statistical analysis. Only LuHvLuHv was rated lower than LuHvNe in pre-use usability. Therefore, there does seem to be some evidence to support the hypotheses that expectations do influence the perception of visual appeal and usability, especially pre-use but these patterns were not found in the statistical tests. The results can be explained by the cognitive dissonance theory. If participants reacted to the expectations differently (i.e. some ignored while others embraced them) then the lack of significant results makes as much sense as the lack of impact from the expectations. It may be that the lack of significantly different results was due to the opposing expectations (high levels of one variable and low levels of the other) could have been confusing. Future studies should perhaps examine larger sample sizes, and stronger wording from the confederate, peer pairing of confederate to increase their influence over the subject.

The next chapter is the General Discussion chapter. It first summarizes all of the findings across this thesis, study by study. Then, given all of the findings, implications for website design are given. This is followed by the theoretical implications. Finally, the limitations are summarized. The final chapter is the conclusion. This is followed by the references and appendices.

Chapter 8. Discussion

This thesis has examined the effect of expectations on usability and visual appeal using a set of controlled experiments, in a website genre where participants did not have highly developed mental models and the website was gender and age neutral. A less developed mental model was required so that expectations could more readily be experimentally controlled, without the possible influence of prior experience. The ecommerce market would benefit from a positive user response for their websites. In particular, this research would be particularly important for government websites, where public opinion is not always positive. Government city websites were thus chosen since they contain neutral information. For example, pet registration, garbage days, and city pictures and attractions are all fairly age and gender neutral topics.

In order to observe the impact of expectation on usability and visual appeal, we manipulated expectations to create cognitive dissonance. We had four research hypotheses based on this theory. Cognitive dissonance is a disagreement of information causing stress, and people strive to reduce the stress by changing the way they think about the issue. The cognition that is most resistant to change is most likely the most recent behaviour (Harmon-Jones et al., 2009). In this thesis, the most recent behaviour was the experience of the expectation. Therefore, if expectations influence visual appeal and usability, then participants should agree to the expectation given, affecting the perception and rating of these variables accordingly. Therefore the six research questions were:

(1) *What happens when visual appeal and usability levels are incongruent?*
(2) *What happens when visual appeal and usability levels are congruent?*
(3) *What effect do verbal expectations have on usability and visual appeal?*
(4) *What effect do textual expectations have on visual appeal and usability*?
(5) *Are perceived and objective usability influenced by expectations?*
(6) *Do expectations influence visual appeal?*

A review of the preliminary studies results is presented first. Then, the summary of the main findings in the Main Studies 1, 2, and 3 are presented respectively. This is followed by a discussion on the implications of the results on the research hypotheses. Subsequently, a discussion on the implications for theory is presented. Next, the limitations are given.

Preliminary Studies Review

The purpose of the preliminary studies was to select a suitable stimulus website. To control for expectations, we needed a website domain that had less developed mental models, to exclude the influence of past experiences. A website genre was required that was age and gender neutral. Hence, to meet this requirement a city-themed website – namely city councils and city tourism websites were examined.

The main results of the five preliminary studies showed that participants were less familiar (i.e. less developed mental models) with city council websites than with city tourism websites. Moreover, participants rated the city council genre more negatively in visual appeal, stress, etc. Yet, they rated the Gold Coast city council website as the overall prettiest website. This underestimation further suggested that their mental models were less developed for city council websites. Therefore, the Gold Coast city council website was chosen for further testing.

Since the Gold Coast city council website was chosen primarily on the basis of its high visual appeal results, the usability needed to be verified. This was done via user and expert based usability testing. Both resulted in relatively high usability, considering the inexperience with the website genre.

Next, the original Gold Coast website needed to be manipulated so that there would be four versions of it that varied in high and low levels of visual appeal and usability. One website was used in order to eliminate unnecessary confounding variables of using different city council websites varying in usability and visual appeal, which might include biases for preference towards different cities and not others. Also, it would have altered the positioning of items on the page by default, possibly altering the complexity of the tasks, making some tasks relevant and others not. Moreover, it would have changed aspects of usability that might have interfered with visual appeal as well – making it hard to have independent levels the two variables. For example, different fonts could have impacted the visual appeal and usability between websites, since some are prettier but harder to read (e.g. Blackadder ITC). To account for the other unforeseen factors that could influence visual appeal and usability, one website was chosen for the studies. This website was manipulated to produce four versions that varied in usability and visual appeal in a highly controlled manner. The manipulations were also tested with users. The low usability manipulation was not low enough the first time around, so it was re-manipulated and re-tested to be significantly harder to use than the original website.

It took more manipulations to make usability worse than it did to make visual appeal worse. However, the preliminary studies successfully resulted in a set of four fully functional websites that differed only in visual appeal and usability. The four website versions were: high in both of usability and visual appeal (HuHv also referred to as easy and pretty), low in both (LuLv, also referred to as hard and ugly), high in usability and low in visual appeal (HuLv, also referred to as easy and ugly), and low in usability and high in visual appeal (LuHv, also referred to as hard and ugly).

Main Study 1 Discussion of Results

The purpose of this study was to see if expectations influence the visual appeal, perceived and objective usability. To test this, the easy and pretty, and hard and ugly versions of the website were used, with three different levels of expectations: high expectation of visual appeal and usability, low in both, and no expectations (control

condition). Thus, there were six conditions in this phase: (1) easy/pretty website with easy/pretty expectations, HuHvHe, (2) easy/pretty website with hard/ugly expectations, HuHvLe, (3) easy/pretty website without expectations, HuHvNe, (4) hard/ugly website with easy/pretty expectations, LuLvHe, (5) hard/ugly website with hard/ugly expectations, LuLvLe, and (6) hard/ugly website with no expectations, LuLvNe. These expectations were chosen because they were either completely congruent or completely incongruent with the website usability and visual appeal levels. For example, the HuHv website would have congruent (He), incongruent (Le), or no expectations (Ne). This allowed for a more focused examination of the impact of expectations, controlling for the impact usability and visual appeal have on each other. Given the work done on textual user reviews and their influence on trust and their impact of prospective buyers (e.g. Gefen et al., 2003; Smith et al., 2005) outlined in Chapter 2, the expectations were implemented textually only, merged into the task descriptions since it was thought to be enough to influence participants. This was changed for Main Study 2, discussed in the next section. To examine the research hypotheses above, causational and correlational statistics were computed, along with brief qualitative analysis.

The participant feedback summary is given first. It was done in order to gain an understanding of what participants were thinking and to see if expectations were strong enough to influence their opinions of the website. This is followed by the beanplot summary, which was done to examine the general data trends. Statistics were then applied to the data to examine which conditions differed from each other. These included the causational and correlational statistical result summaries.

Study 1: Participant Feedback Discussion

Equal numbers of participants across the three conditions in the hard and ugly website complained about the usability. For the easy and pretty websites, the high expectations condition complained the most about the usability with the control group complaining the least. This suggests that the high expectations group actually had the highest expectations, and the subjects were disappointed that the site was not in fact miraculously easy to use. The greatest number of usability praises came from the control conditions, which had no expectations and were thus the most objective in their assessments of the usability, particularly in the easy and pretty website.

For visual appeal, participants in the LuLvLe complained the most, suggesting that they had the highest expectations, and were thus the most disappointed by the visual appeal. The control condition received the most positive feedback for both visual appeal and usability. This suggests that the written form of expectations was impacting users, but the significance of the impact needed to be examined quantitatively, which was done both using graphs and statistics.

Study 1: Beanplot Summary

The beanplots showed slight variations in the means across the hard/ugly website conditions where the control condition was rated highest pre- and post-use visual appeal and pre-use perceived usability, with the low expectations condition rated the lowest. These results coincide with participants‘ feedback, and show evidence that written expectations were impacting the experience of the website as well.

Post-use, the trend in perceived usability changed slightly, with participants in the high expectations group rating it as easiest and low expectations group rating it as hardest. Therefore, a small trend did emerge post-use, supporting the first two hypotheses. Further analysis was necessary in order to ascertain the statistical importance of these findings.

Study 1: Statistical Analysis Summary

Given that the assumptions testing revealed that some variables were non-normal and had heterogeneous variance, that the sample size per condition was relatively small (n=10), and that some variables were discrete while others where binary, the statistics applied were non-parametric. Chiefly, Kruskal-Wallis and Fisher‘s Exact tests were applied where appropriate.

Tuch and colleagues (2012) found that affective experiences with a website‘s usability mediate visual appeal. However, the results in this study found that mood was not related to visual appeal or usability. No evidence was found to suggest that mood was a factor and it was thus not measured in future studies in this thesis.

The results showed that the hard/ugly website with low expectations was significantly rated lower in pre- and post-use visual appeal ratings, compared to the control pre- and post-use visual appeal ratings, respectively. In other words, participants in the hard/ugly website with low expectations condition (LuLvLe) perceived it to be uglier than the control group, irrespective of use. Also, participants in the easy/pretty website with high expectations interacted with the website more, since the number of clicks per task was one higher than it was for the same website with low expectations. This suggests that participants used the website less when they were told that it was hard to use, seemingly uninterested or slightly deterred from using it. Therefore, the statistics suggest that written expectations had an impact on both the perceptions and actions of people using the websites. However, this impact was not as large or widespread as anticipated. The next step was to examine correlations see if any other relations existed between the variables.

Study 1: Correlational Analysis Summary

In HuHvHe, usability and visual appeal were correlated pre- and post-use, along with pre- and post-use usability. In LuLvLe, pre- and post-use usability were correlated. HuHvNe showed relationships between pre- and post-use for visual appeal, and pre- and

post-use usability. The LuLvNe condition had significant correlations between pre-use visual appeal and usability, and with pre- and post-use visual appeal. Both the HuHvHe and LuLvNe conditions showed that pre-use usability was correlated with post-use visual appeal. This would suggest that the impression pre-use usability participants managed to influence their post-use ratings of visual appeal, but no causality can be established with correlations. No correlations were found in conditions where the expectation was incongruent with the website's actual visual appeal and usability levels. While causality cannot be inferred from correlations, evidence exists to answer the research question, that indeed, expectations do impact visual appeal and usability. Mainly, the control conditions seem to be behaving the same way as many studies in the literature (e.g. Tractinsky et al., 2012). The conditions in which expectations and website visual appeal and usability levels are congruent also obtain similar results to the literature. However, when the expectations and website levels are incongruent (i.e. when there is dissonance), correlations disappear. This discussion will be brought up in the implications to theory section below.

Main Study 2 Discussion of Results

In the previous study, textual expectations did affect visual appeal ratings and the number of clicks per task. Previous work suggests that the use of confederates (e.g. Asch, 1956; Milgram, 1963) showed that verbal instruction such as WOM can largely influence product reputation (e.g. Ellison & Fudernberg, 1995). Therefore, the purpose of this study was to see if adding a confederate to verbally reinforce the textual expectations would strengthen the implementation of expectations, resulting in a greater impact on usage and user perception of visual appeal and usability. This study repeated the three conditions (He, Le, and Ne) from the easy/pretty (HuHv) website.

First, we discuss the participant feedback summary to gain insight into what participants were thinking and if the expectations impacted their opinions. This is followed by the beanplot summary so as to illustrate the visual appeal and usability scale response spreads. Then, the summary of the statistical results is presented next. These include the causational and correlational statistical result summaries.

Study 2: Participant Feedback Summary

Contrary to the previous study, the feedback was not mixed in this study. While the website presented to both groups of participants was the same, the majority of participants in the high usability and visual appeal condition (He) thought that the website was pretty, whereas the majority of participants in the Le condition thought the opposite, that it was ugly. Participants in the low expectations condition were much more critical of the website's usability and visual appeal levels, with none of them mentioning that the easy/pretty site was actually easy to use. These findings strongly supported the first two research hypothesis.

Study 2: Beanplot Summary

Beanplot analysis was undertaken to examine the general spread and behaviour of the data. It was also done to see if people's opinions supported their ratings, since retrospective-interview and self-report can be hard to rely on.

Pre-use visual appeal was perceived to be slightly prettier in the He condition than in the control condition, which was perceived to be slightly prettier than the Le condition. Post-use, the participants in the control condition lowered their ratings of visual appeal while the ratings stayed identical to what they were pre-use in the He condition. This strongly suggests that the high expectation affected perceptions after use, since use (along with boredom and other factors that can interfere) did not lower the results, as it did in the control group. Participants in the Le condition seem to have somewhat split opinions with some ratings staying where they were pre-use while others found it to be quite a bit uglier after they interacted with it. Thus, the verbal reinforcement of the textual implementation of expectations affected participants' perceptions of visual appeal pre- and post-use.

For usability, the He condition was rated as the easiest to use and the Le condition rated as the hardest, both pre- and post-use. The effect of use was more pronounced with perceived usability than with visual appeal because visual appeal was not affected post-use in the He condition. Pre-use perceived usability ratings were all slightly higher than the post-use ratings. This suggests that use itself influenced the perceived usability ratings of the website. Between conditions, He was perceived to be the easiest to use and that Le was perceived to be the hardest. The verbal implementation of expectations via confederate did affect usability perceptions, and the impact lasted post-use as well. Therefore, the first two hypotheses were supported by the findings in the beanplots. To verify the significance of the findings, statistical analysis was applied next.

Study 2: Statistical Analysis Summary

As in the previous study, the assumptions testing revealed that normality and homogeneity of variance were not unilaterally met. In addition, and again similar to the previous study, the sample size per condition was small and there were discrete and binary variables. Therefore, the statistics applied were non-parametric: Kruskal-Wallis and Fishers Exact tests.

The results showed that HuHvHe and HuHvLe (i.e. the easy/pretty website but with high and low expectations) differed in pre-use usability, post-use visual appeal, and post-use usability. This means that the same website was differently rated, depending on what the confederate and tasks descriptions said before the experiment. Specifically, participants rated the website better when they were told it was going to be easy and pretty, and they rated it as worse when they were told the opposite.

For objective usability, the average number of clicks per task, average task completion time, and proportion of passes (success rate) differed between the HuHvHe and HuHvLe conditions. More precisely, the low expectations condition received more

clicks, took nearly double the time, and had a lower success rate than the high expectations condition, doing the same tasks and using the same website. Therefore, there is substantial evidence to support all three research hypotheses. Participants rated the same website as prettier and easier to use when they were told that it was going to be well made, pretty, and usable. Moreover, they struggled more with the website when completing the information retrieval tasks when told that the website was hard to use. Therefore, verbally reinforced textual expectations influence both how participants viewed and interacted with the website, when visual appeal and usability levels are congruent.

Study 2: Correlational Analysis Summary

Across all three conditions, pre- and post-use correlations existed between visual appeal and perceived usability. These results are congruent with the general results of Main Study 1, as well as with the literature. Only one other correlation was significant: Pre-use visual appeal was highly and positively correlated with post-use visual appeal, only in the HuHvHe condition. This means that participants did not significantly change their minds or attitudes after use about visual appeal, in the high expectations condition. These results strongly support the cognitive dissonance theory, which is explained in the implications to theory section below.

Main Study 3 Discussion of Results

The previous two studies worked with website versions and expectation levels that had congruent visual appeal and usability levels. In this study, the purpose was to extend this research by examining the influence of expectations on visual appeal and usability when they are incongruent. Thus, the easy/ugly (HuLv) and hard/pretty (LuHv) versions of the website were used. The three expectation conditions were: high usability and low visual appeal (HuLv), low usability and high visual appeal (LuHv), and no expectations (Ne) which was the control condition. This created a two by three design with six conditions: (1) easy but ugly website with congruent expectations, HuLvHuLv, (2) easy but ugly website with hard but pretty expectations, HuLvLuHv, (3) easy but ugly website with no expectations, HuLvNe, (4) hard but pretty website with easy but ugly expectations, LuHvHuLv, (5) hard but pretty website with hard but pretty expectations, LuHvLuHv, and (6) hard and ugly website with no expectations, LuHvNe.

The same outline structure and reasoning used in the previous two studies was applied here as well. Participant feedback is discussed first, followed by the beanplot summary. The quantitative analysis discussion is presented last.

Study 3: Participant Feedback Summary

Participants were able to differentiate between usability and visual appeal and state which one of the two needed to be improved. However, they did mention that the worse variable slightly lowered the value of the better variable. For example, poor usability brought down the rating for visual appeal as well, due to frustration. In addition, higher expectations lead to slightly more positive feedback towards the corresponding variable, and negative expectations allowed participants to be more hostile in their feedback but only towards the corresponding variable, leading towards support for the hypotheses. Thus, while the actual variable level obviously would impact the feedback, evidence suggests that the expectations did as well. Participant feedback offered strong evidence that incongruent expectations were impacting usability and visual appeal judgements, supporting the research hypotheses.

Study 3: Beanplot Summary

High visual appeal in both the website (i.e. actual) and expectations (i.e. experimental) increased ratings of usability and visual appeal, most notably pre-use. The highest rated usability levels came from the conditions in which usability was said to be highest. The lowest ratings of these two variables came from the conditions in which expectations were low. Thus, expectations did seem to impact the perception of visual appeal and usability, more vividly pre-use, supporting the hypotheses.

Study 3: Statistical Analysis Summary

For the same reasons as in the previous two studies, non-parametric statistics were applied here as well. Partial statistical evidence exists for the second hypothesis: perceived usability was rated lower when expectations were set to be low, especially between the control group and the low-expectation group, which rated the website as harder to use. This difference was only found pre-use, suggesting that the impact of the expectation was only strong enough to influence ratings after use.

These findings suggest that lower expectations were more effective and persuading participants than high expectations. This is in contrast to Kamins el al.‘s (1997) results in work they did that examined marketplace rumour transmission. In general, Kamins el al., (1997) found that communications that were labeled as _umours‘ were deemed less credible than regular WOM communications, and were less likely to be re-told to others. Conversely, positive rumours were re-told since the messenger was also received with more liking. However, there was a personal factor that influenced relaying rumours. Not surprisingly, people liked spreading positive rumours about themselves more so than negative ones. They also preferred to spread bad rumours about disliked individuals. For neutral parties, both good and bad rumours were equally spread (Kamins el al., 1997). While the _umours‘ in this thesis‘ study were controlled and spread textually and verbally, their impact seemed to be greater

when they were negative, which is in contrast to Kamins el al., (1997)'s findings that negative rumours were less credible than positive ones.

No significant results were found for objective usability. This finding, or lack of statistical significance, suggests that low visual appeal created just as much difficulty as low usability in completing the tasks. In other words, both low visual appeal and low usability equally obstructed ease of use. Expectations did not seem to significantly impact use. Thus, there is insufficient statistical evidence to conclude that the third research hypothesis is true, since there is insufficient evidence to state that participants struggled more when expectations were low or that they did better when expectations were high. Thus, the third research hypothesis is not supported; expectations did not seem to significantly impact use.

Study 3: Correlational Analysis Summary

Across all conditions, visual appeal and perceived usability were positively and moderately correlated pre- and post-use, with the post-use correlation being slightly weaker. In the three HuLv conditions, participants did not significantly change their opinions on usability or visual appeal after having used the website, and usability judgements may have been swayed by the website's visual appeal prior to using it. This suggests that even though the website was created to be easy to use, participants judged it as hard because it was ugly, even after experiencing it. Similarly in all three of the LuHv conditions, ratings did not significantly change after use and the rating of pre-use usability was reliant on the website's visual appeal. Thus, even though the website visual appeal and usability levels were not congruent, they were similarly judged.

In the easy but ugly website control condition, usability ratings were not affected with use, but visual appeal was, and participants separately judged these two variables. In the hard but pretty website control condition, visual appeal and usability were rated differently from each other, both pre- and post-use (i.e. not correlated).

In the easy but ugly website with congruent expectations, use did not impact usability and visual appeal ratings, and perceived usability was based on the website's visual appeal, pre- and post-use. Even though the website was created and empirically tested to be easy to use, participants judged it as hard because it was ugly, even after having used it. In the hard but pretty website with congruent expectations, usability and visual appeal differed before use, but visual appeal was rated lower after use. Thus, these two variables were rated similarly after use. This could be due to having lower opinions of usability before use, and then having the frustration of using the website lower the visual appeal of the website after use (supported by participant feedback).

In the easy but ugly website with hard but pretty expectations (i.e. opposite/incongruent), first impressions of visual appeal did not change after use. This was not the case in the hard but pretty website with easy but ugly expectations, where participants judged visual appeal and usability similarly, but only pre-use.

Implications of Findings

Based on the main studies‘ results, the following conclusions can be made for the research questions. The first research question was: *Do expectations influence visual appeal?*

The answer is a clear *yes*. There was evidence of this across all three main studies, but particularly in the first two. In Main Study 1, the hard/ugly website with low expectations was rated significantly lower in pre- and post-use visual appeal ratings, compared to the control pre- and post-use visual appeal ratings, respectively. In Main Study 2, the high and low expectation conditions in the HuHe website differed in post-use visual. The same website was differently rated, depending on the expectation.

The second research question was: *Are perceived and objective usability influenced by expectations?*

Again, the answer is *yes*. In Main Study 1, participants in the easy/pretty website with low expectations clicked through the website less when they were told that it was hard to use, seemingly deterred from using it. In Main Study 2, HuHvHe and HuHvLe differed in pre- and post-use usability, where participants rated the website better when they were told it was going to be easy and pretty, and they rated it as worse when they were told the opposite. For objective usability, the low expectations condition made more clicks, took nearly double the time, and had a lower success rate than the high expectations condition, doing the same tasks and using the same website. In Main Study 3, pre-use perceived usability was rated lower when expectations for usability were set to be low as compared to the control group. Therefore, evidence exists to support the hypothesis that, yes, perceived and objective usability both are impacted by expectations.

The third research question was: *What effect do textual expectations have on visual appeal and usability*?

In order to address this question, we first need to examine *if* textual expectations were strong enough to influence usability and/or visual appeal. Perceived usability was not affected by textual expectations. However, from Main Study 1, we can conclude that negative (i.e. low expectations in both visual appeal and usability; Le) textual expectations do have an effect on pre- and post-use visual appeal. Participants in the low textual expectation condition with the hard/ugly website actually thought that the website was uglier than the control group, and this opinion lasted after having experienced the website. Positive textual expectations did not seem to significantly influence perceptions of either visual appeal or usability. Textual expectations also impacted the number of clicks (i.e. objective usability; performance measure) made in the website. Specifically, low textual expectations deterred people from exploring the website, as if they were not interested in clicking through the interface or were deterred by the negative reference in the task description. Therefore, it seems that evidence does exist to support the claim that textual expectations do impact objective usability and visual appeal. It would also suggest that positive texts do not have the same impact as

negative textual expectations – negative ones significantly affecting participants' perceptions and system interaction.

The fourth research question was: *What effect do verbal expectations have on usability and visual appeal?*

Verbal expectations alone were not tested. Only verbally reinforced textual expectations were, thus we can only comment on the combined effect. Based on the results of Main Studies 2 and 3, there is compaelling evidence that, *yes*; both visual appeal and usability (perceived and objective) are affected by the combination of verbally and textually implemented expectations. In Main 2, for the easy/pretty website with high and low expectations, pre- and post-use usability, and post-use visual appeal were affected. The combined expectations affected participants' perceptions more so than textual expectations alone, rating the website better when they were told it was going to be easy and pretty, and they rated it as worse when they were told the opposite. The effect lasted after use as well. In addition, the combined expectations also affected objective usability, where participants in Le made more clicks, took nearly double the time, and had a lower success rate than those in He, doing the same tasks and using the same website. This means that the same website was differently rated, depending on what the confederate and task descriptions said before the experiment. However, this effect was more evident when usability and visual appeal levels were congruent because in Main Study 3, where usability and visual appeal were at incongruent levels, there are less significant results. Mainly, perceived usability was rated lower when expectations were set to be low, only found pre-use, suggesting that the impact of the expectation was only strong enough to influence ratings after use. Therefore, textual and verbal expectations influenced usability and visual appeal, more so in the congruent conditions than in the incongruent ones.

The fifth research question was: *What happens when visual appeal and usability levels are congruent?*

To answer this question, we refer to the results in Main Studies 1 and 2. In these two studies, visual appeal and usability were both either high or low (i.e. congruent; HuHv and LuLv). In both studies, there was evidence to conclude that expectations influence both how participants viewed and interacted with the website. This may have occurred because, here, visual appeal and usability complemented each other (i.e. were the same level), which meant that there was no cognitive dissonance within the expectation or the website (just between the expectation and the website in two conditions). This simplified the understanding of the expectation, strengthening the impact of the expectation and its effect on experiencing the website. The website was either entirely good or entirely bad, with good, bad, or no expectations. This also meant that there were no confusing conflicts.

The sixth research question was: *What happens when visual appeal and usability levels are incongruent?*

The answer to this question lies in Main Study 3 where one of the two was high and the other was low (i.e. LuHv and HuLv). Given the participant feedback, beanplots, statistical and correlational analysis, the following implications can be made. An ugly

website seemed to lower the usability rating, and evidence of this existed in the participant feedback since a repeating sentiment was that ~~–e~~olours distract from use." A pretty website did not always affect usability ratings before use but both ratings drop after having used a hard-to-use website. Thus, the frustration of using a hard website lowers the visual appeal of the website, after use. Overall, an ugly website is terrible from the beginning, but a hard-to-use website will initially have good ratings, but eventually will be too annoying for the visual appeal to make a difference. The lack of significance across many of the statistical tests, nevertheless, suggests that upon hearing and experiencing the websites, people were reacting differently. Had the expectation impacted everyone equally, there may have been significant results (i.e. clear differences between groups with different expectations). Thus, expectations may be impacting participants but the reactions to the expectations may not be so predictable. More on this topic is presented in the theoretical implications section below.

Summary. Evidence from participant feedback, the general graph behaviour, statistical and correlational calculations all point to the conclusion that expectations do impact visual appeal, objective and subjective usability. Visual appeal and perceived usability differed between conditions in the same website, pre- and post-use, based on a pre-set expectation. Furthermore, the addition of the verbal implementation of expectations via confederate further impacted how participants interacted with the website more, struggling more with it when they were told it was going to be hard. These results suggest that expectations do impact the perception and use of a website.

Implications for Website Design

According to Taebi, Aldabbas, and Clarkson (2013), visual appeal and usability are both important, but they play different roles in website perception. Visual appeal impacts the first impression, fostering the initial attraction to a website, whereas usability becomes more important with use (Taebi et al., 2013). Both visual appeal and usability are thus equally important and both need to be improved to enrich user experience (Taebi et al., 2013). Without high visual appeal, users would reject using the website right from the get-go, and usability maintains their loyalties. However, both of these variables are impacted by expectations, as shown by the results in this thesis. High expectations have positive effects on users, while negative expectations can lower trust and even price mark-ups in an online market environment (e.g. Cabral & Hortacsu, 2006; Houser & Wooders, 2006).

For website design, unfortunately this means that how well a website is made is not the only factor that influences what people think about it. As demonstrated in this research through the use textual and verbal messages, a bad reputation can turn people against your website, even if the reputation is not true. However, making a usable and pretty website in conjunction with having a good reputation is what website owners should strive to achieve, since this combination yields the best ratings.

This impact might also happen in social networking and online reviews. Amazon is currently suing over 1000 people over false reviews left on their website. Users were able to write reviews on products in exchange for $5 per review. This was misused for personal gain and to falsely promote companies. The results in this thesis suggest that Amazon may be right – user reviews (i.e. textual descriptions) do indeed influence people‘s opinions of a product.

The impact of these results extends to other areas of research that use websites and possibly even other technologies. For example, someone who is reluctant to use technology (such as an elderly person) may need to use an app for medical reasons such as tweaking a pace maker. The reputation of the technology can encourage the individual to use and even like the product more, making the possibly steep learning curve slightly more enjoyable. More research would need to be done to examine the impact of expectations with other technologies.

To overcome this, one should invest in marketing to give a website a more positive reputation right from the beginning. Good marketing offers the opportunity to influence people before they use a website and, according to the results of this study, the influence may be strong enough to outweigh the impact of the actual website. In this study, participants were forced to use a specific site, whereas in real life there are thousands of websites to choose from and competition can be fierce. If you advertise, people will (1) know about it, (2) know something *good* about it, (3) be willing to check your website out, and (4) like it more before and after they use it. However, the positive impact of marketing can only last after use if the website is truly well designed. Thus, spending time and money on the development of a visually appealing and easy to use website may be just as important as advertising (be it print, radio, TV, or online).

Overview of Theoretical Implications

As mentioned in Chapter 3, *cognitive dissonance* is the stress caused to the individuals who experience contradictions in their understanding of information, beliefs, or values. The results of Main Studies 1 and 2 both directly support the cognitive dissonance theory, in that participants clearly agreed with most recent information (i.e. the expectation; Harmon-Jones et al., 2009). Expectations influenced participants‘ perceptions of visual appeal and usability, so much so that participant interaction with the websites also changed as a result of the expectation. These results were strengthened when verbal expectations were added in addition to the textual ones, and when the visual appeal and usability levels were congruent. This may have occurred because the congruence made the message simple and straightforward which in turn made it easier to understand and set the expectation. Thus, the dissonance was more apparent when the expectation was either completely positive or completely negative, and then the website either agreed or disagreed with the expectation. Therefore, given the results, participants were influenced by the expectation, rating the website accordingly. The cognitive dissonance theory readily explains this phenomenon since participants agreed with the most recent information.

While it is relatively easy to predict an individual‘s actions when there is no disagreement, it is a lot harder to do so when there is dissonance. There were very few significant findings in Main Study 3, with the correlation results appearing to be random. While not immediately apparent, the findings in Study 3 support the cognitive dissonance theory as well. The participants all internalized the expectations and experienced dissonance. In the presence of this dissonance and with conflicting expectations, participants all reacted differently. As previously mentioned, when dissonance occurs an individual strives to achieve consonance by reducing the inconsistency. Reducing the inconsistency can be done with at least one of four ways: (1) add or (2) increase the importance of the information causing dissonance, or can (3) take away or (4) reduce the importance of the information causing dissonance, in order to reduce the dissonance. Given that there are four known options and no exclusive expectation of response, these four options, the randomness of the results makes sense, since people reacted to the dissonance differently. Dissonance reduction is measurable by noting attitude change, usually in the direction of the cognition most resistant to change (Harmon-Jones et al., 2009). A person‘s attitudes are most likely to change to concur with that person‘s most recent actions, so as to avoid further dissonance. However, when the message is mixed, it is quite likely that the reaction to the expectation depends on the participant and how they manage the dissonant information. In the case of Main Study 3, there was no evidence of a single pattern that everyone applied.

Summary of Limitations

Threat to construct validity: the mood scale. The first main study had a problem with instrumentation: The mood scale, SAMS, that was employed presented ambiguities making it unreliable ambiguous, creating opportunities for misunderstandings. The subjective interpretation of the images in the SAMS scale and the lack of main effects with mood indicated that there was no further need to examine mood in this thesis. There were no other threats to construct validity in this thesis.

Threat to statistical validity: sample size per condition. Potentially, there was one limiting factor to the analysis throughout the thesis. A relatively small sample size (ten participants per condition) may have been insufficient to capture some significant differences that a larger sample could have shown. Approximately one hour was spent with each participant, which allowed a substantial amount of time for a thorough examination of their opinions and actions. It was difficult to schedule and re-schedule participants and the two confederates, who were full-time PhD students. Participants were often late or did not show up, making it hard to finish data collection on time. This also influenced the confederates‘ schedules, who tried to be as flexible as possible. Funding also restricted this decision as each participant was thanked for their time with a $20 gift certificate, and the two confederates were paid for their time as well.

Four threats to internal validity. (1) Expectation Formation: Another possible limitation for the first study in particular was the assumption that expectations can be formed in a matter of seconds, by reading a short task description, in an unfamiliar physical environment (i.e. ecological validity for the formation of expectations). It may be that expectations are formed over time.

(2) Location: Unfamiliarity of the location and experimenter could have also influenced trustworthiness of the expectation, lowering its impact.

(3) Expectations: In Main Study 1, expectations were set by participants reading a task description that outwardly stated what they were intended to believe. In the following two studies, the expectations were implemented both textually and verbally. It may have been more effective to give the expectation less overtly. For Main Study 3, another possibility is that having two different expectations (i.e. one that is high and one that is low) was a bit confusing for participants. However, based on their feedback, they were able to differentiate between visual appeal and usability, suggesting that two pieces of information were not hard to follow or understand.

(4) Confederate Familiarity: Given that significant differences were found even with these limitations, it is hypothesized that the results would be more compelling had a familiar face (e.g. friend, family, teacher or tutor) been the one giving them the expectation, in a familiar environment (e.g. at home, in a classroom, on the phone, or even in a message on Facebook or via email).

Threats to external validity. There were no pre-tests to indicate the possibility of a reaction or interaction effect during testing. Each participant was given one treatment, so multiple treatment interference was not a concern in this thesis. Participant recruitment and selection was random and participants were only screened for eye-sight and colour-blindness, given that the colours and images in the website are important in order to ascertain the appropriate visual appeal level. Having met the 20/20 vision requirement, participants were randomly assigned to conditions to eliminate the possibility of selection biases. Also, the use of a confederate is a widely acceptable method in experimental studies. Thus, the results of this study are generalizable.

Summary

This chapter summarized the main findings from each of the studies. These started with the review of the preliminary studies. Then, the summary of the main findings in the Main Studies 1, 2, and 3 were presented respectively. A general discussion on the implications of the results on the research questions was examined next. Subsequently, a discussion on the implications for theory was presented. This chapter concluded with the general limitations. The next chapter is the Conclusion chapter. It examines this thesis‘ contributions and suggests future work that should be examined to further the understanding of the impact of expectations on usability and visual appeal.

Chapter 9. Conclusion

There are many business and service providers who would benefit from consumers having a positive attitude towards the visual appeal and usability of their websites. This is certainly the case for service websites. This is also the case with government websites where public opinion is often not generally favourable, yet many of their services are easily attainable through online interactions rather than face-to-face.

Most research to date has focused on ecommerce websites, even though other types (e.g. government, corporate, personal blogs, etc.) are just as commonly found online, and users still have high expectations of them as well (Burtuskova & Krejcar, 2013). The value in terms of effectiveness and efficiency to both the website owner and the user is substantial. With textual user reviews and verbal word-of-mouth communications influencing users in their purchasing decisions, their trust, loyalties, etc. examining their impact on website perception and use was lacking in current literature. Thus, this thesis examined the effect of expectations on usability and visual appeal, in a website genre where participants do not have highly developed mental models and the website is gender and age neutral. In particular, the primary focus of this thesis was the examination of government websites.

In this thesis, we manipulated website levels of visual appeal and usability. In addition, expectations were controlled – manipulating them to create cognitive dissonance. Cognitive dissonance is a disagreement of information, causing stress that needs to be reduced. According to the theory of cognitive dissonance, the stress if often reduced with the individual‘s agreement to their most recent action or cognition (Harmon-Jones et al., 2009). The most recent behaviour was the experience of the expectation. Therefore, the theory of cognitive dissonance states that new information can create expectations which impact behaviour. In this thesis, we induced different expectations of website usability and visual appeal to examine the impact on the perception and use of the website.

A series of preliminary studies was performed first, in order to obtain a website that was unfamiliar to participants, from a genre that they would have had little experience with. Tourism and city council websites were examined and based on the preliminary study results, this was later reduced to just city council websites. The prettiest and easiest to use website was chosen and the website was manipulated to create several versions of it, ranging in usability and visual appeal. The manipulations were user tested and verified. The main studies were then undertaken with participants interacting with the website to solve information retrieval tasks under controlled conditions. We used textual descriptions of the website to seed users with expectations in written form (hence, written expectations). The results showed that written usability and visual appeal expectations influence objective and subjective usability, and visual appeal. The next section discusses the results and contributions in more detail.

Results and Contributions

We designed and ran a set of experiments to measure the impact of expectation on usability and visual appeal. In the five preliminary studies, we developed an instrument to test familiarity and expectations for city websites. This allowed us to target a less familiar government website (the Gold Coast city council website) and develop an empirically chosen and tested website data sample that statistically varied in usability and visual appeal. A less familiar genre was necessary in order to control for and set expectations in the main studies.

The initial round of experiments tested the easy/pretty and hard/ugly websites, with easy/pretty, hard/ugly, or no textual expectations. In Main Study 1, the results revealed that pre- and post-use visual appeal ratings were significantly lower within the hard/ugly website, when expectations were low (compared to the control). For objective usability, the average number of clicks per task significantly differed within the easy/pretty website, where participants interacted more with the website that had the positive expectations. Therefore, it is possible to affect users‘ perceptions of website visual appeal and alter some aspects of objective usability textually implemented expectations.

Main Study 2 re-tested the easy/pretty website with a confederate who verbally reinforced the written expectations. The findings showed that pre- and post-use perceived usability and post-use visual appeal statistically differed between the easy/pretty and the hard/ugly websites. Moreover, for objective usability, the low expectations group of the easy/pretty website differed from the high expectations group in the average number of clicks per task, average completion time per task, and the average number of passed tasks (where the low expectations group struggled more). Thus, the combination of the verbal implementation of the expectation in addition to the written form was highly successful in influencing participants‘ perceptions and experiences with the city council website.

Main Study 3 examined the easy/ugly and the hard/pretty websites, with easy/ugly, hard/pretty, or no expectations. Expectations were set both textually and verbally. Results were weaker than anticipated, since only pre-use perceived usability was found to vary in the hard/pretty website between the control and low usability, high visual appeal conditions. However, the qualitative data suggested that expectations did influence participants. Therefore, the findings suggest that people are complex and sometimes they ignore new information, rather than accept it. The next step would be to investigate what factors determine whether an individual will accept or reject the expectation.

Thus, the key contribution was the novel approach to examining usability and visual appeal, by manipulating the expectations being experienced. It was done in a highly controlled experiment, with 15 experimental conditions. Evidence from participant feedback, the general graph behaviour, statistical and correlational calculations all point to the conclusion that the combination of textual and verbally implemented expectations impact visual appeal, objective and subjective usability.

Visual appeal and perceived usability differed between conditions in the same website, pre- and post-use, based on a pre-set expectation. It also impacted how participants interacted with the website, struggling more with it when they were told it was going to be hard. Therefore, expectations impact the perception of usability and visual appeal, and the use of websites. Overall, this research contributes to an improved understanding of the relationship of usability and visual appeal by added understanding of the degree to which expectations affect these variables, in a web environment.

Implications

Implication for Technology

By demonstrating the influence of verbal and textual expectations on how one views a website could open positive opportunities for other forms of encouragement for the use of newer technologies. This could be someone who is reluctant to use technology or an elderly person who may need to use an app for medical reasons, such as tweaking a pace maker or tracking their diet. Future work should examine how influencing expectations impacts the acceptance of technology.

Moreover, future work should examine the effectiveness of websites that may already be striving to increase the acceptance of technology. For example, on-going work by Theng et al. (2012) designed a checklist to help foster trust in users, in health and nutricion websites. They found that graphics were used the most to increase trust, where as social cues were used the least. However, they did not mention the success of these means, just the frequency of implementation by developers. Therefore, more research needs to be done on a more diverse set of websites and web domains, on how to better foster technology acceptance.

Implications for Web Design

For website design, the results of this thesis imply that how well a website is made, regarding usability and visual appeal, is not the only factor that influences what people think about it and how they interact with it. Indeed, your reputation precedes you. A bad reputation can turn people against your website, even if the reputation is not true. To overcome the negative impact this research has demonstrated that there is argument for investing in marketing to give a website a more positive reputation right from the beginning. Other research in the field states that spending time and money on the development of a visually appealing and easy to use website is crucial. We add that it is as important to advertise, whether through traditional media, social media or word of mouth, it is important to build reputation to create optimistic preconceptions of both usability and visual appeal making initial usage a more positive experience.

Theoretical Implications Overview

The results of Main Studies 1 and 2 both support the cognitive dissonance theory, in that participants must have reduced the dissonance and agreed with the most recent information. Based on participant feedback, objective and subjective measurements, when visual appeal and usability levels were congruent, then the results offered unambiguous support of the cognitive dissonance theory. While it was relatively easy to predict an individual‘s actions when there was no disagreement, it was impossible to do so when there was dissonance. While not immediately apparent, the findings in Main Study 3 supported the cognitive dissonance theory as well: Participants internalized the expectations and reacted differently to the dissonance. Participants each used different methods to reduce the dissonance. Some added or increased the importance while others took away or reduced the importance of the information causing dissonance. Given these four options, random and insignificant results are almost to be expected. Therefore, all of the results in this thesis can be readily explained by the cognitive dissonance theory.

Future work

The first part of the future work section describes ways in which the limitations mentioned in the previous chapters could be minimized for future studies. This is followed by a discussion on the natural progression of the work, as possible studies to come are outlined.

Reducing the Possible Effect of the Outlined Limitations

No main effects were found for mood. However, the mood scale posed a potential threat to construct validity, in the first main study where SAMS was used. SAMS requires the user to interpret a set of images which caused some ambiguity and inconsistent results. If future studies are interested in mood, then they should consider using a different scale or add clear labels in text above each image. However, it could be that mood did not affect any of the other measured variables. In any case, mood was not a central interest in this thesis, so we did not examine other scales.

The relatively small sample size per condition (n=10) posed a potential threat to the statistical validity. Specifically, had there been more participants, we may have acquired more statistically significant results. Yet, statistically significant results were obtained even with the small sample size. Thus, having more participants per condition may strengthen the support for our hypotheses. Regardless, future studies should strive to have more participants to obtain more reliable results. Alternatively and in addition to, future work could examine only one website version with more participants per condition to gain further insight. In addition, future work should include automated testing processes so that participants could do the test online, individually, and at their own convenience.

Other Websites and Participants

Government websites were used because they gave the thesis a boundary and defined purpose. This work should be extended to inspect experiences on different website genres, such as ecommerce and social media, among others. Furthermore, different populations are needed in order to determine the generalizability of the results. These can include children and elders. Cross-cultural studies should be done to examine if some are more susceptible to accepting new information with regards to website visual appeal and usability.

Using Confederates

Given the success of the addition of confederates in the main studies, future studies should include a confederate as well. This is particularly true if they would like to increase their chances of making a greater impact on their participants. Also, the confederate used in this thesis was unfamiliar to participants. Being in a laboratory environment with unfamiliar faces and locations may have posed a threat to internal validity. Future studies should examine the impact of expectations using the influence of a friend or a family member, via text message, email, or in person, and in the participant‘s home or other such frequented locations. An increase of familiarity and trustworthiness may strengthen the impact of expectations on visual appeal and usability. In addition, the current confederate was an individual with a similar demographic (i.e. student). The impact of the confederate may have been greater had the confederate been a person with some authority, such as a professor, doctor, or police officer. Alternatively, using a more famous individual such as a Youtuber or celebrity, that the participant knows of and perhaps looks up to, may increase the impact of expectations and this should also be examined in future research. Furthermore, alternative methods to implement expectations less overtly (i.e. subliminal) should be investigated for their impact on user perception and interaction with websites.

Other Expectation, Visual Appeal, and Usability Levels

This thesis examined expectation levels that were entirely congruent or entirely incongruent with the website usability and visual appeal levels. For example, with a HuHv website, the expectations were HuHv (congruent) and LuLv (incongruent). Future work should examine what influence only partially congruent expectations have on visual appeal and usability. For example, for the HuHv website, the expectations would be HuLv and LuHv. This would give more insight on to what the influence is of visual appeal and usability on one another, in the presence of expectations that are true for one factor but not for another.

Moreover, the visual appeal and usability levels could be manipulated to more nuanced degrees to make the bad websites more realistically bad. The low conditions in this thesis were created to be terrible which may influence the findings since no real developer would make the usability and visual appeal choices we did (at least we

hope!). Thus, while having more nuanced degrees of usability and visual appeal levels may yield less significant findings since the differences would inherently be less obvious, it would further the understanding of the impact of expectations on visual appeal and usability for more ‗naturally occurring' websites.

Other Forms of Use

We used performance-based information retrieval tasks with clearly defined instrumental goals (find specific information). Future research should examine what would happen under different interaction types (Hassenzahl & Ulrich 2007; van Schaik & Ling 2009). Mainly, would the same results occur if the participants were just browsing (i.e. absence of a goal)?

Other Devices

Recent technological advances have seen the introduction of portable devices to enable viewing of websites, including devices with different sized screens and modes of interaction. As an extension of the research, applicability of the findings on devices varying in screen size should be examined. Some researchers have found that the user experience changes when viewing sites on different screen sizes and through different modes of interaction (Cyr, Head, Larois, & Pan, 2009). Cyr and colleagues (2009) suggested that visual designs influence satisfaction, perceived usefulness, and ease of use of wireless devices. Geissler and colleagues (2006) demonstrated that home page length, number of graphics and links, amount of text, and use of animation impacted perceptions of complexity. This suggests that since small screens have less real estate, the information presented would be denser and thus could appear as more complex than on larger screen displays. Chan and colleagues (2002) examined usability in different mobile (or wireless) device platforms and proposed wireless interface design guidelines. One study by Tarasewich (2002) examined mobile device usability and found that content legibility, quick sequential presentation, device user interaction, and browser types affected the usability of smaller mobile screens. Aesthetics and usability were found to be equally important when designing for a pleasant mobile device user experience (Tarasewich, 2002). However, the effect of expectation has not yet been examined on mobile platforms.
Evidently, more research needs to be done to understand the applicability of the results on different screen sized mobile devices such as mobile phones and tablet devices.

Other Theories to Examine

Another theory related to visual complexity comes from Berlyne (1974). This theory states that a moderately complex stimulus is the most pleasing because too simple stimuli are boring and too complex stimuli cause stress. If the incongruent levels of visual appeal and usability (both in the website and in the expectations) were too

complex, then it is possible that Berlyne's theory could predict what will happen in such cases. However, a precise definition and validity of measurement of visual complexity is lacking in the website evaluation domain, so the application of this theory may not be easy or ideal.

Last Remarks

Evidence to support all three research hypotheses was found in this thesis. When visual appeal and perceived usability levels were congruent then they were rated as higher when the expectation was set to be high, and lower when the expectation was set to be low. In addition, participant performance was also affected by expectations. The addition of verbally enforced expectations impacted the perception and use of a website, more so than just the written task descriptions on their own. Future studies could include a confederate as well, given their success in strengthening the implementation of expectations. The lack of significantly different results in the last main study suggests that incongruent levels of visual appeal and usability are internalized differently by participants and their reaction to the dissonance is unpredictable without further analysis on other factors such as personality traits.

References

Albert, W. & Tedesco, D. (2010). Reliability of Self-Reported Awareness Measures Based on Eye Tracking. *Journal of Usability Studies*, *5*(2), 50 – 64.

Anand, P. (2015). Amazon sues 1,000 people for writing fake reviews. Published: Oct 19, 2015. HTTP://WWW.MARKETWATCH.COM/STORY/AMAZON-SUES-1000-PEOPLE-FOR-WRITING-FAKE-REVIEWS-2015-10-19

Apple. (1996). *Web design guidelines*. Retrieved 17 Feb, 2012, from: http://www.usability.ru/sources/AppleWeb.pdf

Arnheim, R. (1954). Art and visual perception: A psychology of the creative eye. Los Angeles: University of California Press.

Aronson, E., Wilson, T. D., & Akert, R. M. (2010). Social Psychology (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men,* Pittsburgh, PA: Carnegie Press, 177-190.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193, 35–35.

Asch, S. E. (1956). Studies of independence and conformity. A minority of one against a unanimous majority. *Psychological Monographs*, *70*(9), 1–70.

Ba, S. & P.A., Pavlou. (2002). Evidence of the effect of trust in electronic markets: Price premiums and buyer behaviour. *MIS Quart, 26*(3), 243-267.

Baecker, R. M., Grudin, J., Buxton, W. A., & Greenburg, S. (2000). Readings in Human-Computer Interaction: Toward the Year 2000. 2nd Ed. Morgan Kaufmann Publishers, Inc. 573-659.

Bargas-Avila, J., Oberholzer, G., Schmutz, P. De Vito, M., & Opwis, K. (2007). Usable error message presentation in the World Wide Web: Do not show errors right away. *Interacting with Computers*, *19*(3), 330-341.

Bartlett, E. F. (1932). *Remembering*, Cambridge University Press, Cambridge, UK.

Bartuskova, A. & Krejcar, O. (2013). Evaluation framework for user preference research implemented as web application. In ICCCI 2013, LNAI 8083, C., Badica, N.T. Nguyen, and M. Brezovan (Eds.), 537-548.

Bartuskova, A., & Krejcar, O. (2014). Design Requirements of Usability and Aesthetics for e-Learning Purposes. In *Advanced Approaches to Intelligent Information and Database Systems*, Springer International Publishing, 235-245.

Bastien, J. M. C. (2010) Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics, 79*, e18–e23.

Bekker, M., Baauw, E., & Barendregt, W. (2008). A comparison of two analytical evaluation methods for educational computer games for young children. *Cognition, Technology, and Work, 10*, 129–140.

Ben-Ari, M. & Yeshno, T. (2006). Conceptual models of software artifacts. *Interacting with Computers*, *18*(6) 1336-1350.

Ben-Bassat, T., Meyer, J, Tractinsky, N. (2006). Economic and subjective measures of the perceived value of aesthetics and usability. ACM T Comput-Human Int 13(2), 210-234.

Bennett, M. J. (1980). *Heuristics and the weighting of base rate information in diagnostic tasks by nurses*. Unpublished Doctoral Thesis, Monash University, Melbourne, Australia.

Bevan, N. (2009a). What is the difference between the purpose of usability and user experience evaluation methods? UXEM'09 Workshop, INTERACT 2009, Uppsala, Sweden. http://www.nigelbevan.com/cart.htm

Bevan, N. (2009b). Extending Quality in Use to Provide a Framework for Usability Measurement. *Journal of Usability Studies*, *4*(3).

Bevan, N. (2009c). Criteria for selecting methods in user-centred design. *Proceedings in I-USED'09 Workshop, INTERACT 2009*, *Uppsala, Sweden*.

Bias, R. G. (1994). Pluralistic usability walkthrough: coordinated empathies. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods*. (pp. 63-76). New York, NY: Wiley and Sons, Inc.

Blass, T. (1999). "The Milgram paradigm after 35 years: Some things we now know about obedience to authority". *Journal of Applied Social Psychology, 29* (5), 955–978.

Blijlevens, J. (2011). Typically the Best? Perceived Typicality and Aesthetic Appraisal of Product Appearances. PhD thesis, Delft University of Technology, The Netherlands.

Boren, M.T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice, *IEEE Transactions on Professional Communication*, *43*(3), 261–278.

Bowers, C., Cannon-Bowers, J., & Hussain, T. (2009). Considering User Knowledge in the Evaluation of Training System Usability. *Proceedings in Human Centered Design First International Conference, HCD 2009, San Diego, CA, USA*.

Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: Self-Assessment Manikin and the semantic differential. *B&w Thu. & Exp. Psvchrar, 25*(1), 49-59.

Brady L., & Philips, C. (2003). Aesthetics and usability: a look at color balance. *Usability News, 5*(1).

Brajnik, G. (2000). Automatic web usability evaluation: what needs to be done? In the *Proceedings of 6th Human Factors and the Web Conference 2000*. Available at: users.dimi.uniud.it/~giorgio.brajnik/papers/hfweb00.html

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives, *Journal of Abnormal and Social Psychology*, 52, 384-9.

Brooke, J. (1986). System Usability Scale (SUS). Digital Equipment Corporation, UK.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London: Taylor and Francis.

Bruun, A., Gull, P., Hofmeister, L. & Stage, J. (2009). Let Your Users Do the Testing: A Comparison of Three Remote Asynchronous Usability Testing Methods. *Proceedings in 27th international conference on Human factors in computing*

systems. 27. CHI: Human Factors in Computing Systems, Boston, MA, USA, 1619-1628.

Card, S. K., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction. Lawrence Erlbaum Associates, Hillsdale, Nj.

Chan, S. S. Fang, X. Brzezinski, J. Zhou, Y. Xu, S., & Lam J. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research, 3(*3).

Chandy R. & Gu, H. (2012). Identifying Spam in the iOS App Store, in Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, WebQuality _12, ACM, pp. 56–59.

Chang HC., Lai, HH., Chang YM (2007). A measurement scale for evaluating the attractiveness of a passenger car form aimed at young consumers. *Int J Ind Ergonom, 37*(1), 21-30.

Chattratichart, J. & Lindgaard, G. (2008). A comparative evaluation of heuristic-based usability inspection methods. In proceedings of *CHI '08 Extended Abstracts on Human Factors in Computing Systems,* April 28 – May 3, San Jose, USA, pp. 2213-2220.

Chawda, B., Craft, B., Cairns, P., Rüger, S., & Heesch, D. (2005). Do Attractive Things Work Better? An exploration of search tool visualisations. In Proceedings of 19th British HCI group annual conference.

Chevalier J. A. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Choi, Y. J. (2009). Providing Novel and Useful Data for Game Development Using Usability Expert Evaluation and Testing. In *Proceedings of Sixth International Conference on Computer Graphics, Imaging and Visualization,* 129-132.

Coltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the Effectiveness of Interactive Map Interface Designs: A Case Study Integrating Usability Metrics with Eye-movement Analysis. *Cartography and Geographic Information Science*, *36*(1), 5-17.

Craik, K. J. W. (1952). *The Nature of Explanation*. Cambridge University Press, Cambridge, UK.

Cramer, D. & Howitt, D. (2004). The SAGE dictionary of statistics, London: SAGE.

Cramer, D. (1998). Fundamental statistics for social research. London: Routledge.

Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring Human Images in Website Design: A Multi-Method Approach. *MIS Quarterly, 33*(3), 539-566.

Davis, F. D. (1989), "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly 13 (3): 319–340.

Davis, F. D.; Bagozzi, R. P.; Warshaw, P. R. (1989), "User acceptance of computer technology: A comparison of two theoretical models", *Management Science, 35*: 982–1003.

De Angeli, A., Sutcliffe, A., and Hartmann, J. (2006). Interaction, Usability and Aesthetics: What influences users‘ preferences? 271-280. University Park, PA, USA.

de Jong, D.T. (2014). Editorial Menno, Editor, 61(3), *Technical Communication* 145-146.

Dion, K. K., Berscheid, E., & Walster, E. (1972).What is beautiful is good. *Journal of Personality and Social Psvchology, 24*, 285-90.

Doane, D. P. & Seward, L. E. (2011). Measuring Skewness. *Journal of Statistics Education, 19*(2), 1-18.

Druckman J. N. Kam C. D. (2009). Students as Experimental Participants: A Defense of the ―Narrow Data Base"*. SSRN: http://ssrn.com/abstract=1498843

Duan, W., Gu, B., & Whinston, A. B. (2008). Do Online Reviews Matter? – An Empirical Investigation of Panel Data, *Decision Support Systems*, *45*(4), 1007–1016.

Duan, W., Gu, B., & Whinston, A. B. (2008). The Dynamics of Online Word-of-Mouth and Product Sales – An Empirical Investigation of the Movie Industry. *Journal of Retailing, 84*, 233–242.

Dutton, D. (2002). Aesthetic Universals, in The Routledge Companion to Aesthetics, edited by Berys Gaut and Dominic McIver Lopes. http://www.denisdutton.com/universals.htm

Eagly, A. H., Ashmore, R. D., Makhijani, M. G. & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109-128.

Ellison, G. & Fudenberg, D. (1995). Word-of-mouth communication and social learning. *The quarterly Journal of Economics, 110*(1), 93-125.

Farooq, U., & Zirkler, D. (2010). API peer reviews: a method for evaluating usability of application programming interfaces. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, *USA*, 207-210.

Feagin, S. F. (1995). Beauty. In R. Audi (Ed.), The Cambridge dictionary of philosophy Cambridge, England: Cambridge University Press, p. 66.

Festinger, L. & Carlsmith, J. M. (1959) Cognitive consequences of forced compliance, *Journal of Abnormal and Social Psychology*, *58,* 203-211.

Filonik, D. & Baur, D. (2009). Measuring aesthetics for information visualization, Information Visualization, 13th International Conference, 579-584.

Fischhoff, B. (1975). Hindsight = foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288-299.

Fogg, B.J., Kameda, T., Boyd, J., Marshall, J., Sethi, R., Sockol, M., Trowbridge, T. (2002*). Stanford-Makovsky Web Credibility Study 2002: Investigating what makes Web sites credible today.* A Research Report by the Stanford Persuasive Technology Lab in collaboration with Makvosky & Company. Stanford University. Availabe at www.webcredibility.org.

Forrester Research.: The State of Retailing Online (2006). The 9th annual shop.org study. www.clickz.com/3611181.

Fox, N. S., Brennan J. S., Chasen, S. T., & Chasen, B. (2008). "Clinical estimation of fetal weight and the Hawthorne effect". Eur. J. Obstet. Gynecol. *Reprod. Biol. 141*(2), 111–4.

Gefen, D., Karahanna, E., Straub, D., (2003). Trust and TAM in online shopping: An integreated model. MIS Quart., 27(1), 51-90.

Geissler, G. L., Zinkhan, G. M., & Watson, R. T. (2006). The influence of home page complexity on consumer attention, attitudes, and purchase intent. *Journal of Advertising, 35*(2), 69-80.

Gentner, B. R. & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity, in D. Gentner & A.L. Stevens (Eds.), *Mental Models*. Erlbaum, Hillsdale, NJ.

Gould, E. W. (2009). Intercultural Usability Surveys: Do People Always Tell "The Truth"? In N. Aykin (Ed.), *Internationalization, Design, LNCS 5623* (pp. 254–258). Berlin, Germany: Springer-Verlag.

Gray, W. D., &. Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*, 203–261.

Gregory, Y. (1996). *Random House Dictionary of Popular Proverbs and Sayings*, p. 21.

Gulliksen, J., Boivie, I., & Goransson, B. (2006). Usability professionals — current practices and future development. *Interacting with Computers*, *18*, 568–600.

Hall, R. H. & Hanna, P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour and Information Technology, 23(3),* 183-195.

Harmon-Jones, E., Amodio, D., & Harmon-Jones, C. (2009). Action-Based Model of Dissonance: A Review, Integration, and Expansion of Conceptions of Cognitive Conflict. In Mark P. Zanna, editor: Advances in *Experimental Social Psychology*, *41,* Burlington: Academic Press, 119-166.

Harper, B. D. & Norman, K. L. (1993). Improving user satisfaction: The Questionnaire for User Interaction Satisfaction. *Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference,* Santa Monica, CA. Human Factors and Ergonomics Society, pp. 225-233.

Hartmann, J., De Angeli, A., Sutcliffe, A. (2008). Framing the user experience: Information biases on website quality judgment. Proceedings of the Conference on Human Factors in Computing Systems CHI 2008: ACM Press.

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. pp. 396.

Hartmann, J., Sutcliffe, A., & De Angeli, A. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction (TOCHI),* 15(4), 15.

Hashim N. H., Murphy, J., & Law. R. (2007). A Review of Hospitality Website Design Frameworks. In Proceedings of the *International Conference of Information and Communication Technologies in Tourism 2007*, in Ljubljana, Slovenia, 219-230.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction, 19*(4), 319-349.

Hekkert, P, Snelders, D, van Wieringen, PCW (2003). Most advanced, yet acceptable: typicality and novelty as joint predictors of aesthetic preference. Brit J Psychol, 94, 111-124.

Hirschfeld, G. & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. Ergonomics. 2015;58(2):310-20. doi: 10.1080/00140139.2014.965228. Epub 2014 Oct 14.

Hoaglin, D. C., Iglewiez, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labelling. Journal of American Statistical Association, 81, 991-999.

Hoon, L., Vasa, R., Schneider, J., & Grundy, J. (2013). An analysis of the mobile app review landscape: trends and implications. Swinburne Library. Swinburne University of Technology. Faculty of Information and Communication Technologies.

Hoon, L., Vasa, R., Schneider, J.-G., & Mouzakis, K. (2012). ―A Preliminary Analysis of Vocabulary in Mobile App User Reviews," in Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12).

IBM (2008). *Design Principles*. Retrieved 17 Feb, 2012 from: http://www-01.ibm.com/software/ucd/designconcepts/designbasics.html

Ilmberger, W. Schrepp M. & Held. T. (2008). Cognitive processes causing relationship between aesthetics and usability, lecture notes in computer science. HCI and usability for education and work, Springer, 5298, 43-54.

Ip, C., Law, R., & Lee, H., (2011). A review of website evaluation studies in the tourism and hospitality fields from 1996 to 2009. International Journal of Tourism Research, 13(3), 234 – 265.

Ivory, M. Y., & Hearst, M. A. (2001). State of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys, 33*(4), 470-516.

Jaspers, M. W. M. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International journal of medical informatics, 78*, 340–353.

Jeffries, R., & Desurvire, H. (1992). Usability testing vs. heuristic evaluation: Was there a contest? *SIGCHI Bulletin, 24*(4), 39-41.

Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques. In: S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings CHI'91* (pp. 119–124). New York: ACM Press.

Johnson, R. R., Salvo, M. J., & Zoetewey, M. W. (2007). User-Centered Technology in Participatory Culture: Two Decades ―Beyond a Narrow Conception of Usability Testing". *IEEE Transactions on Professional Communication*, *50*(4), 320-332.

Johnson-Laird, P. (1983). *Mental Models*. Harvard University Press, Cambridge, MA.

Kairies (2012). Joy. Aesthetic theories. http://www.nngroup.com/articles/flash-99-percent-bad/. Viewed Nov, 2015.

Kamins, M. A., Folkes, V.S., Perner, L. & Kamins, M. A. (1997). Consumer responses to rumors: Good news, bad news. *Journal of Consumer Psychology*, *6*(2), 165-187.

Katz, A. (2010). Aesthetics, usefulness and performance in user-search-engine interaction. *Journal of Applied Quantitative Methods, 5*(3), 424-445.

Knox & Inkster. (1968). Postdecision dissonance at post time*, Journal of Personality and Social Psychology*, *8*, 319-323

Kurosu M., & Kashimura, K. (1995). Apparent usability vs inherent usability: experimental analysis on the determinants of the apparent usability. *Conference companion on human factors in computing systems.* ACM New York, 292-293.

Lavie, T. & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human Computer Studies, 60*, 269-298.

Law, L. & Hvannberg, E. (2002). Complementarity and Convergence of Heuristic Evaluation and Usability Test: A Case Study of Universal Brokerage Platform. *NordiCHI*.

Lawrence, D. & Tavakol, S. (2007). Balanced Website Design Optimising Aesthetics*, Usability and Purpose*. Spinger-Verlag, London.

Lee, S. & Koubek, R. J. (2010). Understanding user preferences based on usability and aesthetics before and after actual use, *Interacting with Computers, 22*(6), 530-543.

Litvin, S., Goldsmith, R., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458-468.

Livio, M. (2002). *The Golden Ratio: The Story of Phi, The World's Most Astonishing Number*. New York: Broadway Books.

Löfgren, K. (2000). Teacher Education, Statistical Methodologies and the Construction of Knowledge, In *Pierre Bourdieu. Four-Volume Set edition, SAGE Masters in Modern Social Thought series,* 212-228.

Ludden, G.D.S., Schifferstein, H. N. J. & Hekkert, P. (2012) Surprise & Emotion: a longitudinal study of responses to visual – tactual incongruities in products. The International Journal of Design, 6(1), pp 1-10.

Maguire, M. (2001). Methods to support human-centered design. *International Journal of Human-Computer Studies, 55,* 587–634.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: a taxonomy of intrinsic motivations for learning. In: Snow, R. E., & Farr, M. J. (Eds), *Aptitude, learning and interaction III cognitive and affective process analysis.* Erlbaum, Hillsdale.

Maritain, J. (1966). Beauty and imitation. In M. Rader (Ed.), A modern book of Esthetics (3rd ed.), New York: Holt, Rinehart & Winston. pp. 27-34.

Martin, W. E., & Bridgmon, K. (2012). *Quantitative and Statistical Research Methods: From Hypothesis to Results*. Somerset, NJ: Wiley.

McCarney, R., Warner J., Iliffe, S., van Haselen, R., Griffin, M., Fisher, P., Warner, I., Van Haselen, G., Fisher, (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC Med Res Methodol, 7,* 30.

McLellan, S., Muddimer, A., & Peres, S. C. (2012) The Effect of Experience on System Usability Scale Ratings. *Journal of Usability Studies, 7*(2) pp 56-67.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971, 1992). Manual for the profile of mood states. San Diego: Educational and Industrial Testing Service.

Medlock, M. C., Wixon D., Terrano, M., Romero, R., & Fulton, B. (2002). Using the RITE Method to improve products: a definition and a case study. *Usability Professionals Association*, Orlando, FL.

Meinald, T., Thielsch, Engel1, R., & Hirschfeld, G. (2015). Expected usability is not a valid indicator of experienced usability. *Peer J computer science*.

Milgram, S. (1963). Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology, 67* (4), 371–8.

Milgram, S. (1974a). Obedience to Authority; An Experimental View. Harpercollins.

Milgram, S. (1974b). "The Perils of Obedience". Harper's Magazine. Archived from the original on 2011-05-14. Abridged and adapted from Obedience to Authority.

Mook, D. G. (1983). In Defense of External Invalidity. *American Psychologist, 38,* 379-387.

Moshagen, M. & Thielsch, M. (2012). A short version of the visual aesthetics of websites inventory. Behaviour and Information Technology, 1-7.

Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human Computer Studies, 68*, 689-709.

Moshagen, M., Musch, J., & Goritz, A. S. (2009). A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance, *Ergonomics, 52*.

Nakarada-Kordic, I. Lobb B. (2005) Effect of perceived attractiveness of web interface design on visual search of web sites. In proceedings of the 6th ACM SIGCHI – CHINZ '05. ACM: New York, 25-27.

Newell, A. & Card, S. (1995). The Prospects of Psychological Science in Human-Comptuer Interaction. Human-Computer Interaction 1(3), 251-267.

Nielsen, J. (1993). Usability Engineering, *Academic Press*, Cambridge, MA. SF: Morgan Kaufman.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen, & R. L. Mack (Eds.), *Usability inspection method,* John Wiley & Sons, New York, pp. 25–62.

Nielsen, J. (2000) „Designing Web Usability", USA: New Riders, ISBN 1-56205-810-X

Nielsen, J. (2002), http://www.useit.com/ viewed on 29th of July 2004.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces, *Proceedings of the ACM CHI'90 Conference.* Seattle, WA, 249-256.

Nisbett, R. E & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84,* 231-259.

Nordstokke, D. W. & Zumbo, B. D. (2010). A new nonparametric Levene test for equal variances. *Psicologica, 31*(2), 401-430.

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with

assessment and evaluation data. Practical Assessment, Research & Evaluation, 1(5).

Norman, D. (2004).Emotional design: Why we love (or hate) everyday things. New York: Basic Books.

Olson, G. M., & Olson, J. S. (2003). Human-Computer Interaction: Psychological Aspects of the Human Use of Computing. *Annual Review of Psychology, 54*, 491–516.

Olson, J. & Olson G. (1990). The Growth of Cognitive Modeling in Human-Computer Interaction Since GOMS. Human-Computer Interaction 5, 221-265.

Oulasvirta, A. (2004). Task demands and memory in web interaction: A levels of processing approach. *Interacting with Computers*, *16*(2), 773-789.

Pacioli, L. (1509). *De divina proportione*, Luca Paganinem de Paganinus de Brescia (Antonio Capella), Venice.

Partala, T., & Kangaskorte, R. (2009) The Combined Walkthrough: Measuring Behavioral, Affective, and Cognitive Information in Usability Testing. *Journal of Usability Studies*, *5*(1), 21-33.

Parthasarathy, R., & Fang, X. (2013). Introducing Emotional Interfaces to Healthcare Systems. In *Human-Computer Interaction. Applications and Services* (pp. 150-162). Springer Berlin Heidelberg.

Parush, A., Kramer, C., Foster-Hunt, T., Momtahan, K., Hunter, A., & Sohmer, B. (2011). Communication and team situation awareness in the OR: Implications for augmentative information display. *Journal of Biomedical Informatics, 44,* 477–485.

Pavlou, P.A. & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, *17*(4), 392-414.

Pawson, M., & Greenberg, S. (2009). Extremely Rapid Usability Testing. *Journal of Usability Studies*, *4*(3), 124-135.

Plott, C. R. (1991). Will Economics Become an Experimental Science? *Southern Economic Journal, 57,* 901-919.

Polson, P., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive Walkthroughs: a method for theory-based evaluation of user interfaces, *International Journal of Man-Machine Studies*, *36*, 741–773.

Prisacari, A., & Holme, T. (2013). Using eye-tracking to test and improve website design. In *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools* (pp. 389-398). Springer Berlin Heidelberg.

Quinn, J. M., & Tran, T. Q. (2010). Attractive phones don‘t have to work better: Independent effects of attractiveness, effectiveness, and efficiency on perceived usability. In CHI ‘10: Proceedings of the 28th international conference on human factors in computing systems. NY: New York, pp. 353–362.

Razali, N. M. & Wah, Y. B. (2011). Power comparisons of ShapiroWilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Read, H. (1972). The meaning of art. London: Faber & Faber.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*(4), 364-382.

Roth, S. P., Schmutz, P., Pauwels, S. L., Bargas-Avila, J. A., & Opwis, K. (2010). Mental models for web objects: Where do users expect to find the most frequent objects in online shops, news portals, and company web pages? *Interacting with Computers*, *22,* 140-152.

Santayana, G. (1955). The sense of beauty. New York: Dover. (Original work published in 1896).

Sauer, J., & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. Appl Ergonomics, 40, 670-677.

Sauro, J. (2013). A single-item measure of website usability: Comments on Christophersen and Konradt (2011). Interacting with computers, 25(4), 271-277.

Schank R. C. & Abelson R. P. (1977). Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures (Chap. 1-3), L. Erlbaum, Hillsdale, NJ.

Schenkman, B.N. Jonsson FU (2000). Aesthetics and preferences of web pages. Behav Inform Technol 19 (5) 367-377.

Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise, *International Journal of Human-Computer Interaction, 9*(3), 213–234.

Sears, A., & Hess, D. J. (1999). Cognitive Walkthroughs: understanding the effect of task description detail on evaluator performance. *International Journal of Human-Computer Interaction, 11*(3), 185–200.

Shaik, A. & Lenz, K. (2006). *Where's the search? Re-examining user expectations of web objects*. Retrieved 17 Feb, 2012 from: http://www.surl.org/usabilitynews/81/webobjects.asp 17 Feb 2012.

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52*(3/4), 591-611.

Sinreich, D., Gopher, D. Ben-Barak, S., Marmur, Y., & Menchel, R. (2005). Mental models as a practical tool in the engineer's toolbox . *International Journal of Production Research*, ***43***, 2977- 2996.

Slavkovic, A., & Cross, K. (1999). Novice heuristic evaluations of a complex interface. In the *CHI '99 extended abstracts on Human factors in computing systems*, (pp. 304-305). New York: ACM Press.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. Journal of interacting marketing, 19(3), 15-37.

Sonderegger A., Sauer J. and Eichenberger J. (2014). Expressive and classical aesthetics: two distinct concepts with highly similar effect patterns in user-artefact interaction. *Behaviour & Information Technology, 33*, 1180-1191.

Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. *Appl Ergnomics, 41*, 403-410

Sonderegger, A., Uebelbacher, A., Pugliese, M., & Sauer, J. (2014). The influence of aesthetics in usability testing: the case of dual-domain products. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* April (pp. 21-30). ACM.

Sonderegger, A., Uebelbacher, A., Pugliese, M., & Sauer, J. (2014). The influence of aesthetics in usability testing: the case of dual-domain products. In proceedings of CHI 2014, Toronto, Canada, 21-30

Sonderegger, A., Zbinden, G. Uebelbacher, A, Sauer, J. (2012). The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. Ergonomics 55(7), 713-730.

Spencer, R. (2000). The streamlined Cognitive Walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of CHI 2000: ACM Annual Conference on Human Factors in Computing Systems*, (353-359). New York: ACM Press.

Sutcliffe, A. G. (2001). Heuristic evaluation of website attractiveness and usability. In GIST technical report G2001/1: Proceedings 8th Workshop on Design, Specification and Verification of Interactive Systems Dept of Computer Science, University of Glasgow, Glasgow, 188-199.

Svahnberg, M., Aurun, A., & Wohlin, C. (2008). Using students as subjects - an empirical evaluation. In proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement ESEM '08. 288-290

Taebi, O., Aldabbas, H., & Clarskon, M. (2013). Users‘ perception towards usability and aesthetics design of travel websites. *In Proceedings of The international Conference on E-Commerce & Information Technolog, EcomIT & GBM, 117,* Sri Lanka.

Tarasewich, (2002). P. Wireless devices for mobile commerce: User interface design and mobility, in: B. Mennecke, T.J. Strader, (eds), Mobile Commerce: Technology, Theory, and Applications, Hershey, PA: Idea Group Publishing, 26–50.

Thielsch et al. (2015), PeerJ Comput. Sci., DOI 10.7717/peerj-cs.19

Thielsch M.T., Blotenberg, I., & Jaron, R. (2013). User evaluation of websites: from first impression to recommendation. *Interacting with computers, 26*(1), 89-102.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

Thurgood, C., Whitfield, A. T. W., & Patterson, J. (2011). Towards a visual recognition threshold: New instrument shows humans identify animals with only 1 ms of visual exposure. *Vision Research, 51*, 1966–1971.

Thuring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *Int J Psychol. 42,* 253-264.

Toffler, A. (1970). *Future Shock*. Bradley, London UK.

Tooby, J., & Cosmides, L. (2001). Does Beauty Build Adapted Minds? Toward an Evolutionary Theory of Aesthetics, Fiction, and the Arts. *SubStance, 30*(1&2), 6-27.

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. *Proceedings of ACM SIGCHI* (pp. 115-122), New York, NY, USA. ACM.

Tractinsky, N. (2004). Toward the study of aesthetics in information technology, in: Proceedings Twenty-Fifth International Conference on Information Systems, 771–780.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2), 127-145.

Tuch, A. N., Roth, S. P., Hornbaek, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior, 28*(5), 1596-1607.

Van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2009). Evaluating municipal websites: A methodological comparison of three think-aloud variants. *Government Information Quarterly, 26*, 193–202.

Van den Haak, M. J., de Jong, M. D. T., Schellens, P. J. (2003) Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour and Information Technology, 22*(5) 339–351.

Van der Heijden, H. (2003). Factors influencing the usage of websites: the case of a generic portal in the Netherlands. Information & Management, 40(6), 541-549.

Van Schaik, P, & Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. Int J Hum Comp St, 67(1), 79-89.

Varela, M., Maki, T., Skorin-Kapov, L., & Hossfeld, T. (2013, July). Towards an understanding of visual appeal in website design. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on* (pp. 70-75). IEEE.

Vasa, R., Hoon, R., Mouzakis, K., & Noguchi, A. (2012). A preliminary analysis of mobile app user reviews. OZCHI 2012, Melbourne, Australia, 241-244.

Vermeulen I. E., & Seegers, D. (2009). –Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration," *Tourism Management, 30*, pp. 123–127.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspections methods*, (pp. 105-140). New York: Wiley.

Wickens, C. D., Lee, J. D., Liu, Y., and Becker, S. E. G. (2004). An introduction to human factors engineering (2nd ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Williams, R. (1994). The Non-Designers Design Book. Peachpit Press.

Yang, C. Y. (2009). Website Designer as an Evaluator: A Formative Evaluation Method for Website Interface Development. In J.A. Jacko (Ed.), *Human-Computer Interaction*, Part I, HCII 2009, LNCS 5610, Springer, 372–381.

Ye, Q., Law, R., & Gu, B. (2009). The Impact of Online User Reviews on Hotel Room Sales, *International Journal of Hospitality Management*, *28*(1), 180–182.

Theng, Y., Ying Qin Goh, L., Tin, M. T., Sopra, R., & Kumar, S. K. P. (2012). Trust cues fostering initial consumers' trust: usability inspection of nutrition and healthcare websites. In Proceedings of the *2nd ACM SIGHIT International Health Informatics Symposium, IHI 2012*. New York, NY, 807-812.

Yu, G. L. (2013). Research on the Optimization of Flexibility Selection of Limb Aesthetics in the Dance Movements Based on Dynamics and Geometry Morphology. *Applied Mechanics and Materials*, *380*, 4872-4876.

Zettl, H. (1999). Sight, Sound, Motion: Applied media aesthetics. Belmont, CA: Wadsworth.

Appendices

Appendix A: Ethics Approvals

Appendix A1: First Ethics Approval

SUHREC Project 2013/011 Understanding the effect of expectation on the relationship of usability and aesthetics in a web environment on different screen sized devices
Prof Christopher Pilgrim, FICT/ Ms Milica Stojmenovic
Approved Duration: 21/03/2013 To 31/08/2013 [Adjusted]

I refer to the ethical review of the above resubmitted and revised project protocol undertaken on behalf of Swinburne's Human Research Ethics Committee (SUHREC) by SUHREC Subcommittee (SHESC2) at a meeting held on 8 February 2013. Your response to the review as e-mailed on 26 February 2013 was reviewed by SHESC2 delegates.

I am pleased to advise that, as submitted to date, the project may proceed in line with standard on-going ethics clearance conditions here outlined.

- All human research activity undertaken under Swinburne auspices must conform to Swinburne and external regulatory standards, including the National Statement on Ethical Conduct in Human Research and with respect to secure data use, retention and disposal.
- The named Swinburne Chief Investigator/Supervisor remains responsible for any personnel appointed to or associated with the project being made aware of ethics clearance conditions, including research and consent procedures or instruments approved. Any change in chief investigator/supervisor requires timely notification and SUHREC endorsement.
- The above project has been approved as submitted for ethical review by or on behalf of SUHREC. Amendments to approved procedures or instruments ordinarily require prior ethical appraisal/ clearance. SUHREC must be notified immediately or as soon as possible thereafter of (a) any serious or unexpected adverse effects on participants and any redress measures; (b) proposed changes in protocols; and (c) unforeseen events which might affect continued ethical acceptability of the project.
- At a minimum, an annual report on the progress of the project is required as well as at the conclusion (or abandonment) of the project.
- A duly authorised external or internal audit of the project may be undertaken at any time.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance or you need a signed ethics clearance certificate, citing the SUHREC project number. A copy of this clearance email should be retained as part of project record-keeping.

Best wishes for the project.
Yours sincerely
Kaye Goldenberg

Secretary, SHESC2, Administrative Officer (Research Ethics), Swinburne Research (H68), Swinburne University of Technology, P O Box 218, HAWTHORN VIC 3122, Tel +61 3 9214 8468

Appendix A2: Second Ethics Approval

SUHREC 2013/075 Understanding the effect of expectation on the relationship of usability and aesthetics in a web environment on different screen sized devises.
Prof C Pilgrim Ms Milica Stojmenovic FICT
Approved duration: 07/05/2013 To 31/12/2013 [Adjusted]

I refer to the ethical review of the above project protocol undertaken on behalf of Swinburne's Human Research Ethics Committee (SUHREC) by SUHREC Subcommittee (SHESC2) at a meeting held on 19th April 2013. Your response to the review as e-mailed on 6 May 2013 was reviewed.

I am pleased to advise that, as submitted to date, the project may proceed in line with standard on-going ethics clearance conditions here outlined.

- All human research activity undertaken under Swinburne auspices must conform to Swinburne and external regulatory standards, including the current *National Statement on Ethical Conduct in Human Research* and with respect to secure data use, retention and disposal.

- The named Swinburne Chief Investigator/Supervisor remains responsible for any personnel appointed to or associated with the project being made aware of ethics clearance conditions, including research and consent procedures or instruments approved. Any change in chief investigator/supervisor requires timely notification and SUHREC endorsement.

- The above project has been approved as submitted for ethical review by or on behalf of SUHREC. Amendments to approved procedures or instruments ordinarily require prior ethical appraisal/ clearance. SUHREC must be notified immediately or as soon as possible thereafter of (a) any serious or unexpected adverse effects on participants and any redress measures; (b) proposed changes in protocols; and (c) unforeseen events which might affect continued ethical acceptability of the project.

- At a minimum, an annual report on the progress of the project is required as well as at the conclusion (or abandonment) of the project.

- A duly authorised external or internal audit of the project may be undertaken at any time.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance. The SUHREC project number should be quoted in communication. Chief Investigators/Supervisors and Student Researchers should retain a copy of this email as part of project record-keeping.

Best wishes for project.

Yours sincerely,

Ann Gaeth
Administration Officer (Research Ethics), Swinburne Research (H68), Swinburne University of Technology, P O Box 218, HAWTHORN VIC 3122, (+61 3 9214 83567 +61 3 9214 5267

Appendix A3: Third Ethics Approval

SUHREC 2013/075 Understanding the effect of expectation on the relationship of usability and aesthetics in a web environment on different screen sized devises.
Prof C Pilgrim, Ms M Stojmenovic, FSET
Approved duration: Extended to 31/12/2014 [Modified: January 2014, March 2014, July 2014]

I refer to your further request for ethics clearance for modifications to the approved protocol concerning a larger participant cohort and a variation to the research method (use of a confederate). The request, as emailed on 25 July 2014, was put to a delegate of the SUHREC Subcommittee (SHESC2) for consideration.

I am pleased to advise that, as modified to date, the project has approval to continue in line with ethics clearance conditions previously communicated and reprinted below.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance citing the project number. A copy of this email should be retained as part of project record-keeping.

As before, best wishes for the project.

Yours sincerely,
Astrid Nordmann
Secretary, SHESC2

Dr Astrid Nordmann
Research Ethics Executive Officer
Swinburne Research (H68)
Swinburne University of Technology
PO Box 218, Hawthorn, VIC 3122
Tel: +613 9214 3845
Fax: +613 9214 5267
Email: anordmann@swin.edu.au

Appendix A4: Fourth Ethics Approval

SUHREC 2013/075 Understanding the effect of expectation on the relationship of usability and aesthetics in a web environment on different screen sized devises.
Prof C Pilgrim, Ms M Stojmenovic, FSET et al
Approved duration: Extended to 31/12/2014 [Modified: January 2014, March 2014]

I refer to your further request for ethics clearance for modifications to the approved protocol concerning a larger participant cohort and a variation to the research method. The request, as emailed on 20 February 2014, was put to a delegate of the SUHREC Subcommittee (SHESC2) for consideration.

I am pleased to advise that, as modified to date, the project has approval to continue in line with ethics clearance conditions previously communicated and reprinted below.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance citing the project number. A copy of this email should be retained as part of project record-keeping.

As before, best wishes for the project.

Yours sincerely

Keith for
Secretary, SHESC2

Keith Wilkins
Secretary, SUHREC & Research Ethics Officer
Swinburne Research (H68)
Swinburne University of Technology
P O Box 218
HAWTHORN VIC 3122
Tel +61 3 9214 5218
Fax +61 3 9214 5267

Appendix B: Universal Forms and Metrics

Appendix B1: Demographics Questionnaire

Demographics

Thank you for signing up for this study of website visual appeal and usability. The purpose of this short questionnaire is to acquire some general background information. This questionnaire is strictly confidential. You are in no way obliged to complete this survey. Please return this survey to me as soon as you have completed it. In advance, I thank you for your time.

Please circle the answer(s) that best suits you.

1) Age: Up to 17 18-30 31 or over

2) Gender: Male Female

3) Were you born in Australia? Yes No

4) What do you use the Internet regularly for? (Circle more than one option if appropriate)

Banking Shopping Entertainment Study News Social Travel

5) How familiar are you with the purposes of Local City Councils?

Not very Somewhat Very

You will hear from me shortly to confirm a time and place for the main part of the study.

Appendix B2: Consent Form

Web Usability and Visual Appeal Study

Participant Consent Form

Principal Investigators: Professor Chris Pilgrim and Milica Stojmenovic, Faculty of Information and Communication Technologies.

1. I consent to participate in the project named above. I have been provided a copy of the Project Information Statement to which this consent form relates and any questions I have asked have been answered to my satisfaction.
2. I acknowledge that:
    a. My participation is voluntary and that I am free to withdraw from the project at any time without explanation;
    b. The Swinburne project is for the purpose of research and not for profit;
    c. Any identifiable information about me which is gathered in the course of and as the result of my participating in this project will be (i) collected and retained for the purpose of conducting this project;
    d. My anonymity is preserved and I will not be identified in publications of otherwise without my express written consent.

By signing this document I agree to participate in this project.

I acknowledge that I have received a gift card.

Name of Participant: ……

Signature & Date: ……

Faculty of Information & Communication Technologies

John Street Hawthorn
Victoria 3122 Australia

PO Box 218 Hawthorn
Victoria 3122 Australia

Telephone +61 3 9214 8731
Facsimile +61 3 9214 5916

http://www.swin.edu.au/ict/

ABN 13 628 586 699
CRICOS Provider 00111

Appendix B3: Project Information Sheet

Web Usability and Visual Appeal Study

Project Information Statement and Tasks

Thank you for participating in this experiment. The purpose of this experiment is to examine the usability and visual appeal of certain types of websites to help us select the most and least visually appealing, as well as the least and the most usable webpages for future studies.

Investigators

This experiment is part of a project being conducted by **Milica Stojmenovic** who is a full time PhD student under the supervision of **Professor Chris Pilgrim** who is a full-time member of staff of the Faculty of Information and Communication Technologies, and under the supervision of **Professor Gitte Lindgaard**, a part-time member of the Faculty of Design, from Swinburne University of Technology.

Participants

The participants for this experiment are being drawn from the general Swinburne community. All participation is voluntary and does not have any bearing on student results or staff employment.

A consent form must be signed prior to involvement in this experiment.

You will receive a $20 Gift Card upon completion of the session to compensate you for your time.

Description of the Experiment

The full experiment session will take approximately one hour.

There will be several sections to this experiment:

- You will be asked to fill out a brief mood questionnaire.
- In a practice trial, you will be asked to view and rate a website, including the homepage and two other pages from the same website. Each of these sites will be quickly flashed to you. Once you have seen all three sites, you will be asked to rate the website as a whole, on usability and visual appeal using the scales found in front of you.
- Then, you will be asked to rate two separate sets of 26 websites on usability and visual appeal. Please rate the websites as quickly as possible to ensure that your ratings are reflect your first impression. To ensure that the experiment takes approximately an hour, after a period of time, you may be asked to stop the task.
- At the end of each set of webpages, you will be asked to complete a short checklist and questionnaire about your expectations for a given genre of webpages.

Participant Rights and Confidentiality

- This experiment is designed to examine the usability and visual appeal of websites and not to test your abilities.
- You may leave the experiment at any time if you feel uncomfortable or concerned about any of the experimental tasks or surveys.
- No video or audio recording will be made.
- The information obtained will be treated in strict confidence in accordance with Swinburne's Policy on the Conduct of Research. There will be no recorded details of the identity of participants connected with the recorded data or survey results. All data will be securely stored

during and after analysis. All recordings and data will be destroyed after 5 years. Only the Investigators will have access to this data.

- Participant consent forms will be stored separately to any data collected. This prevents data-matching and preserves anonymity.

Research Outputs

- Aggregate data only may be used in subsequent academic publications. These data will not identify any participants. A copy of any publication or report will be made available to participants upon request.
- The outcomes of this project will provide the empirical basis for further research into the relationship between usability and aesthetics in a web domain.

If you have any questions, or require any follow up, after the completion of this experiment, please contact **Professor Chris Pilgrim** at cpilgrim@swin.edu.au or 9214 5231. Your query will be responded to promptly. If you do not feel well, please see Swinburne Health Care services.

This project has been approved by or on behalf of Swinburne's Human Research Ethics Committee (SUHREC) in line with the National Statement on Ethical Conduct in Human Research. If you have any concerns or complaints about the conduct of this project, you can contact:
Research Ethics Officer, Swinburne Research (H68),
Swinburne University of Technology, PO Box 218, HAWTHORN VIC 3122.
Tel (03) 9214 5218 or +61 3 9214 5218 or resethics@swin.edu.au

Faculty of Information & Communication Technologies

John Street Hawthorn
Victoria 3122 Australia

PO Box 218 Hawthorn
Victoria 3122 Australia

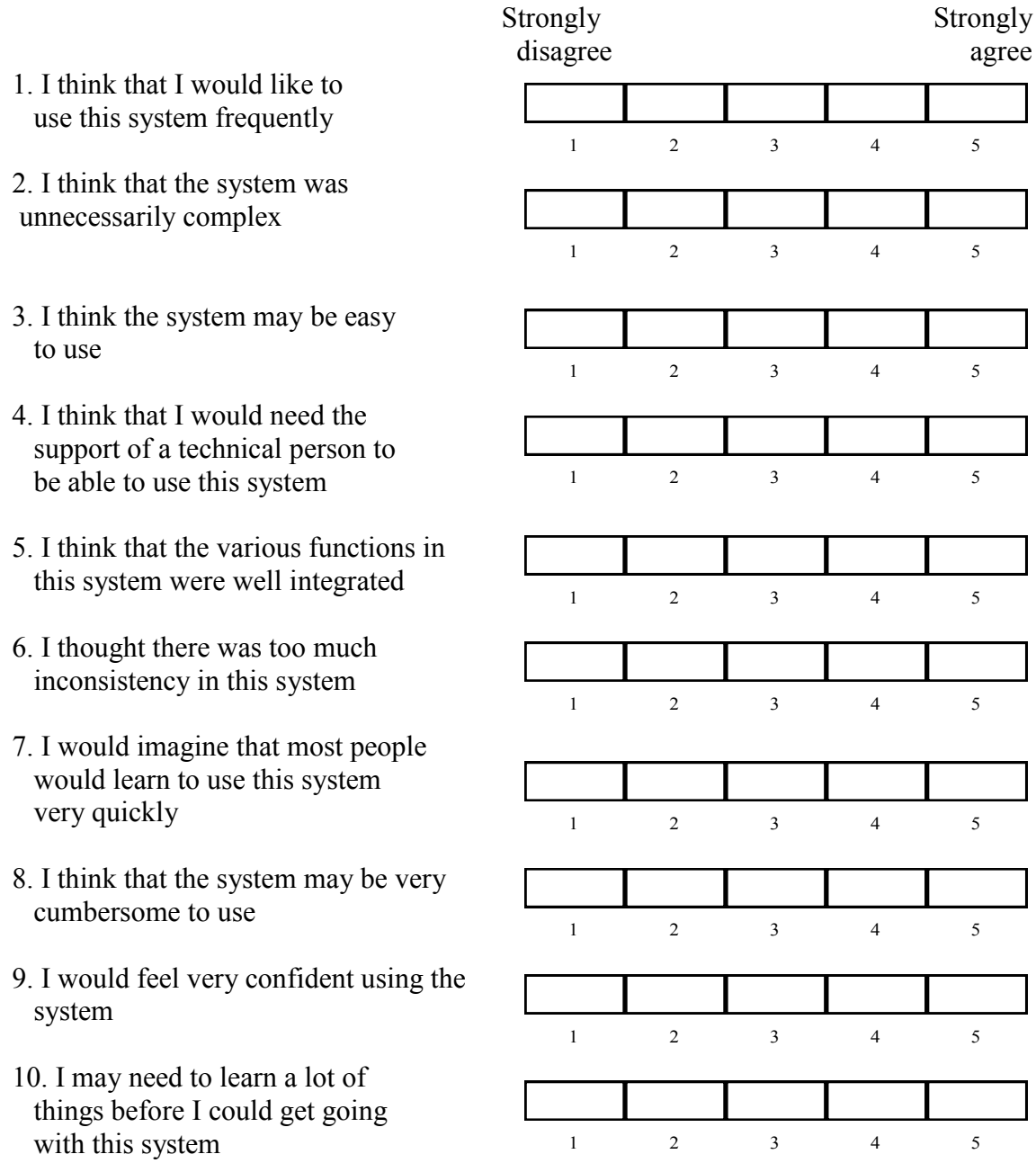
Telephone +61 3 9214 8731
Facsimile +61 3 9214 5916

http://www.swin.edu.au/ict/

ABN 13 628 586 699
CRICOS Provider 00111

Appendix B4: The modified System Usability Scale
© Digital Equipment Corporation, 1986.

(modified) System Usability Scale (SUS)



Appendix B5: The visual appeal questionnaire used called VisAWI-S.

Visual Appeal

Facet 1: Simplicity
Everything goes together on this site.

Strongly Disagree	Disagree	Weak Disagree	Neither	Weak Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Facet 2: Diversity
The layout is pleasantly varied.

Strongly Disagree	Disagree	Weak Disagree	Neither	Weak Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Facet 3: Colourfulness
The colour composition is attractive.

Strongly Disagree	Disagree	Weak Disagree	Neither	Weak Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Facet 4: Craftsmanship
The layout appears professionally designed.

Strongly Disagree	Disagree	Weak Disagree	Neither	Weak Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Appendix C: Preliminary Studies

Appendix C1: Expectation checklist and graded semantic differential questions

Item Checklist

Which of the following Items would you expect to be found on a <*website genre*> website? Please check all that may apply:

__Photo(s) of People
__Photo(s) of Objects and Animals
__Photo(s) of Scenery
__Login
__Shopping Cart
__Calendar
__Logo
__Advertisements
__Weather
__Translator
__Save Button
__Print Button
__Videos
__News
__Discounts and Specials
__Help
__Payment options
__Games/Puzzles
__Contact Information
__Import/Export Functions (ex. Email)
__Order Tracking
__Search
__Social Media (ex. Facebook)
__Navigation (ex. Breadcrumbs)
__Other:_____
_____

Overall Expectations

Please circle the words that you associate most with <*website genre*>websites.

a. Usable	1	2	3	4	5	Unusable
b. Stressful	1	2	3	4	5	Relaxing
c. Complicated	1	2	3	4	5	Easy
d. Enjoyable	1	2	3	4	5	Frustrating
e. Boring	1	2	3	4	5	Exciting
f. Fast	1	2	3	4	5	Sluggish
g. Inefficient	1	2	3	4	5	Efficient
h. Bad	1	2	3	4	5	Good
i. Pretty	1	2	3	4	5	Ugly

Appendix C2: Task Description

Tasks

1. First, you will be asked to fill out a brief mood survey.
2. Practice Round: Please view the three webpages that will be shown for less than a second to you. Please be vigilant. Once you have seen the webpages, please rate them on usability and visual appeal using the online scales. Please rate the webpages quickly to ensure that your answer reflects your first impression. That will conclude the practice round.
3. Please feel free to ask any questions of clarification before starting the formal study.
4. Round 1, Part 1: Once you have completed rating the practice website, you may continue to the real experiment, where you will repeat the same viewing and rating procedure for 26 webpages.
5. Round 1, Part 2: Once the 26 webpages have been viewed and rated, you will be asked to fill out an expectation questionnaire, asking you to choose components of a webpage from a checklist that you would expect to find in a given genre, and to give your attitudes towards the genre.
6. Round 2: Please repeat the instructions for Round 1, Part 1 and 2, but with a different set of webpages presented to you.

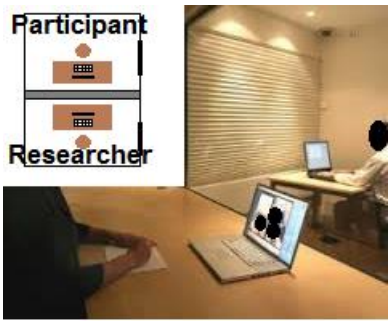
Appendix C3: List of the websites participants saw and rated in the first preliminary study.

All websites were retrieved in July, 2012.

1. Adelaide
    a. www.adelaidecitycouncil.com
    b. www.southaustralia.com/regions/adelade-city.aspx
2. Albury
    a. www.alburycity.nsw.gov.au
    b. www.visitalburywondonga.com
3. Ballarat
    a. www.ballarat.vic.gov.au
    b. www.visitballarat.com.au
4. Bendigo
    a. www.bendigo.vic.gov.au/Home
    b. www.bendigotourism.com
5. Brisbane
    a. www.brisbane.qld.gov.au
    b. www.queenslandholidays.com.au/destinations/brisbane/48-hours-48-hot-spots/48-hours-48-hot-spots_home.cfm
6. Burnie
    a. www.burnie.net/Home
    b. www.discoverburnie.net
7. Cairns
    a. www.cairns.qld.gov.au
    b. www.visitcairns.com.au
8. Canberra
    a. www.act.gov.au
    b. www.visitcanberra.com.au
9. Darwin
    a. www.darwin.nt.gov.au
    b. www.travelnt.com/darwin-and-surrounds.aspx
10. Devonport
    a. www.devonport.tas.gov.au
    b. www.devonporttasmania.travel
11. Geelong
    a. www.geelongaustralia.com.au
    b. www.visitgeelongbellarine.com.au
12. Gold Coast
    a. www.goldcoast.qld.gov.au
    b. www.visitgoldcoast.com
13. Hobart
    a. www.hobartcity.com.au
    b. www.welcometohobart.com.au
14. Launceston
    a. www.launceston.tas.gov.au
    b. www.visitlauncetontamar.com.au
15. Mandurah
    a. www.mandurah.wa.gov.au
    b. www.visitmandurah.com

16. Melbourne
    a. www.melbourne.vic.gov.au
    b. www.visitmelbourne.com
17. New Castle
    a. www.newcastle.nsw.gov.au
    b. www.visitnewcastle.com.au
18. Perth
    a. www.cityofperth.wa.gov.au
    b. www.experienceperth.com
19. Sunshine Coast
    a. www.sunshinecoast.qld.gov.au
    b. www.visitsunshinecoast.com.au
20. Sydney
    a. www.cityofsydney.nsw.gov.au
    b. www.sydney.com
21. Toowoomba
    a. www.toowoombarc.qld.gov.au
    b. www.toowoomba.com
22. Townsville
    a. www.townsville.qld.gov.au
    b. www.townsvilleholidays.info
23. Tweed
    a. www.tweed.nsw.gov.au
    b. www.tweedtourism.com.au
24. Wagga Wagga
    a. www.wagga.nsw.gov.au
    b. www.waggawaggaaustralia.com.au
25. Wodonga
    a. www.wodonga.vic.gov.au
    b. www.mackayregion.com
26. Wollongong
    a. www.wollongong.nsw.gov.au
    b. www.visitwollongong.com.au

Usability laboratory in Swinburne, separated by a one-way mirror:



Appendix C4: Phase 1 Results

Table 1. Percent of participants (expected) and percent of websites (actual) with a website component, per genre.

Category	Expected City Council	Expected Tourism	Actual City Council	Actual Tourism
Advertisements	16. 7%	40.0%	0.0%	20.0%
Calendar	53.3%	76.7%	12.0%	20.0%
Contact Details	90.0%	93.3%	100%	100%
Discounts	20.0%	76.7%	0.0%	24.0%
Help	76.7%	60.0%	0.0%	8.0%
Import/Export	23.3%	26.7%	8.0%	36.0%
Login	20.0%	26.7%	0.0%	4.0%
Logo	83.3%	80.0%	100%	88.0%
Navigation	70.0%	60.0%	64.0%	44.0%
News	93.3%	50.0%	80.0%	16.0%
Order tracking	0.0%	13.3%	0.0%	0.0%
Payment	26.7%	46.7%	76.0%	24.0%
Photos of objects	43.3%	80.0%	44.0%	76.0%
Photos of people	70.0%	86.7%	68.0%	100%
Photos of scenery	63.3%	100%	72.0%	100%
Print button	30.0%	43.3%	20.0%	28.0%
Save button	6.7%	20.0%	4.0%	0.0%
Search	80.0%	80.0%	100%	100%
Shopping Cart	0.0%	16.7%	0.0%	20.0%
Social media	33.3%	53.3%	28.0%	76.0%
Translator	20.0%	33.3%	8.0%	24.0%
Videos	36.7%	73.3%	4.0%	16.0%
Weather	33.3%	73.3%	40.0%	44.0%

Table 2. Usability and visual appeal, non-parametriccorrelations

			AesthAv	UsabAv
Spearman's rho	AesthAv	Correlation Coefficient	1.000	.643**
		Sig. (2-tailed)	.	.000
		N	1559	1559
	UsabAv	Correlation Coefficient	.643**	1.000
		Sig. (2-tailed)	.000	.
		N	1559	1559

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Highest and lowest visually appealing websites, with means and genres.

Top 6 (three per genre)	Bottom 6 (three per genre)
Gold Coast (5.966; city council)	Toowoomba (2.55833; tourism)
Sunshine Coast (5.925; tourism)	Brisbane (3.74167; tourism)
Adelaide (5.84167; tourism)	Tweed (3.9666; city council)
Gold Coast (5.68966; tourism)	Newcastle (4.0583; tourism)
Bendigo (5.54167; city council)	Perth (4.075; city council)
Devonport (5.30833; city council)	Ballarat (4.30833; city council)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
F1Simplicity	.108	30	.200*	.972	30	.606
F2Diversity	.074	30	.200*	.991	30	.996
F3Colourfulness	.094	30	.200*	.984	30	.915
F4Craftmanship	.080	30	.200*	.984	30	.913
@1Usefrequently	.105	30	.200*	.960	30	.302
@2Complex	.162	30	.044	.945	30	.123
@3Easyuse	.058	30	.200*	.984	30	.925
@4Techsupport	.149	30	.085	.885	30	.004
@5Wellintegrated	.152	30	.075	.945	30	.124
@6Inconsistency	.081	30	.200*	.961	30	.332
@7Learnquickly	.113	30	.200*	.968	30	.481
@8Cumbersome	.175	30	.020	.892	30	.005
@9Confident	.084	30	.200*	.973	30	.620
@10Learnbefore	.094	30	.200*	.943	30	.111

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Appendix C5: HE++ Tools

A. Usability heuristics:

1 Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2 Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

3 User control and freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

4 Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5 Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place.

6 Recognition rather than recall

Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

7 Flexibility and efficiency of use

Accelerators may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

8 Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

9 Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

10 Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

B. A list of specific problem areas to ***focus*** on (see below). Research has found that most problems with websites fall into one of these areas. It may therefore be easier to spot a problem if you look for problems in these specific areas.

Problem areas

Area 1: Graphics (e.g. symbols, buttons, links, icons, maps)
Area 2: Information content
Area 3: Formatting and Layout (e.g. font size, white space, alignment)
Area 4: System efficiency and functionality (e.g. down load time)
Area 5: Navigation
Area 6: Wording (e.g. category names, language use)
Area 7: Help and error messages

Appendix C6: Phases 3-5: User-based usability testing task list.

For participants:

Tasks

1. You have a pet dog and need to know if it needs to be registered and microchipped in the Gold Coast. Are these mandatory in the Gold Coast? _____.

2. Parts of Gold Coast use recycled water. Is this water safe for you to drink? _____.

3. Gold Coast offers a lower price for water usage to some residents. Does this apply to you? _____.

4. You're worried that a property you would like to live in may be noisy since it is near the airport. Who do you contact for information about this type of noise? _____.

5. Who is eligible for free vaccinations? - _____.

6. How many beaches are located in the Gold Coast? _____.

7. You always wanted to have a beach wedding. How many people would you be able to invite if you had your wedding on one of Gold Coast's beaches? _____.

8. What breakfast restaurant is highly recommended and won an award? _____.

9. You love going shopping. How many shopping centres are there in the city? ____.

10. Your office will be in Robina. What city division is this in? _____.

11. Who is the Councillor representing your division? _____.

12. Your boss has asked you to put up a temporary billboard and banner advertisements of the business around Gold Coast. Do you need a license or permit for this? _____.

13. If you had a complaint to make to the council, what phone number would you need to call?
_____
__.

14. Your boss is planning to build another office building in the Gold Coast. He asked you go visit the Planning Advice Center to get some information. When is it open during the week?
_____.

15. You own a jet ski and would love to use it at Gold Coast‘s beaches. Is there a legislation for jet skiing? _____-
____.

Tasks- Optimal paths to task completion
Not given to participants

1. You have a pet dog and need to know if it needs to be registered and microchipped in the Gold Coast. Are these mandatory in the Gold Coast?
   _____.

   Community tab > Pets (change to: Animals, Companions, Four-Legs), Caring for Pets (change to: Animal Aid, Four-Leg Sympathy, Saving Companions), and Pet Laws and Registration (change to Animal Rights, Companion Jury, Four-Leg Rules).
   Dog Registration (change to: Woof Accounting), Cat Registration (change to: Meow Accounting).
   Registering Dogs and Cats (Change to: Animal Accounting, Companion Catalogue, Four-Legged Recording)

2. Parts of Gold Coast use recycled water. Is this water safe for you to drink?
   _____.

   Environment Tab > Waste & Recycling (Garbage and Rejuvenation), Waste & Recycling Initiatives (Garbage and Rejuvenation Advances, Dirt & Reprocessed Enterprises), Water & Sewerage (H2O & Septic Tanks, Liquid Dumps, Marine Dirt), Water Services (H2O Facilities, Liquid Accommodations, Marine Amenities), Sewerage & Recycled Water (Septic Tanks & Rejuvenated H20, Dumps & Invigorated Liquid, Liquid Dirt & Reprocessed Marine Fluids), Water & Sewerage Projects (H2O & Septic Tank Projects, Liquid and Dirt Assignments), Water Quality (Liquid Class, H2O Excellence) .

3. Gold Coast offers a lower price for water usage to some residents. Does this apply to you?
   _____.

   Council > Council Rates (Assembly Taxes, Jury Fees), Water Rates & Billing (H2O Taxes & Payments, Liquid Fees & Receipts, Marine Balances), Water Pricing (H2O Value, Liquid Costs, Marine Appraisal), Water Account Enquiries (H2O Bank Questions, Liquid Account Requests, Marine Tab Review), Online Water and Wastewater Rate Notice Enquiry (Web H20 and Dirty H2O Cost Advice, Internet Liquid and Junk Tax Questions, WWW. Fluid Garbage Fees Suggestions).

4. You‘re worried that a property you would like to live in may be noisy since it is near the airport. What contact information do you need to ask about this type of noise?
   _____
   ___.

Community > Community Concerns (Public Anxieties, Group Fears, Civic Worries), Neighbourhood Issues (District Disturbance, Area Problems, Zone Disputes)

5. Who is eligible for free vaccinations? -
_____.

Community > Health (Fitness, Strength, Shape), Community Health & Wellbeing (Neighbourhood Fitness & Happiness, Area Strength and Joy), Environmental Health (Surroundings and Fitness, Biosphere Cleanliness)

6. How many beaches are located in the Gold Coast?
_____.

The Gold Coast > Beaches & Foreshores (Sand and Cliffs, Seawater & Fjords), Gold Coast Beaches (Yellow Sand Water, Sparkling H2O, Community Water Outlets).

7. You always wanted to have a beach wedding. How many people would you be able to invite if you had your wedding on one of Gold Coast‘s beaches?
_____.

8. What breakfast restaurant is highly recommended and won an award?
_____
__.

The Gold Coast > Food, Wine, Dining (Milk, Red Water, and Nutrition; Sustenance and Alcohol, Provisions and Juice, Pots and Pans)

9. Imagine you love to shop. How many shopping centres are there in the city?
____.

The Gold Coast > Gold Coast Attractions (Yellow River Views, Waterway Sights), Attractions and Activities (Views and Do‘s, Fun Actions, Magnetisms and Happenings) and Shopping & Markets (Purchases and Dealers, Buy and Bargain, Deal and Dealt) Gold Coast Shopping (Yellow Water Purchase, H2O Deals, Looking for Something Special), Gourmet Shopping (Yum Foods, Good Food Deals, Where to go for Expensive Food), Markets (Cheap Food, Fresh Food, Real Food)

10. Your office will be in Robina. What city division is this in?
_____.

Council > Councillors & Divisions (People and Areas, Jury Members and Quarters, Workers and Land), Mayor & Councillor Profiles (Boss and Employees, Meet Them, Who‘s Running the Show), Council Divisions (Political Assembly, Assembly Areas, Jury Regions), Council Elections (Jury Dates, Choosing Members, City Voting)

11. Who is the Councillor representing your division?
_____.

12. Your boss has asked you to put up a temporary billboard and banner advertisements of the business around Gold Coast. Do you need a license or permit for this? _____.

Council and Online Services > Permits & Licencing (Commandments and Rules, Do‘s and Don‘ts, Law and Order)

13. If you had a complaint to make to the council, what phone number would you need to call?
_____
__.

Contact Council (View, Save, Talk)
Council and online services > Make a complaint (Judge, Help Needed, Be Heard)

14. Your boss is planning to build another office building in the Gold Coast. He asked you go visit the Planning Advice Center to get some information. When is it open during the week?
_____.

Planning and Building > Planning Enquiries (Thinking Questions, Organization Studies, Forecasting Explorations)

15. You own a jet ski and would love to use it at Gold Coast‘s beaches. Is there a legislation for jet skiing? _____-
____.

The Gold Coast > Sport & Recreation (Game and Rebirth, Diversion and Regeneration) Sports Clubs (Game Membership, Leisure Center, Other Clubs), Sport & Leisure Activities (Game and Fun Jobs, Diversion and Ease Events, Hobby and Vacation Things)

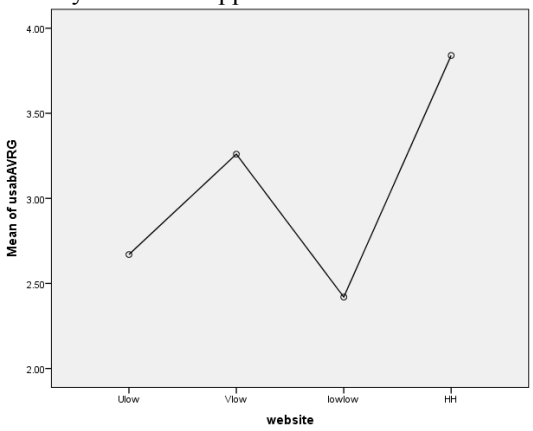
Appendix C7: Phase 3 Results

Table 4. The average results per task in the first user-centred usability test.

Task	Success	Hints	Clicks	Hovers	Time
1	70%	0	3.5	2.2	01:36.4
2	80%	0.3	5.5	4.5	02:23.2
3	90%	0.8	7.1	4.9	02:52.9
4	90%	0.8	4.3	6.3	02:38.8
5	100%	0.6	2.2	2.5	02:10.5
6	100%	0	1.4	2.3	00:36.9
7	100%	0.4	2.7	3.2	01:19.8
8	100%	0.1	3.1	1.6	01:58.1
9	100%	0	2	1.4	00:30.1
10	90%	0.4	2.4	4.8	01:47.9
11	100%	0.1	2.5	1	00:37.3
12	90%	0.9	4.7	8.8	02:54.1
13	90%	0.2	3.4	2.5	01:12.6
14	60%	1.1	7.5	4.9	03:13.6
15	100%	0.4	7.1	4.4	02:25.7

Appendix C8: Phase 4 Results

usability and visual appeal verification results



usabAVRG

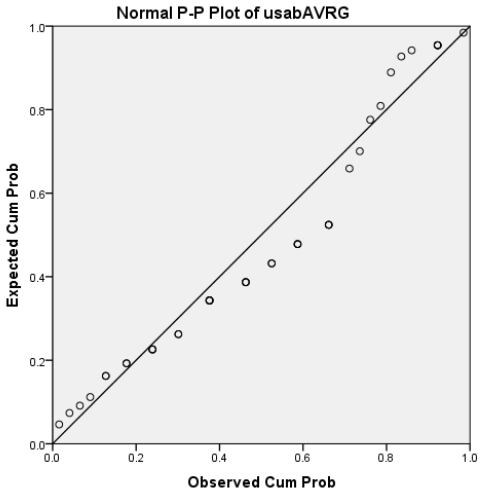
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Ulow	10	2.6700	.43982	.13908	2.3554	2.9846
Vlow	10	3.2600	.89094	.28174	2.6227	3.8973
lowlow	10	2.4200	.51597	.16316	2.0509	2.7891
HH	10	3.8400	.78344	.24775	3.2796	4.4004

Test of Homogeneity of Variances

usabAVRG

Levene Statistic	df1	df2	Sig.
3.122	3	36	.038

Variances are not equal at alpha of 0.05 but are at alpha 0.01. Given the small number of participants in each condition, the variance would most likely be equally distributed for each of the website versions. Therefore, we conclude that variance is homogenous.



The normal probability plot appears to be normally distributed for the usability ratings. We can therefore assume that the assumption holds for normality.

ANOVA

usabAVRG

	Sum of Squares	df	Mean Square	F	Sig.

Between Groups	12.095	3	4.032	8.637	.000
Within Groups	16.805	36	.467		
Total	28.900	39			

The anova table shows that the usability levels differ between website versions, at alpha <0.01.

Post Hoc Tests
Multiple Comparisons

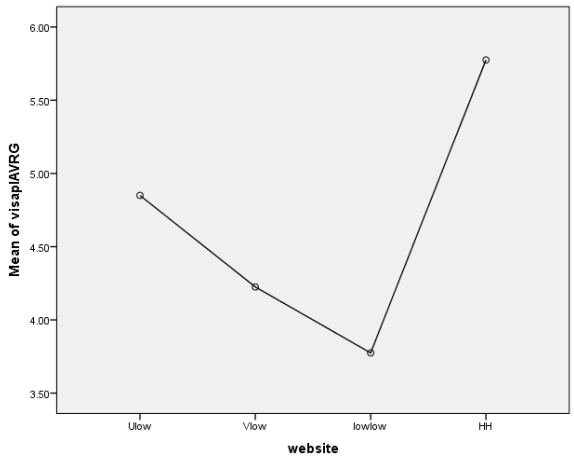
	(I) website	(J) website	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	95% Confidence Interval
						Lower Bound	Upper Bound
Tukey HSD	Ulow	Vlow	-.59000	.30555	.233	-1.4129	.2329
		lowlow	.25000	.30555	.845	-.5729	1.0729
		HH	-1.17000*	.30555	.003	-1.9929	-.3471*
	Vlow	Ulow	.59000	.30555	.233	-.2329	1.4129
		lowlow	.84000*	.30555	.044	.0171	1.6629*
		HH	-.58000	.30555	.247	-1.4029	.2429
	lowlow	Ulow	-.25000	.30555	.845	-1.0729	.5729
		Vlow	-.84000*	.30555	.044	-1.6629	-.0171*
		HH	-1.42000*	.30555	.000	-2.2429	-.5971*
	HH	Ulow	1.17000*	.30555	.003	.3471	1.9929*
		Vlow	.58000	.30555	.247	-.2429	1.4029
		lowlow	1.42000*	.30555	.000	.5971	2.2429*
Bonferroni	Ulow	Vlow	-.59000	.30555	.368	-1.4431	.2631
		lowlow	.25000	.30555	1.000	-.6031	1.1031
		HH	-1.17000*	.30555	.003	-2.0231	-.3169*
	Vlow	Ulow	.59000	.30555	.368	-.2631	1.4431
		lowlow	.84000	.30555	.056	-.0131	1.6931
		HH	-.58000	.30555	.394	-1.4331	.2731
	lowlow	Ulow	-.25000	.30555	1.000	-1.1031	.6031
		Vlow	-.84000	.30555	.056	-1.6931	.0131
		HH	-1.42000*	.30555	.000	-2.2731	-.5669*
	HH	Ulow	1.17000*	.30555	.003	.3169	2.0231*
		Vlow	.58000	.30555	.394	-.2731	1.4331
		lowlow	1.42000*	.30555	.000	.5669	2.2731*

*. The mean difference is significant at the 0.05 level.

The results are as follows for visual appeal:

visaplAVRG

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Ulow	10	4.8500	1.19140	.37676	3.9977	5.7023
Vlow	10	4.2250	1.40164	.44324	3.2223	5.2277
lowlow	10	3.7750	1.50670	.47646	2.6972	4.8528
HH	10	5.7750	.70168	.22189	5.2730	6.2770



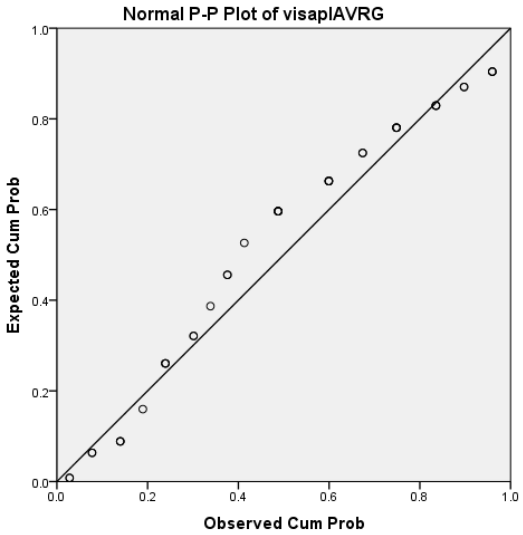
Descriptives

Test of Homogeneity of Variances

visaplAVRG

Levene Statistic	df1	df2	Sig.
2.064	3	36	.122

Since p=.122 > 0.05, the null hypothesis cannot be rejected with 95% confidence. Therefore, the assumption for equal variance is held for visual appeal ratings.



The normal probability plot shows that visual appeal is normally distributed.

ANOVA

visaplAVRG

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	22.517	3	7.506	4.885	.006
Within Groups	55.319	36	1.537		
Total	77.836	39			

The anova table shows that the visual appeal levels differ between website versions, at alpha <0.01.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: visaplAVRG

	(I) website	(J) website	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	95% Confidence Interval
						Lower Bound	Upper Bound
Tukey HSD	Ulow	Vlow	.62500	.55437	.675	-.8680	2.1180
		lowlow	1.07500	.55437	.230	-.4180	2.5680
		HH	-.92500	.55437	.355	-2.4180	.5680
	Vlow	Ulow	-.62500	.55437	.675	-2.1180	.8680
		lowlow	.45000	.55437	.849	-1.0430	1.9430
		HH	-1.55000*	.55437	.039	-3.0430	-.0570*
	lowlow	Ulow	-1.07500	.55437	.230	-2.5680	.4180
		Vlow	-.45000	.55437	.849	-1.9430	1.0430
		HH	-2.00000*	.55437	.005	-3.4930	-.5070*
	HH	Ulow	.92500	.55437	.355	-.5680	2.4180
		Vlow	1.55000*	.55437	.039	.0570	3.0430*
		lowlow	2.00000*	.55437	.005	.5070	3.4930*
Bonferroni	Ulow	Vlow	.62500	.55437	1.000	-.9228	2.1728
		lowlow	1.07500	.55437	.362	-.4728	2.6228
		HH	-.92500	.55437	.623	-2.4728	.6228
	Vlow	Ulow	-.62500	.55437	1.000	-2.1728	.9228
		lowlow	.45000	.55437	1.000	-1.0978	1.9978
		HH	-1.55000*	.55437	.049	-3.0978	-.0022*
	lowlow	Ulow	-1.07500	.55437	.362	-2.6228	.4728
		Vlow	-.45000	.55437	1.000	-1.9978	1.0978
		HH	-2.00000*	.55437	.006	-3.5478	-.4522*
	HH	Ulow	.92500	.55437	.623	-.6228	2.4728
		Vlow	1.55000*	.55437	.049	.0022	3.0978*
		lowlow	2.00000*	.55437	.006	.4522	3.5478*

*. The mean difference is significant at the 0.05 level.

Appendix C9: Preliminary Study 4 Usability Manipulations

These were changed, everywhere the links occurred on the website.

16. You have a pet dog and need to know if it needs to be registered and microchipped in the Gold Coast. Are these mandatory in the Gold Coast?
_____.

Community tab > Pets (change to: Animals, Companions, Four-Legs), Caring for Pets (change to: Animal Aid, Four-Leg Sympathy, Saving Companions), and Pet Laws and Registration (change to Animal Rights, Companion Jury, Four-Leg Rules).
Dog Registration (change to: Woof Accounting), Cat Registration (change to: Meow Accounting).
Registering Dogs and Cats (Change to: Animal Accounting, Companion Catalogue, Four-Legged Recording)

17. Parts of Gold Coast use recycled water. Is this water safe for you to drink?
_____.

Environment Tab > Waste & Recycling (Garbage and Rejuvenation), Waste & Recycling Initiatives (Garbage and Rejuvenation Advances, Dirt & Reprocessed Enterprises), Water & Sewerage (H2O & Septic Tanks, Liquid Dumps, Marine Dirt), Water Services (H2O Facilities, Liquid Accommodations, Marine Amenities), Sewerage & Recycled Water (Septic Tanks & Rejuvenated H20, Dumps & Invigorated Liquid, Liquid Dirt & Reprocessed Marine Fluids), Water & Sewerage Projects (H2O & Septic Tank Projects, Liquid and Dirt Assignments), Water Quality (Liquid Class, H2O Excellence) .

18. Gold Coast offers a lower price for water usage to some residents. Does this apply to you?
_____.

Council > Council Rates (Assembly Taxes, Jury Fees), Water Rates & Billing (H2O Taxes & Payments, Liquid Fees & Receipts, Marine Balances), Water Pricing (H2O Value, Liquid Costs, Marine Appraisal), Water Account Enquiries (H2O Bank Questions, Liquid Account Requests, Marine Tab Review), Online Water and Wastewater Rate Notice Enquiry (Web H20 and Dirty H2O Cost Advice, Internet Liquid and Junk Tax Questions, WWW. Fluid Garbage Fees Suggestions).

19. You‘re worried that a property you would like to live in may be noisy since it is near the airport. What contact information do you need to ask about this type of noise?
_____
___.

Community > Community Concerns (Public Anxieties, Group Fears, Civic Worries), Neighbourhood Issues (District Disturbance, Area Problems, Zone Disputes)

20. Who is eligible for free vaccinations? -
_____.

Community > Health (Fitness, Strength, Shape), Community Health & Wellbeing (Neighbourhood Fitness & Happiness, Area Strength and Joy), Environmental Health (Surroundings and Fitness, Biosphere Cleanliness)

21. How many beaches are located in the Gold Coast?
_____.

The Gold Coast > Beaches & Foreshores (Sand and Cliffs, Seawater & Fjords), Gold Coast Beaches (Yellow Sand Water, Sparkling H2O, Community Water Outlets).

22. You always wanted to have a beach wedding. How many people would you be able to invite if you had your wedding on one of Gold Coast‘s beaches?
_____.

23. What breakfast restaurant is highly recommended and won an award?
_____
__.

The Gold Coast > Food, Wine, Dining (Milk, Red Water, and Nutrition; Sustenance and Alcohol, Provisions and Juice, Pots and Pans)

24. Imagine you love to shop. How many shopping centres are there in the city?
____.

The Gold Coast > Gold Coast Attractions (Yellow River Views, Waterway Sights), Attractions and Activities (Views and Do‘s, Fun Actions, Magnetisms and Happenings) and Shopping & Markets (Purchases and Dealers, Buy and Bargain, Deal and Dealt) Gold Coast Shopping (Yellow Water Purchase, H2O Deals, Looking for Something Special), Gourmet Shopping (Yum Foods, Good Food Deals, Where to go for Expensive Food), Markets (Cheap Food, Fresh Food, Real Food)

25. Your office will be in Robina. What city division is this in?
_____.

Council > Councillors & Divisions (People and Areas, Jury Members and Quarters, Workers and Land), Mayor & Councillor Profiles (Boss and Employees, Meet Them, Who‘s Running the Show), Council Divisions (Political Assembly, Assembly Areas, Jury Regions), Council Elections (Jury Dates, Choosing Members, City Voting)

26. Who is the Councillor representing your division?
_____.

27. Your boss has asked you to put up a temporary billboard and banner advertisements of the business around Gold Coast. Do you need a license or permit for this? _____.

Council and Online Services > Permits & Licencing (Commandments and Rules, Do‘s and Don‘ts, Law and Order)

28. If you had a complaint to make to the council, what phone number would you need to call?
_____
__.

Contact Council (View, Save, Talk)
Council and online services > Make a complaint (Judge, Help Needed, Be Heard)

29. Your boss is planning to build another office building in the Gold Coast. He asked you go visit the Planning Advice Center to get some information. When is it open during the week?
_____.

Planning and Building > Planning Enquiries (Thinking Questions, Organization Studies, Forecasting Explorations)

30. You own a jet ski and would love to use it at Gold Coast‘s beaches. Is there a legislation for jet skiing? _____-
____.

The Gold Coast > Sport & Recreation (Game and Rebirth, Diversion and Regeneration) Sports Clubs (Game Membership, Leisure Center, Other Clubs), Sport & Leisure Activities (Game and Fun Jobs, Diversion and Ease Events, Hobby and Vacation Things)

Appendix C10: Preliminary Study 5 Usability Re-manipulations

council = {"Board", "Assembly", "Committee", "Congress", "Politics", "Government", "Law", "Jury"};
theGoldCoast = { "The City", "Streets", "Miscellaneous", "About", "Life" };
community = { "People", "Public", "Us", "Neighbourhood", "Open", "Civic" };
planningAndBuilding = { "Development", "Infrastructure", "Brick by Brick" };
environment = { "Parks and Beaches", "Nature", "Flora and Fauna", "Setting", "Surroundings", "Atmosphere", "Biosphere" };
business = { "Job", "Stocks", "Money", "Corporate", "Professional", "Commerce " };
councilAndOnlineServices = { "Online Acts", "Help", "Rules", "Services" };
pets = { "Animals", "Companions", "Four-Legs"};
public static string[] caringForPets = { "Animal Aid", "Four-Leg Sympathy", "Saving Companions" };
petLawsRegistrations = { "Animal Rights", "Companion Jury", "Four-Leg Rules" };
public static string[] registeringDogsAndCats = { "Animal Accounting", "Companion Catalogue", "Four-Legged Recording" };
wasteRecyclingInitiatives = { "Garbage & Rejuvination Advances", "Dirt & Reprocessed Enterprises"};
waterSewerage = { "H2O & Septic Tanks", "Liquid Dumps", "Marine Dirt" };
public static string[] waterServices = { "H2O Facilities", "Liquid Accommodations", "Marine Amenities" };
sewerageRecycledWater = { "Septic Tanks & Rejuvenated H2O", "Dumps & Invigorated Liquid", "Liquid Dirt & Reprocessed Marine Fluids" };
waterSewerageProjects = { "H2O & Septic Tank Projects", "Liquid & Dirt Assignments"};
waterQuality = { "Liquid Class", "H2O Excellence"};
councilRates = { "Assembly Taxes", "Jury Fees" };
waterRatesBilling = { "H2O Taxes & Payments", "Liquid Fees & Receipts", "Marine Balances" };
waterPricing = { "H2O Value", "Liquid Costs", "Marine Appraisal" };
waterAccountInquiries = { "H2O Bank Questions", "Liquid Account Requests", "Marine Tab Review" };
onlineWaterWasteWaterRateNoticeEnquiry = { "Web H2O and Dirty H2O Cost Advice", "Internet Liquid and Junk Tax Questions", "WWW. Fluid Garbage Fees Suggestions" };
communityConcerns = { "Public Anxieties", "Group Fears", "Civic Worries" };
neighbourhoodIssues = { "Disrict Disturbance", "Area Problems", "Zone Disputes" };
health = { "Fitness", "Strength", "Shape" };
healthCommunityWellbeing = { "Neighbourhood Fitness & Happiness", "Area Strength and Joy" };
environmentalHealth = { "Surroundings and Fitness", "Biosphere Cleanliness", };
beachesForeshores = { "Sand and Cliffs", "Seawater & Fjords", "Shape" };
goldCoastBeaches ={ "Yellow Sand Water","Sparkling H2O", "Community Water Outlets"};
foodWineDining = { "Milk, Red Water and Nutrition", "Sustenance and Alcohol", "Provisions and Juice", "Pots and Pans" };
goldCoastAttracitions = { "Yellow River Views", "Waterway Sights"};

public static string[] attractionsActivities = { "Views and Do's", "Fun Actions", "Magnetisms and Happenings" };
shoppingMarkets = { "Purchases and Dealers", "Buy and Bargain", "Deal and Dealt" };
goldCoastShopping = { "Yellow Water Purchase", "H2O Deals", "Looking for something Special" };
gourmetShopping ={"Yum Foods", "Good Food Deals", "Where to go for Expensive Food"};
markets = { "Cheap Food", "Fresh Food", "Real Food" };
councillorsDivisions = { "People and Areas", "Jury Members and Quarters", "Workers and Land" };
mayorCouncillorProfiles = { "Boss and Employees", "Meet Them", "Who's Running the Show" };
councilDivisions = { "Political Assembly", "Assembly Areas", "Jury Regions" };
councilElections = { "Jury Dates", "Choosing Members", "City Voting" };
permitsLicencing = { "Commandments and Rules", "Do's and Dont's", "Law and Order" };
contactCouncil = { "View", "Save", "Talk" };
makeComplaint = { "Judge", "Help Needed", "Be Heard" };
planningEnquiries = { "Thinking Questions", "Organization Studies", "Forecasting Explorations" };
sportRecreation = { "Game and Rebirth", "Diversion and Regeneration"};
sportClubs = { "Game Membership", "Leisure Center", "Other Clubs"};
sportLeisureActivities = { "Game and Fun Jobs", "Diversion and Ease Events", "Hobby and Vacation Things" };

#region remove breadcrumbs
#region shuffle menu items

Appendix D: Main Study 1

Appendix D1: Shorter Task List

Tasks

Please use the website to find the information listed below. Briefly write the answer to each question when you think you have enough information. When you complete a task please proceed immediately to the next task.

1. Parts of Gold Coast use recycled water. Is this water safe for you to drink? _____.

2. Gold Coast offers a lower price for water usage to some residents. Does this apply to you? _____.

3. You're worried that a property you would like to live in may be noisy since it is near the airport. Who do you contact for information about this type of noise? _____.

4. Who is eligible for free vaccinations? - _____.

5. How many beaches are located in the Gold Coast? _____.

6. What breakfast restaurant is highly recommended and won an award? _____.

7. You love going shopping. How many shopping centres are there in the city? ____.

8. Your office will be in Robina. What city division is this in? _____.

9. Who is the Councillor representing your division? _____.

10. You own a jet ski and would love to use it at Gold Coast's beaches. Is there a legislation for jet skiing? _____.

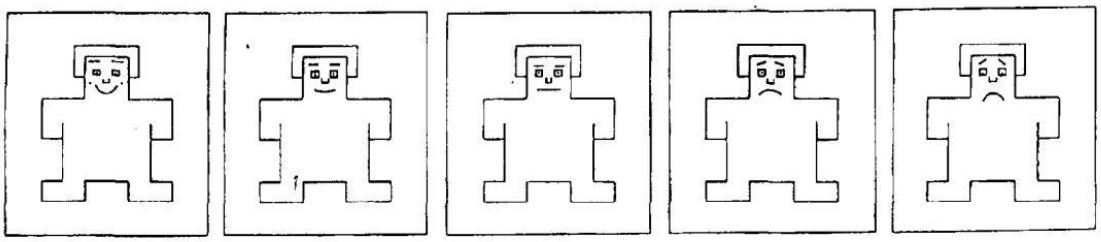
Appendix D2: The mood questionnaire used called SAM.

Mood Questionnaire

Participant number:_____

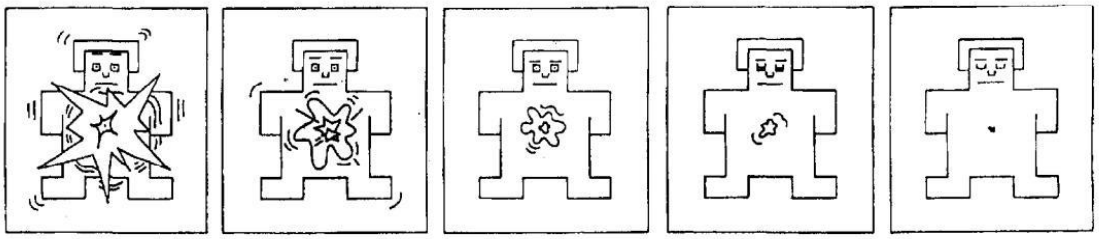
Please circle the picture that best suits your current mood.

Happy Sad



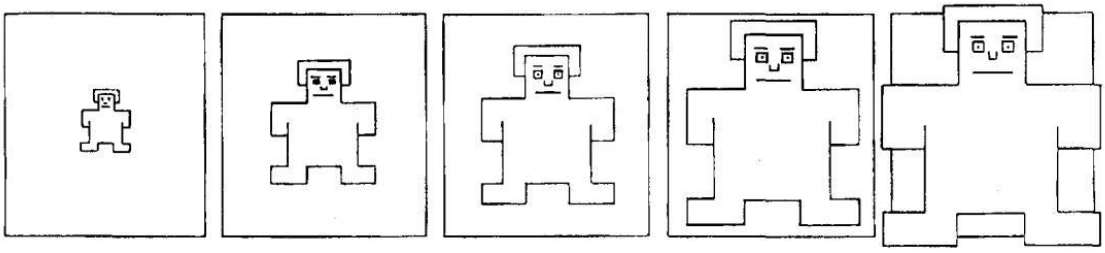
Please circle the picture that best suits your current mood.

Anxious Calm



Please circle the picture that best suits your current mood.

Insecure Confident



Appendix D3: Textual Expectations Main 1 and 2

Setting Expectations Too High (HH)

Welcome to Gold Coast, Australia‘s greatest travel destination! Your boss was delighted with your work and decided to promote you to senior manager of the company in Gold Coast. You are bound to love it there and the job‘s pay is great. Before you start packing and head off, you‘re going to check the city‘s city council website out, to get some information which will help you get ready for the move. Recent surveys have found that the website is as beautiful as the gorgeous city. People are finding it incredibly easy to use, and they all recommended it to their family and friends. The developers created a professional masterpiece and the website won an award for best city council website in Australia in 2013.

Setting Expectations Too Low (LL)

Welcome to Gold Coast, a big city in Australia. Your boss was not entirely happy with your recent work and demoted you to the Gold Coast branch of the company. You‘re really not happy about going to the Gold Cost and you probably won‘t like it. Before you start packing and head off, you‘re going to check the city council‘s website out, to get some information which will help you pack and get ready for your move. Recent surveys have found that the website was really ugly. People found it incredibly hard to use, and they would never recommend it to their family or friends. The developers created a professional disaster, and the website was voted as worst city council website in Australia in 2013.

Setting No Expectation (Control)

You have some time off from work coming up and you are thinking to take a vacation on the Gold Coast. The city has a population size of approximately 590,000, and is 414.3 km^2 large. The community is varied, with people of all ages and nationalities. English is the main language spoken, although resources are also offered in 13 other languages. The maximum temperature is between 25◦C and 17◦C, with an annual rainfall of 12cm. There are many activities to do in the city, including tours, surfing, and hiking among others. Before you start packing and head off, you are going to check the city‘s website out, to get some information which will help you pack. The programmers made a website representative of the city.

Appendix D4: Main 1 Results

Participants Sixty Swinburne University student volunteers participated in the study. Out of these, ten were randomly assigned to each of the six conditions. 48 of these students were aged 18-30 and 12 aged 31 and above. Thirty-nine were male and 21 female, with 28 born in an English speaking country and 40 speaking it frequently at home. Thirty-five out of the sixty were undergraduate students, 21 masters, and four PhD students. Forty-seven were studying computer science, three design, two each for games development, arts, psychology, and one each for engineering, business, biomedical engineering, and astrophysics and supercomputing.

Stats

HuHv website

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of PreUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.258	Retain the null hypothesis.
2	The distribution of PostUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.967	Retain the null hypothesis.
3	The distribution of PreVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.251	Retain the null hypothesis.
4	The distribution of PostVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.727	Retain the null hypothesis.
5	The distribution of clicks is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.030	Reject the null hypothesis.
6	The distribution of hovers is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.938	Retain the null hypothesis.
7	The distribution of time is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.889	Retain the null hypothesis.
8	The distribution of pass is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.661	Retain the null hypothesis.
9	The distribution of PreMood is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.746	Retain the null hypothesis.
10	The distribution of PostMood is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.800	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

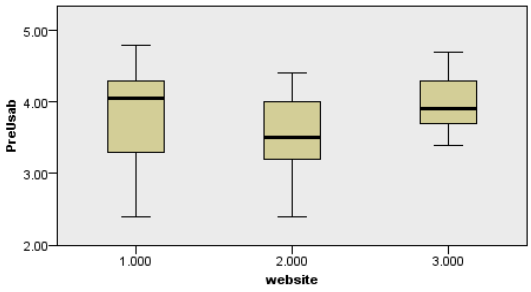
Pairwise Comparisons of website

Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
2.000-3.000	-4.200	3.930	-1.069	.285	.856
2.000-1.000	10.350	3.930	2.634	.008	.025
3.000-1.000	6.150	3.930	1.565	.118	.353

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Independent-Samples Kruskal-Wallis Test



Total N	30
Test Statistic	2.709
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.258

1. The test statistic is adjusted for ties.
2. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

LuLv website

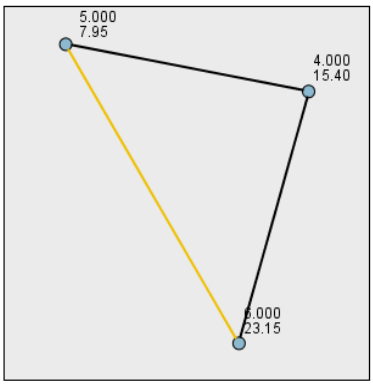
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of PreUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.261	Retain the null hypothesis.
2	The distribution of PostUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.386	Retain the null hypothesis.
3	The distribution of PreVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.001	Reject the null hypothesis.
4	The distribution of PostVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.019	Reject the null hypothesis.
5	The distribution of clicks is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.095	Retain the null hypothesis.
6	The distribution of hovers is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.295	Retain the null hypothesis.
7	The distribution of time is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.642	Retain the null hypothesis.
8	The distribution of pass is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.757	Retain the null hypothesis.
9	The distribution of PreMood is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.902	Retain the null hypothesis.
10	The distribution of PostMood is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.061	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

PreVis

Pairwise Comparisons of website



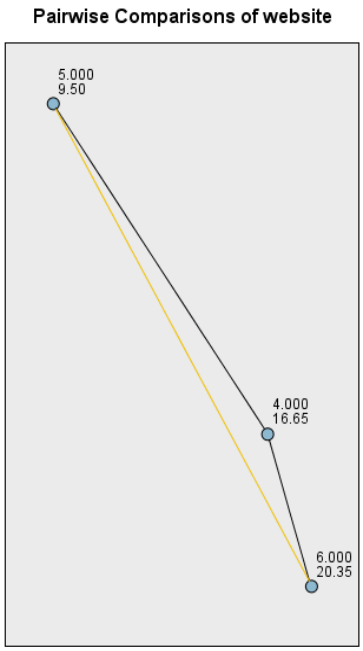
Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
5.000-4.000	7.450	3.916	1.902	.057	.171
5.000-6.000	-15.200	3.916	-3.882	.000	.000
4.000-6.000	-7.750	3.916	-1.979	.048	.143

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

View: Pairwise Comparisons | Test: Kruskal-Wallis | Field(s): PreVis * website(Test 3) | Layout

Post Vis



Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
5.000-4.000	7.150	3.929	1.820	.069	.206
5.000-6.000	-10.850	3.929	-2.761	.006	.017
4.000-6.000	-3.700	3.929	-.942	.346	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

View: Pairwise Comparisons Test: Kruskal-Wallis Field(s): PostVis * website(Test 4) Layout

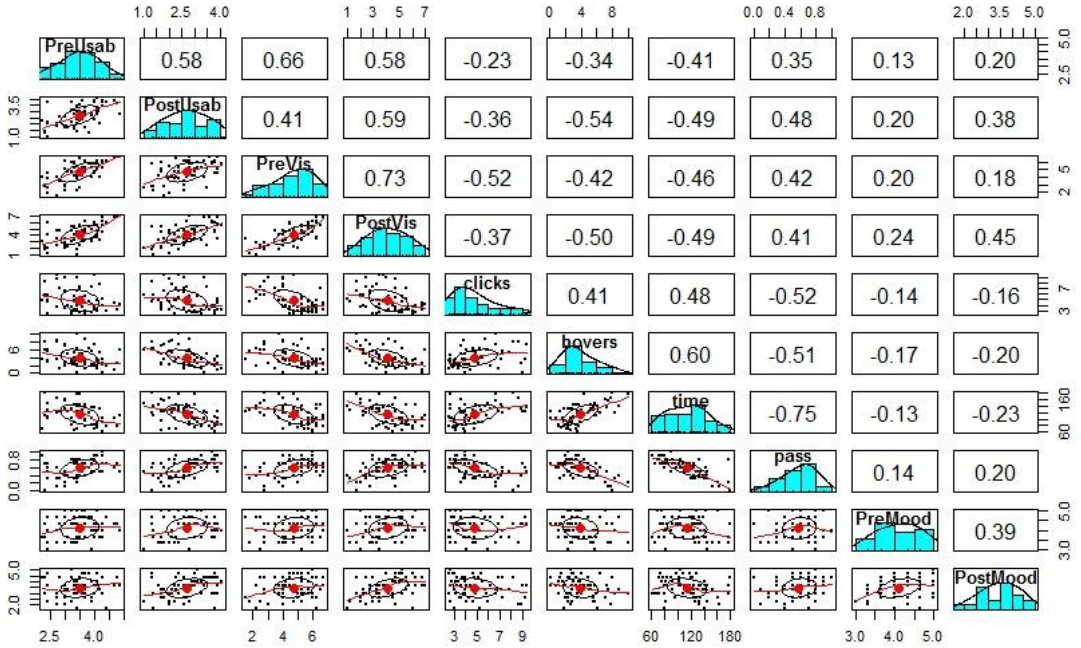


Figure 5. Spearman correlations for all variables and conditions in Main Study 1.0.

table for both HuHv and LuLv . The correlations in Figure 5 are rounded to hundredth, while the ones in Table 5 are slightly more accurate due to the additional decimal point.

Upon separating the data between the two website versions, HuHv and LuLv, the Spearman Correlations can be seen in Tables 6 and 7 respectively. In these tables, the columns and rows are both the measured variables: pre-use perceived usability (PreUsab), post-use perceived usability (PostUsab), pre-use visual appeal (PreVis), post-use visual appeal (PostVis), the average number of clicks, the average number of hovers, the average completion time per task, the proportion of passed tasks, pre-use mood (PreMood), and post-use mood (PostMood).

Table 5. Spearman correlations for all website conditions.

	PostUsab	PreVis	PostVis	Clicks	Hovers	Time	Pass	PreMood	PostMood
PreUsab	.578**	.657**	.582**	-.229	-.344**	-.411**	.348**	.128	.196
PostUsab	-	.411**	.585**	-.363**	-.543**	-.492**	.478**	.196	.378**
PreVis		-	.733**	-.516**	-.417**	-.461**	.416**	.200	.175
PostVis			-	-.371**	-.502**	-.485**	.409**	.244	.448**
Clicks				-	.407**	.477**	-.522**	-.135	-.164
Hovers					-	.597**	-.509**	-.174	-.202**
Time						-	-.747**	-.132	-.228
Pass							-	.144	.198**
PreMood								-	.387**

** Significant at 0.01 (2-tailed).

* Significant at 0.05 (2-tailed).

Table 6. Spearman Correlations for the HuHv website conditions.

	PostUsab	PreVis	PostVis	Clicks	Hovers	Time	Pass	PreMood	PostMood
PreUsab	.535**	.498**	.484**	.310	-.449*	-.397*	.050	-.034	.082
PostUsab	-	-.012	.563**	.132	-.199	-.187	.043	-.079	.387*
PreVis		-	.405*	.198	-.339	-.036	-.194	.190	-.182
PostVis			-	.156	-.240	-.033	-.130	.202	.417*
Clicks				-	-.064	-.178	.012	-.109	-.061
Hovers					-	.174	.244	.086	.095
Time						-	-.354	.135	.153

Pass	-	-.088	-.113
PreMood		-	.333

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

Table 7. *Spearman Correlations for the LuLv website conditions.*

	PostUsab	**PreVis**	**PostVis**	**Clicks**	**Hovers**	**Time**	**Pass**	**PreMood**	**PostMood**
PreUsab	.485**	.615**	.498**	-.003	.015	-.081	.277	.298	.245
PostUsab	-	.246	.411*	.064	-.437*	-.262	.278	.306	.248
PreVis		-	.634**	-.359	.072	-.086	.106	.187	.346
PostVis			-	.104	-.303	-.236	.240	.225	.452*
Clicks				-	-.102	.265	.035	.027	-.089
Hovers					-	.377*	-.423*	-.206	-.274
Time						-	-.729**	-.289	-.310
Pass							-	.294	.223
PreMood								-	.364*

** Significant at 0.01 (2-tailed).
* Significant at 0.05 (2-tailed).

Appendix D5: Main 2

Script for Confederate

Setting Expectations High (HH)
Instructions: Happily come into the room after having finished the test, smile and be cheerful. Say the following:
Wow, that was fun! That was a fantastic website. I didn't realise that government websites could be that good. It was a really good looking website! I loved the colours. I totally zoomed through the questions so quickly, and managed to complete every single one. It was easy, you'll do great.

Setting Expectations Low (LL)
Instructions: Sadly come into the room after having finished the test, frown and be annoyed. Say the following:
Wow, that was terrible! That was a really crappy website. I didn't realise that government websites could be that bad. It was so ugly too! I hated the colours. It took ages to do all of the questions, and I think I only completed one of them. It was so hard, good luck!

Appendix D6: Main 2 Results

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of hovers is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
2	The distribution of hovers is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.655	Retain the null hypothesis.
3	The distribution of click is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
4	The distribution of click is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.008	Reject the null hypothesis.
5	The distribution of time is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
6	The distribution of time is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.004	Reject the null hypothesis.
7	The distribution of pass is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
8	The distribution of pass is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.017	Reject the null hypothesis.
9	The distribution of previs is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
10	The distribution of previs is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.335	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

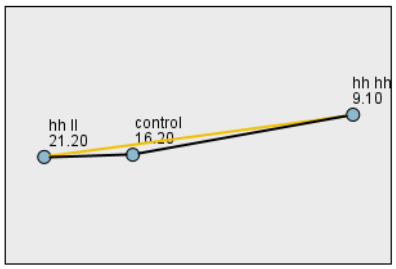
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
11	The distribution of preusab is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
12	The distribution of preusab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.
13	The distribution of postVis is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
14	The distribution of postVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.010	Reject the null hypothesis.
15	The distribution of PostUsab is the same across categories of website.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
16	The distribution of PostUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Clicks

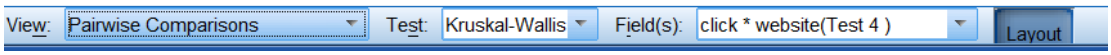
Pairwise Comparisons of website



Each node shows the sample average rank of website.

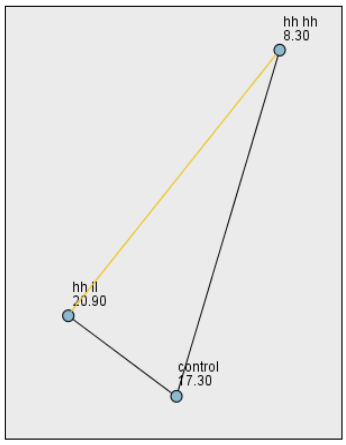
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh hh-control	-7.100	3.926	-1.809	.071	.212
hh hh-hh ll	-12.100	3.926	-3.082	.002	.006
control-hh ll	5.000	3.926	1.274	.203	.608

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.



Time

Pairwise Comparisons of website



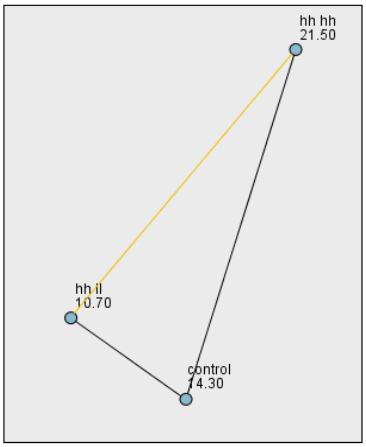
Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh hh-control	-9.000	3.937	-2.286	.022	.067
hh hh-hh ll	-12.600	3.937	-3.200	.001	.004
control-hh ll	3.600	3.937	.914	.361	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Pass

Pairwise Comparisons of website



Each node shows the sample average rank of website.

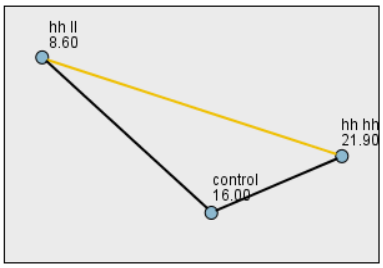
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh ll-control	-3.600	3.864	-.932	.352	1.000
hh ll-hh hh	10.800	3.864	2.795	.005	.016
control-hh hh	7.200	3.864	1.863	.062	.187

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

View: Pairwise Comparisons | Test: Kruskal-Wallis | Field(s): pass * website(Test 8) | Layout

PReUSab

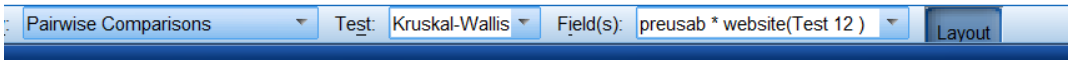
Pairwise Comparisons of website



Each node shows the sample average rank of website.

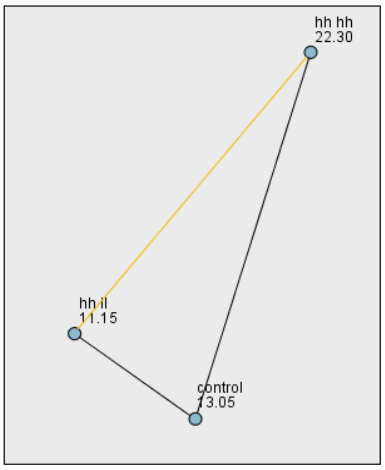
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh ll-control	-7.400	3.921	-1.887	.059	.177
hh ll-hh hh	13.300	3.921	3.392	.001	.002
control-hh hh	5.900	3.921	1.505	.132	.397

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.



PostVis

Pairwise Comparisons of website



Each node shows the sample average rank of website.

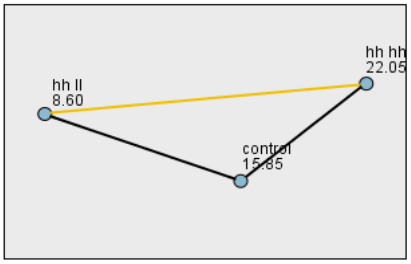
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh ll-control	-1.900	3.913	-.486	.627	1.000
hh ll-hh hh	11.150	3.913	2.850	.004	.013
control-hh hh	9.250	3.913	2.364	.018	.054

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

View: Pairwise Comparisons | Test: Kruskal-Wallis | Field(s): postVis * website(Test 14) | Layout

PostUsab

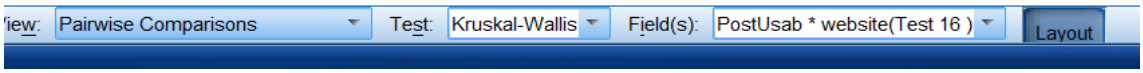
Pairwise Comparisons of website



Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
hh ll-control	-7.250	3.911	-1.854	.064	.191
hh ll-hh hh	13.450	3.911	3.439	.001	.002
control-hh hh	6.200	3.911	1.585	.113	.339

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.



Appendix E: Main Study 3

Appendix E1: Main Study 3 Script for Confederate

Setting HL
Wow, that was different! That was a really easy to use website. I totally zoomed through the questions so quickly, and managed to complete every single one. It was so ugly though! I hated the colours, and the pictures were ugly.

Setting LH
Wow, that was different! That was a really hard to use website. It took ages to do all of the questions, and I think I only completed one of them. It was a really good looking website! I loved the colours, and the pictures were beautiful.

Appendix E2: Main Study 3 Written Expectations

HL

Welcome to Gold Coast, a big city in Australia. You recently got a job there and will be moving quite soon. Before you start packing and head off, you‘re going to check the city‘s city council website out, to get some information which will help you get ready for the move. Recent surveys have found that the website was really ugly. The colours are unprofessional and jarring. However, people are finding it incredibly easy to use, especially given that it is a government website. They said it was easier than doing their taxes.

LH

Welcome to Gold Coast, a big city in Australia. You recently got a job there and will be moving quite soon. Before you start packing and head off, you‘re going to check the city‘s city council website out, to get some information which will help you get ready for the move. Recent surveys have found that the website is as beautiful as the gorgeous city. The colours are very professional and flattering. However, people are finding it incredibly hard to use, even for a government website. They said it was harder than doing their taxes.

Appendix E3: Main Study 3 Results

The Easy/Ugly website = no sig diff

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of hovers is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.648	Retain the null hypothesis.
2	The distribution of clicks is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.219	Retain the null hypothesis.
3	The distribution of time is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.678	Retain the null hypothesis.
4	The distribution of pass is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.467	Retain the null hypothesis.
5	The distribution of preVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.418	Retain the null hypothesis.
6	The distribution of preUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.231	Retain the null hypothesis.
7	The distribution of PostVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.806	Retain the null hypothesis.
8	The distribution of PostUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.502	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

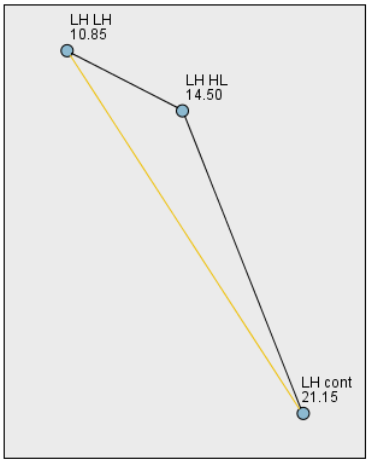
The Hard/Pretty website

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of hovers is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.518	Retain the null hypothesis.
2	The distribution of clicks is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.196	Retain the null hypothesis.
3	The distribution of time is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.823	Retain the null hypothesis.
4	The distribution of pass is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.986	Retain the null hypothesis.
5	The distribution of preVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.413	Retain the null hypothesis.
6	The distribution of preUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.029	Reject the null hypothesis.
7	The distribution of PostVis is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.481	Retain the null hypothesis.
8	The distribution of PostUsab is the same across categories of website.	Independent-Samples Kruskal-Wallis Test	.624	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

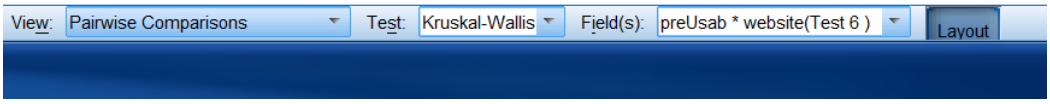
Pairwise Comparisons of website



Each node shows the sample average rank of website.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
LH LH-LH HL	3.650	3.921	.931	.352	1.000
LH LH-LH cont	-10.300	3.921	-2.627	.009	.026
LH HL-LH cont	-6.650	3.921	-1.696	.090	.270

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.



Thesis-Based Publications to Date

Stojmenovic, M. (2012). Understanding appropriateness of websites on different screen sized devices through an investigation of aesthetics and usability, *The 24th ACM Australian Computer-Human Interaction Conference OZCHI'12*, November 26–30, Melbourne, Victoria, Australia, pp. 4.

Stojmenovic, M., Pilgrim, C., & Lindgaard, G. (2014). Perceived and Objective Usability and Visual Appeal in a Website Domain with a Less Developed Mental Model. *The 26th ACM Australian Computer-Human Interaction Conference OZCHI'14*, December 2-5, 2014, Sydney, NSW, Australia, pp. 316-323.