J Supercomput (2019) 75:1783–1830
https://doi.org/10.1007/s11227-018-2313-x

The influence of textual and verbal word-of-mouth on website usability and visual appeal

Milica Stojmenovic[1] · **Robert Biddle**[1] · **John Grundy**[2] · **Vivienne Farrell**[3]

Published online: 16 March 2018

Abstract Word-of-Mouth (WOM) may impact the perception and experience of website usability and visual appeal. This study aimed to highlight the effects of WOM, implemented textually and verbally, on subjective and objective usability and visual appeal in a web environment. This research was spread over three studies and was undertaken using an unfamiliar city council website to exclude the influence of past experiences and to allow for greater control of WOM implementation. The statistical results showed that both visual appeal and objective and subjective usability were influenced via text that established expectations around these and that the results were only more compelling when verbal WOM was added. The result implications show that when the message is simple, such as it usually is in communication on social media and advertising, then it does impact people's perceptions of website visual appeal and usability, which may impact future intentions.

Keywords Visual appeal · Usability · Websites · Word-of-Mouth

1 Introduction

Use of computers has evolved far beyond simply typing input to a computer and viewing the output. Increasingly, computers include a range of sensors and sources of contextual awareness that can dynamically influence what the computer does. More-

✉ Milica Stojmenovic
milica.stojmenovic@carleton.ca

1 Carleton University, Ottawa, ON, Canada

2 Monash University, Melbourne, Australia

3 Swinburne University of Technology, Melbourne, Australia

over, the user can monitor this dynamic behaviour, using explicit input only as needed. Heidmann calls this emerging structure "human–computer cooperation" [1].

Our research has a focus on how external factors influence how a user perceives aspects of how a computer presents information. It has been previously established that factors such as visual appeal affect user perception. Our work is to investigate how user expectation will influence user assessment of the design and presentation of information. Because it affects the ability of user and computer to work together effectively, better understanding of this complex issue is important for human–computer cooperation.

Usable, appealing websites, such as tourism, are increasing in their importance as more and more people rely on them to plan their travels [2] and access essential government-related information. In fact, over a third of searches are related to finding travel information [3]. Governments have also been affected by the explosion of the Internet and have encouraged consumers to move the majority of their enquiries to the web. This is due to the current state of long queues, both physical and on the phone. Massive amounts of available information, and many forms that need to be organised, stored, and searched, need to be easily accessed by diverse groups of people, including the elderly, disabled, and regionalized groups and group coordinators. In addition, multiple people can have the same question, causing the need for repetition of dealing 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 independent enquiry process moving online is a natural progression.

However, why would there be a reputation of a government website to be usable and pretty, given that the everyday consumer's real-life bricks and mortar experience is one of countless documents, queues, and being moved from one counter to another or being transferred on the phone which is often unpleasant or unusable? How do we create the user intention and motivation to use the online interface? 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. Insufficient relevant research has been done on corporate websites that do not have a shopping function, even though such websites are common online, and users do expect them to be pretty and usable with useful content [4].

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 [5], "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". The ISO definition of usability was used here. Moreover, visual appeal is in the field of "aesthetics." Aesthetics is used to describe two different concepts: a pleasant experience and a visual property attributed to objects. Here, aesthetics was focused on a visual property of an object requiring judgment of its appearance [6]. In other words, an object's aesthetic appearance is subject to cognitive judgment, also known as aesthetic appraisal [7]. Measuring aesthetic pleasure/experiences would require physiological measures, such as measuring a person's

heart rate, which were out of scope. Henceforth, aesthetic appraisal is referred to as "visual appeal".

We examined these two critical areas and in particular, the effect of WOM on usability and visual appeal. WOM has been demonstrated to have substantial affect in e-commerce, influencing profit and trust [8]. However, this paper examines the impact of WOM in non-e-commerce websites, to see how it applies to areas that are currently considered to be boring, difficult, and less familiar. In particular, this paper examines the impact of nuanced textual and verbal WOM (what you read and what you hear) on usability and visual appeal on city council websites. This was done on a website genre where participants did not have highly developed mental models (website from [9, 10]), 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 of current counter and phone experiences are not always favourable.

In order to observe the impact of WOM on usability and visual appeal, we created "cognitive dissonance": a disagreement between new information and an individual's thoughts or environments, which may cause stress [11]. When dissonance occurs an individual strives to achieve consonance by reducing the inconsistency. The cognition that is most resistant to change is the one most recent one [12]. Here, the most recent behaviour would be experiencing the textual/verbal WOM. Therefore, new information can create expectations and these expectations can impact perceptions, intentions, and behaviour. Thus, participants were given information about the levels 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.

The key findings were that WOM influenced the perception of visual appeal and usability, more so when visual appeal and usability levels are congruent (i.e. both high, or both low) and when textual WOM was reinforced verbally, irrespective of use.

The rest of the paper is outlined as follows. The next section discusses background literature on usability, visual appeal, and WOM. This is followed by a summary of the preliminary studies done in previous work to obtain the website dataset. Finally, the three main studies in this paper are presented, each with their own introduction, method, results, and discussion sections. The paper ends with the general discussion and conclusion sections.

2 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 [13]. The relationship between visual appeal and usability is said to exist [13] 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 [14]. A phenomenon has been observed, called the "halo effect", in which people and things are judged and characters are assumed based on their appearance [15]. 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 [14]. 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 [16], others have found that they are affected after [17], and still others have found that they are affected throughout system use [13]. Only a small sample of the most central papers illustrating each of these was summarized here.

Tractinsky et al. [13] 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 [18]. Tractinsky et al. [13] 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 [13]. They 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 [19], 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 [20] 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 increasing the importance of usability to participants and decreasing the importance of visual appeal pre-use.

Tuch et al. [21] found that visual appeal did not affect perceived usability; rather, usability affected perceived visual appeal after use, on an online shop. 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 previous study [20], 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.

Thus, the relationship between usability and aesthetics has not yet been decisively defined and may vary depending on context [21], target audience [22], tasks, and mood [23]. 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. WOM's impact on usability and aesthetics has not yet been accounted for. 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.

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. For example, one study [24] 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's [20] results demonstrated significant correlations between perceived visual appeal and perceived usability and usefulness before system use, but not after. Yet, Tuch et al. [21] 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 [25]. Visual appeal helps create the first impression which can lead to an automatic peak in interested towards the website. Upon use, usability becomes more important as it is the factor that keeps users on a particular website [25]. Yet, not many concur with these findings. Thus, the relationship between usability and visual appeal has not yet been defined or generally accepted by researchers.

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 [21]. 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 [4] of both variables as a possible justification for the gap in the literature [21]. For example, one study [13] 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 study [21], 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 [26]. Therefore, to help alleviate these issues, this work 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 the impact 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 [18]. Tuch et al. [21] thus suggested that future research should examine the impact of the affective experience in the aesthetic evaluation. Since this research 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 [18].

McLellan et al. [27] found that prior experience and familiarity with a product increased usability ratings, regardless of product type. Prior experiences shape our mental models 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 knowledge that someone else had a bad experience or an overly good one with a similar system to the one being tested? They would be somewhat familiar and have their own expectations. McLellan et al. [27] work suggests that they would be more proficient in it and that in turn it may impact their liking of the system. Thus, the work

must be done based on first impressions, in an unfamiliar domain because controlling for or competing against developed mental models would be impossible at this stage. The gap in the literature may be filled by examining the initial impact of a controlled set of written and verbal WOMs.

Other website studies exist, but their purposes and topics do not align with this paper's. They tend to strive to achieve better usability, to create websites that will maximize profits, or to analyse the impact of visual appeal and usability on each other. For example, related to city websites, are tourism and hospitality website studies [28]. Ip et al. [28] reviewed a series of website hospitality studies in order to create a website evaluation system that assessed features and effectiveness. However, these are not within the scope of this paper, as we did not want to increase profit margins or create website evaluation methods. Instead, our work strived to further the understanding of website perception and interaction by uncovering the influence of WOM.

Sokkar and Law [29] addressed similar issues but did not present a study of user behaviour. 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 paper examines and supports that aspect of their model.

Overall, the relationship between usability and visual appeal has not yet been decisively defined and may vary depending on context [17, 21–23], target audience [22], tasks, and mood not being accounted for in many cases [23]. 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 and reputations have not been properly accounted for and the impact of WOM on usability and visual appeal has not yet been researched in the HCI community. This work contributes to an improved understanding of the relationship of usability and visual appeal by determining the degree to which WOM affects each variable, in a web environment.

2.1 Word-of-Mouth

Research on the impact of WOM on visual appeal and usability in websites is limited. The relevant pieces of literature found are summarized in this section.

All communication has the common purpose of sharing information. WOM tends be about people's experiences. 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 [30]). Recently, WOM has expanded to include textual and video recorded 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. 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. Yet, since people prefer WOM over other mediums [31], 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 textual and verbal WOM.

2.1.1 Textual WOM

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 [32]. We argue that positive or negative texts as well as verbal communication taint users and alter their perception and interaction with 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 [8]. 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 [33]. 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 [33]. User reviews can be considered online versions of WOM communication. Most online consumers actively look for and readily accept reviews because it effectively manages massive amounts of online information [32, 33]. 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 [32]. These textual reviews can implicitly convey information about perceived quality, ease of use, and usefulness [34]. In fact, positive feedback also increases trustworthiness and price premiums for reputable sellers [35].

The overwhelming majority (97%) of users rely on the textual feedback left by previous buyers before proceeding to purchase something from an unknown seller [33]. They have been found to significantly influence both sales [36] and consumer preferences [37], with about 80% of purchases being influenced by a recommendation [38]. Textual feedback impacts prospective consumers because it covers many aspects of the object being reviewed, and offers evidence of a seller's history which is used to predict the seller's future behaviour in transactions [33].

Pavlou and Dimoka [33] argue that nuanced textual messages can significantly impact trust ratings not only when they are left by a neutral party, but also when trust is intentionally impacted, such as in the case of marketing. This is a big problem in the Apple iOS App Store, where valuable personal information can be stored and profits are in the millions [39]. 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 [39].

From the above mentioned studies, it is clear that textual WOM setting can influence the actions of consumers. Here, it is investigated that positive or negative texts can alter their perception and use of websites (outside of the consumer domain). Thus, we examined if biased task descriptions could impact users when examining a government website. To the best of our knowledge, no other prior work has been done on this. The next section examines related work on verbal WOM.

2.1.2 Verbal WOM

The online environment is overloaded with information, and yet, not much research has been done on online communication [32]. This may not be the case for other fields. Given the importance of textual recommendations, feedback, and reviews, one would expect verbal WOM to have been thoroughly investigated but this is not the case, in HCI. Most research that has been done on WOM is over 20 years old. For example, [40] 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 that study [40] 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.

2.1.3 Confederates

In reference to testing processes, 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. The most famous experiment in psychology that used confederates to influence their participants is described here.

The first known study to use a confederate to sway participants was done in the 1950s 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 [41–43]. 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.

The use of texts and confederates strongly suggests that the influence of textual and verbal communication indeed impact a person's thoughts and actions. To the best of our knowledge, no other prior work has been done on the influence of WOM on websites, and no one else 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 paper examined the impact of textual (i.e. embedded in the task description) and verbal (via confederate) WOM on visual appeal and perceived and objective usability.

3 Preliminary studies

To see if WOM influenced visual appeal and usability, a series of five preliminary studies [9, 10] was done to obtain a website that was relatively unfamiliar to participants. A website with less developed mental models was needed in order to make the manipulation of WOM easier, without the influence of prior experience and existing expectations.

The *first* preliminary study examined 26 tourism and 26 city council websites from which 30 participants rated visual appeal and usability of each website after having briefly seen it. In addition, participants filled in a questionnaire for each of the two website genres, indicating their expectations of what they anticipate they would find on the sites (e.g. shopping carts, image galleries) and their sentiment towards the genre (e.g. easy to use, pretty, boring; [9, 10]). The results revealed that participants were less accurate for city council websites, inaccurately identifying items on the page, and rating them as boring, ugly, and hard to use. Yet, the Gold Coast city council website (Fig. 1) was rated as the prettiest of all websites, defying their expectations of the genres.

Then, in the *second* preliminary study, the empirically chosen website was tested by three usability experts in a heuristic evaluation and deemed it to be easy to use. In the *third* preliminary study, the usability was user-tested, by 10 volunteers, who found it relatively easy to use as well. Consequently, the Gold Coast city council website was both pretty (based on the results in of the first preliminary study) and easy to use (based on the second and third preliminary studies).

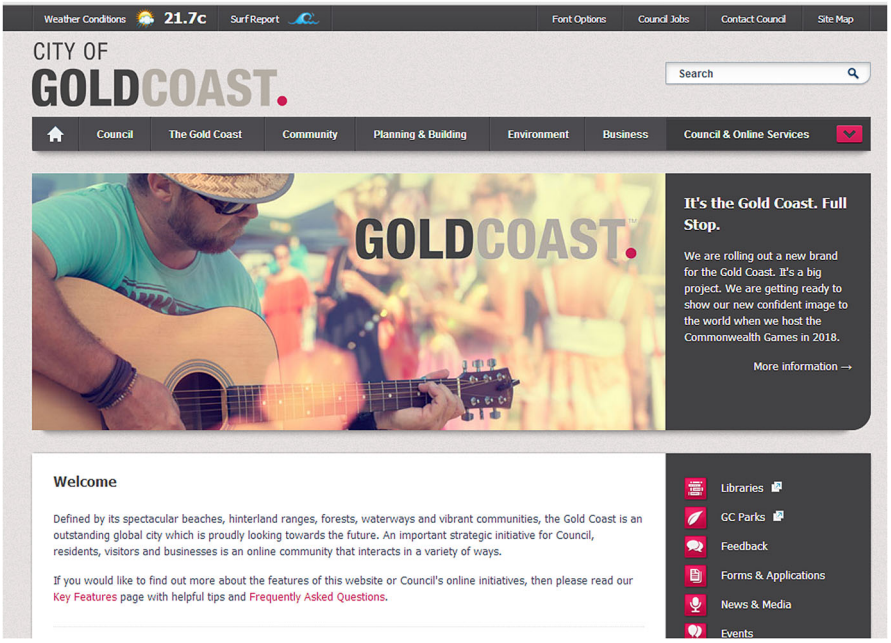

Fig. 1 The "pretty" and "hard" website version

In the *fourth* and *fifth* preliminary studies, the website was manipulated to create three additional versions of it, ranging in usability and visual appeal. To lower the usability level, 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 to alter the consistency and simplicity of the menu. With each click on the website, all the titles in the menu bar changed. 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. For example, the "Council" tab would randomly change to one of: Board, Assembly, Committee, Congress, Politics, Government, Law, or Jury. In addition, title contrast in the dropdown menu that appeared when a user hovered over the top menu bar was lowered. The dropdown menus originally had multiple categories of submenus to choose from. Titles were un-bolded and un-underlined, which perceptually removed the titles and created uncategorized lists under the menus.

For a statistically lower usability level, we further lowered usability. Tuch et al. [21] 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, like it did for the main menu titles. In addition, the location of each of the menu items that were needed 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".

The visual appeal was lowered by changing the main background colour from the original grey-beige to lilac and changing the texts' background from a light off-white to evergreen. We left the textual background white so that with the text 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. The combination of lilac and evergreen was chosen since it was the colour scheme of the Toowoomba tourism website, rated worst in the first preliminary study. In addition, the colours of all the images in the website were inverted to be negatives. Thus, the data sample consisted of the following.

- "Good": The original website was used as the "good" website (i.e. easy to use and pretty, based on the first three preliminary study data). This website was manipulated into three additional versions.
- "Easy/Ugly": Only the manipulations to lower visual appeal were done in this version.
- "Hard/Pretty": Only the manipulations to lower usability were done in this version.
- "Bad": All manipulations were included in this website version, to create a hard and ugly website.

The manipulations were user-tested and statistically verified. Thus, the preliminary studies resulted in the development of an empirically chosen and tested website data sample that statistically varied in usability and visual appeal. 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. More preliminary study details can be seen in [9, 10].

4 Study 1

4.1 The impact of *textual* WOM on congruent visual appeal and usability levels

Study 1 used the good and bad versions of the website, where visual appeal and usability levels were congruent. The purpose was to see if textual WOM influenced the visual appeal, perceived and objective usability (i.e. performance measures per task: clicks, hovers, completion time, and success) of a website, when usability and visual appeal levels were congruent (i.e. both were either good or both were bad). The good and bad website versions were subjected to three WOM conditions: highly positive (i.e. good) textual communication (i.e. task description) of visual appeal and usability, negative (i.e. bad) in both, and no WOM (control). Thus, there were six conditions in this phase: good website with congruent WOM, good website with bad WOM, good website with neutral WOM, bad website with good WOM, bad website with bad WOM, and bad website with neutral WOM.

Based on the theory of cognitive dissonance, if WOM influences visual appeal and usability, then participants should agree to the information provided, and the perceived variables should be reported as either higher or lower, in accordance with the communicated level. Also, if WOM influence perceived usability, then they may

also influence objective usability (in the form of performance measures: clicks, hovers, time, and passes), so that participants struggle more with the website if the textual WOM is low and vice versa. To test these hypotheses, quantitative data were analysed for evidence of the influence of WOM on usability and visual appeal.

4.2 Method 1

4.2.1 Participants 1

Random university students were used because they are a representative sample of the general population and do not pose a threat to external validity [44, 45]. Thus, a sample of 60 (39 males, 21 females; 48 aged 18–30 years, 12 aged 31+) 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, two each for games development, arts, and psychology, and one each for engineering, business, biomedical engineering, and astrophysics and supercomputing. Participants were randomly assigned and individually tested, approximately 1 h per session, ten participants per condition.

4.2.2 Apparatus and location 1

Participants were tested using a Hewlett Packard desktop computer, running Intel® Core™2 Duo CPU with 3 GB of RAM, and a screen resolution of 1290×720. 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 analyse the data statistically and to produce the figures. The study took place in the usability laboratory 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.

4.2.3 Materials 1

First, informed consent and project information forms were given to participants. A demographic questionnaire was administered to determine the participants' background information (e.g. age, gender, and education). The System Usability Scale (SUS; [46]) and the Visual Aesthetics of Websites Inventory—Short version (VisAWI-S; [47]) scales were used for perceived measures of usability and visual appeal, pre- and post-use.

The SUS scale is a short, 10-item, widely used questionnaire. 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 (e.g. strongly agree). Each participant

fills in all ten questions. According to the 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. By having negative questions, it 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 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).

Objective usability in the form of performance measures was acquired per task. These were the number of clicks, the number of hovers, task completion time in seconds with a maximum of 180 s (i.e. 3 min), and success (pass/fail; pass if the answer is correct and within time limit). 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.

As mentioned earlier, two versions of the website were used: good and bad. Three different task descriptions were prepared, a paragraph long each, one had positive text for visual appeal and usability, one was low for both, or neither (control/neutral paragraph, no biased WOM). For example, the high visual appeal, high usability task description was:

> 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.

Participants were not given time or the opportunity to surf or look idly at the website; pre-use measures were based on a 6s exposure to the website's homepage, a page one click in, and a page two clicks into the website (2s/pg.). Two five-slide PowerPoint presentations (one for each of the two versions of the website) were prepared with the first slide being the instructions, the second a "+" focus slide, and the last three screenshots of the interface. Ten information retrieval tasks were given to participants, in random order. An example of a task is: "How many beaches are located in the Gold Coast?"

4.2.4 Design 1

This study adopted a two-by-three (two websites, three reviews transmitted via textual WOM) 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.

4.2.5 Procedure 1

Each participant did the experiment separately in one-on-one sessions with the researchers. Once the given participant was briefed on the purpose and procedure, she/he signed the consent form. Then, the participant was given the nuanced 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 and usability 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.

4.2.6 Data analysis 1

Data were first graphed using beanplots to gain an understanding of any emerging trends. Then, normality and homogeneity of variance were tested. Then, the averages were calculated per condition pre- and post-use for visual appeal and perceived usability. The average results for the objective usability measures were calculated across tasks, per participant. Nonparametric tests were applied: Kruskal–Wallis for main effects and Wilcoxon Mann–Whitney for pairwise comparisons. Kruskal–Wallis is the nonparametric equivalent to ANOVA [48].

4.3 Results 1

The results for all three studies are structured as follows. First, for this and subsequent studies, we represent data with beanplots. Beanplots are a more advanced form of box plots, where the distributions are shown on both sides of the middle bar [49]. They

give the population spread which allows for more accurate conclusions to be drawn. They can also visually present more complicated results.

Second, before we examined causational statistical analysis, we checked the homogeneity and normality assumptions, to ensure that we used the proper statistical tests (i.e. parametric or nonparametric). The normality assumption was tested using Shapiro–Wilk [50, 51] and skewness and kurtosis measures [52, 53]. If the normality assumption was not violated, then parametric tests are used (e.g. the parametric Levene's test, [54]) to examine the homogeneity of variance assumption. However, in this thesis, the normality assumption was violated in every study and thus the nonparametric Levene's test was used [55]. 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, nonparametric 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.

4.3.1 Preliminary beanplot results

Using the results from the SUS and VisAWI-S, beanplots were created to gain a general understanding of the data. Pre-use and post-use visual appeal ratings are shown in Figs. 2 and 3, respectively, and pre-use and post-use perceived usability in Figs. 4 and 5, respectively. The SUS and VisAWI-S scores are presented on the vertical axis. In all of the figures, the grey beanplots on the top relate to the easy and pretty website and the white ones on the bottom relate to the hard and ugly websites. In Figs. 2, 3, 4, and 5, the first columns on the left represents the control condition for control condition (in all four figures), the middle columns are the positive text (i.e. good: easy/pretty) conditions, and the ones on the far right are where the participants were given bad textual WOMs (i.e. bad: hard/ugly). The thick black lines indicate each condition's mean.

For the easy and pretty website, there are slight differences between the three conditions, evidenced by the proximity of the mean lines in Fig. 2. The second column, which corresponds to the good WOM condition, is the highest, followed closely by the control condition, in the first column. Then the lowest is the third column which is the negative text condition (i.e. bad textual WOM).

This finding suggests that participants were impacted by the written WOMs, and perceived the same website differently based on their experimental condition. However, the difference between the means does not appear to be large. There also appears to be a bimodal distribution of data in the pre-use visual appeal ratings of the bad text condition (rightmost, top, grey bean in Fig. 2). This suggests that about half of the participants agreed with the text and rated visual appeal lower.

The hard and ugly website depicts a slightly different story. In the hard and ugly website, the control was rated the prettiest, with the highly good WOM condition in the middle, and bad WOM condition being the lowest. Specifically, participants in the bad WOM condition perceived visual appeal to be on average one point lower (uglier)

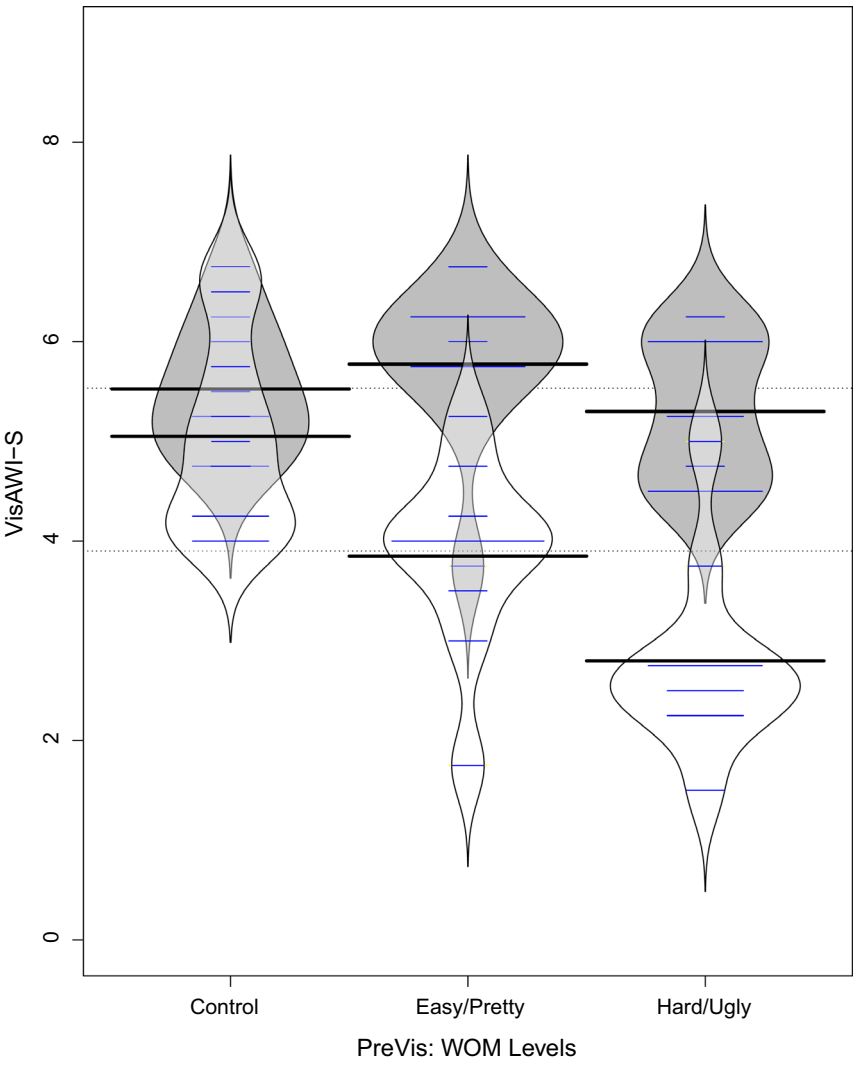

Fig. 2 Beanplot of the pre-use visual appeal results

than participants in the good WOM condition, and two points lower (uglier) than in the control condition. Such a large disparity should have also been found in the visual appeal ratings for the easy and pretty website, but the differences were smaller (Fig. 2). The good WOM condition was most likely rated lower than the control condition because participants experienced some disappointment with the ugly website, still rated it as prettier than participants in the low WOM condition.

Across the easy and pretty website conditions, there are slight changes between pre-use (Fig. 2) and post-use (Fig. 3) and between conditions; altogether dropping roughly half a point in visual appeal after use. A point refers to the 1–7 point VisAWI-S scale used for measuring visual appeal. Participants in the easy and pretty website

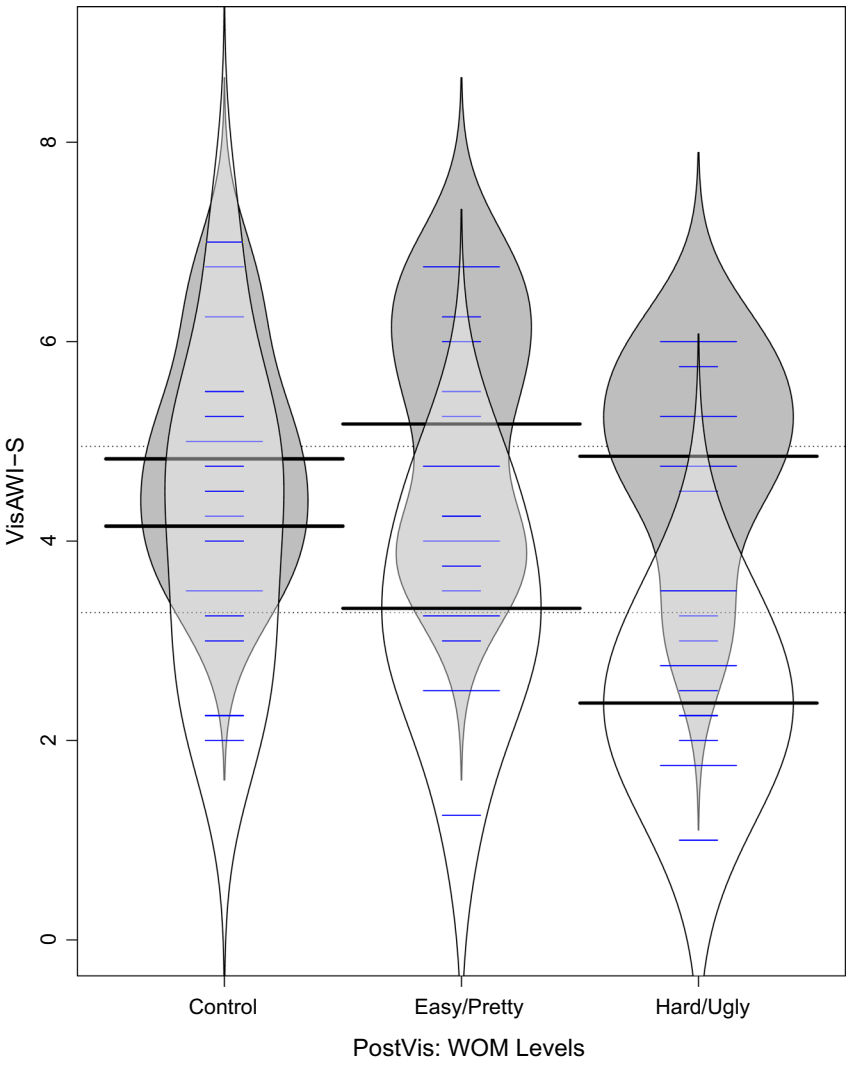

Fig. 3 Beanplot of the post-use visual appeal results

with good WOM (middle upper bean in Fig. 3) appear to be polarized, either rating post-use visual appeal higher or lower than the group's mean. The control and bad WOM conditions in the easy and pretty website seem to be similarly rated, with the majority of the bad WOM participants rating it higher than the control group. One reason this may have occurred was that the bad WOM group seemed to reconsider their bad WOM and think that it was not so bad, whereas the majority in the control group were not overly impressed with the website's beauty.

Post-use visual appeal means in the hard and ugly websites are also similar to the pre-use ratings, but are more normally distributed. In addition, the control condition for the hard and ugly website has a larger spread suggesting that one or two participants

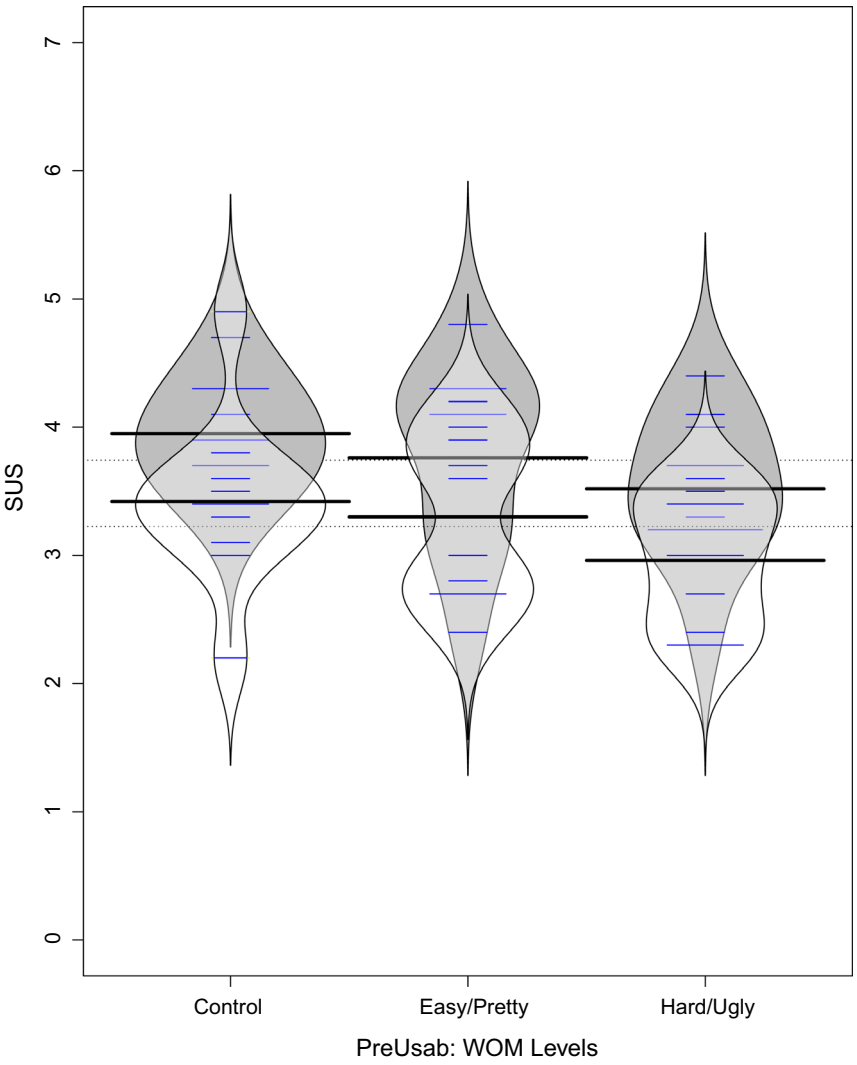

Fig. 4 Beanplot of the pre-use perceived usability results

thought it was a lot prettier and one or two thought that it was a lot uglier than pre-use, but the majority was still centred around the mean. Participants with the negatively phrased text rated visual appeal lower than participants in the other two conditions. Again, the good WOM text condition was most likely rated lower than the control condition because participants experienced disappointment with the ugly website, yet still rated it as prettier than participants in the bad WOM condition.

Altogether, evidence exists that visual appeal differs between conditions in the easy and pretty website, both pre- and post-use.

There seems to be a small yet consistent difference between the groups across all conditions for pre-use usability, as seen in Fig. 4. The control condition seems to be rated as the easiest to use, followed by the good WOM, and then by bad WOM which

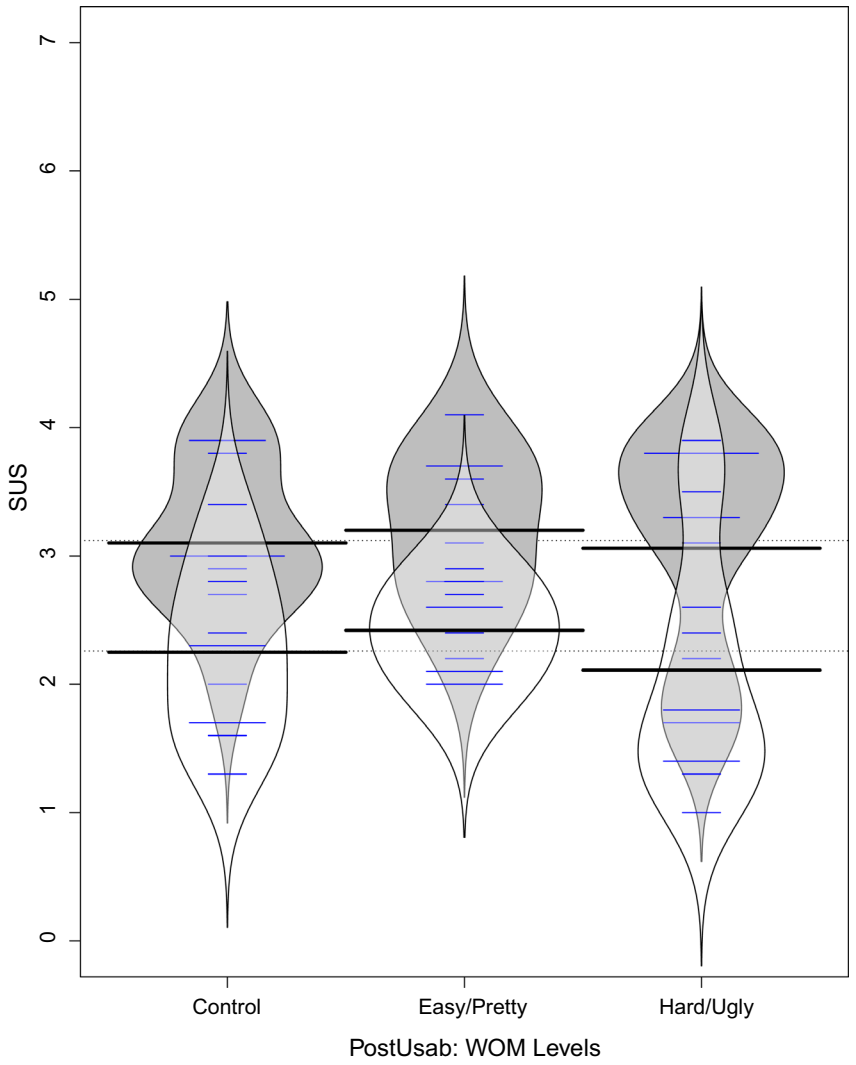

Fig. 5 Beanplot of the post-use perceived usability results

is rated as the hardest, across both website versions. The means to these groups are within a one-point range which can be considered a small difference (the scale is 1–5). However, the difference does appear to be constant. All three conditions for the hard and ugly website appear to have a bimodal distribution, suggesting that some participants either had hope that the website would be easy to use or that they thought it would not be bad. Given that the layout of both websites was not changed (so that visual appeal would not be altered as well), the similarity in usability ratings across conditions appears appropriate pre-use.

After use, ratings of usability dropped, especially for the hard and ugly website, as seen in Figs. 4 and 5. Post-use, in both the easy and pretty website and the hard and ugly, the good WOM condition was rated as the easiest but by a fraction, as seen in

Fig. 5. There also appears to be a small difference between the control and low WOM conditions, where the low WOM condition was rated as lower, in both websites.

4.3.2 Assumptions testing 1

Statistical assumptions for normality and homogeneity were tested 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. These values were computed by SPSS from the data. In summary, the data was not normally distributed and variance was heterogeneous. Thus, nonparametric tests were 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 and Fisher's exact tests on specific groups were done to determine which pairs were significantly different, and they can be one-tailed to determine directionality.

4.3.3 Statistical hypothesis testing 1

Out of the eight statistical sub-hypotheses tested for visual appeal and perceived usability (pre- and post-use, on two different website versions), two were significant. Both pre- ($H = 15.069$, $p < .01$) and post-use ($H = 7.883$, $p < .05$) visual appeal ratings were significantly different within the bad website (i.e. main effects were found in the bad website conditions). Paired comparisons showed that bad website with bad WOM and the same website's control condition differed in visual appeal both pre- ($H = 5.5$, $p < .001$) and post-use ($H = 16$, $p < .05$). In addition, pre-use visual appeal differed between the control and good WOM conditions ($H = 82$, $p < .05$), and between the good and bad WOM ($H = 81$, $p < .05$). Therefore, statistical evidence was found to support the hypothesis: visual appeal does appear to be influenced by textual WOM.

The average number of clicks per task significantly differed ($p < .05$) within the good website (i.e. one main effect found with clicks). Pairwise comparisons showed that the number of clicks were different ($p < .05$) between the good website with good textual WOM ($\bar{x} = 3.98$) and good website with bad textual WOM ($\bar{x} = 3.07$) conditions. This suggests that participants interacted more so with the website that had the positive text 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 possible reason for this finding was that the low WOM may have discouraged the participants from exploring the website.

4.3.4 Correlations study 1

4.3.4.1 Spearman correlations in the control conditions First, we examine the Spearman correlations, ρ, between the two control conditions (of the good and bad websites), in Table 1. In the good website's control condition, both pre- and post-use usability ($\rho = 0.780$, $p < 0.01$) and pre- and post-use visual appeal ($\rho = 0.635$, $p < 0.05$) were

Table 1 Correlations between visual appeal and perceived usability for good and bad websites' *control* conditions, respectively

Good/control				Bad/control			
	PostUsab	PreVis	PostVis		PostUsab	PreVis	PostVis
PreUsab	.780**	.415	.422	PreUsab	.202	.716*	.924**
PostUsab	–	.009	.334	PostUsab	–	−.193	.280
PreVis		–	.635*	PreVis		–	.665*

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

Table 2 Correlations between visual appeal and perceived usability for *good/good* and *bad/bad*, respectively

Good/good				Bad/bad			
	PostUsab	PreVis	PostVis		PostUsab	PreVis	PostVis
PreUsab	.686*	.658*	.782**	PreUsab	.797**	.187	.204
PostUsab	–	.119	.715*	PostUsab	–	.208	.585
PreVis		–	.250	PreVis		–	−.053

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

highly and positively correlated. Perceived usability and visual appeal measures were not significantly correlated with each other in this condition.

In the bad website's control condition, pre-use usability was highly and significantly correlated both with pre- ($\rho = .716$, $p < .05$) and post-use ($\rho = .924$, $p < .01$) visual appeal. Pre- and post-use visual appeal was also highly and significantly correlated with each other ($\rho = .665$, $p < .05$). 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.

4.3.4.2 Correlations when WOM and website levels are congruent All Spearman correlations between visual appeal and perceived usability within conditions where WOM levels were congruent with the actual website levels can be seen in Table 2. In other words, the following two conditions are discussed in this section: the good website with the easy/pretty (i.e. good) WOM and the bad website with the hard/ugly (i.e. bad) WOM.

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

Therefore, in the conditions where the WOM levels of visual appeal and usability were congruent with the website's visual appeal and usability levels, the correlations were positive and strong between visual appeal and perceived usability pre- and post-

Table 3 Correlations between visual appeal and perceived usability for good website with bad WOM (good/bad) and the bad website with good WOM (bad/good), respectively

Good/bad				Bad/good			
	PostUsab	PreVis	PostVis		PostUsab	PreVis	PostVis
PreUsab	.163	.440	.093	PreUsab	.454	.577	.221
PostUsab	–	−.533	.555	PostUsab	–	.470	.164
PreVis		–	.053	PreVis		–	.470

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

use for the good website with equally "good" WOM. However, when the usability and visual appeal were worse in the bad/bad condition, participants attributed the poor website to usability and did not agree on visual appeal.

4.3.4.3 Correlations when WOM and website levels are incongruent All Spearman correlations between visual appeal and perceived usability in conditions where WOM of these were incongruent with the actual website levels are shown in Table 3. Nothing was correlated in the good website with bad WOM and bad website with bad WOM.

While causality cannot be inferred from correlations, there does seem to be some evidence to support the idea that WOM impact visual appeal and usability. The control and congruent conditions seem to be behaving the same way as many studies in the literature. However, when the WOM and website levels are incongruent, the correlations disappear. This may have occurred if the WOM were internalized and acted upon differently.

4.4 Discussion 1

4.4.1 Results summary 1

4.4.1.1 Beanplots The beanplots showed slight variations in the means across the hard and ugly conditions where the control condition was rated highest pre-use and post-use visual appeal and pre-use perceived usability, with low WOM rated the lowest. Post-use, the trend in perceived usability changed slightly, with good WOM rated as easiest and the low WOM condition was rated as hardest. Thus, a small trend did emerge, especially post-use.

In the hard and ugly website with the congruently nuanced text, pre-use and post-use visual appeal ratings were rated as lower than the control use and post-use visual appeal ratings, respectively. This would suggest that participants in the hard and ugly website with congruently nuanced text, perceived the website to be uglier than participants in the control group, irrespective of use. This was also found in the beanplots. Written WOM have an impact on visual appeal since it was rated lower when WOM were set to be low, especially between the control group and the bad textual WOM group, which rated the website as a lot uglier. This difference lasted from pre-use to post-

use, *suggesting that the impact of the opinionated text was strong enough to influence ratings after having been exposed to the website for roughly an hour*. However, this difference was not found in the easy and pretty website, neither was it found between the good and bad conditions of the hard and ugly website, nor with perceived usability.

4.4.1.2 Statistics The pre- and post-use visual appeal ratings hard/ugly website with congruently low WOM differed from the control pre- and post-use visual appeal ratings, respectively. Thus, participants in the bad website with negative WOM condition perceived the same website to be uglier than participants in the control group, irrespective of use. Furthermore, the number of clicks per task varied significantly between the good website with good WOM ($M = 3.98$) and good website with bad WOM ($M = 3.07$) 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 WOM have an impact on both the perceptions and actions of people using the websites.

4.4.1.3 Correlations Overall, correlations between visual appeal and perceived usability were significant and positive pre- and post-use, agreeing with Tractinsky et al.'s [13] results. This was the case in the good/good condition where the WOM was congruent with the website level. However, when the usability and visual appeal were worse in the bad/bad condition, participants attributed the poor website to usability and did not agree on visual appeal. In the case where WOM 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), textual WOM may not have been enough to convince users. The addition of verbal WOM was tested in the next study, below.

4.4.2 Limitations and future studies 1

One limitation in this study was the assumption that the impact of WOM could occur after a short, non-repeated exposure, by reading a short task description, in an unfamiliar physical environment (i.e. ecological validity). Unfamiliarity of the location and experimenter could have also influenced trustworthiness of the text, lowering its value. It may also have also been the case that they did not read the task descriptions, since Rettig [56] found that participants do not always thoroughly read hard-copy or online documents. Another possible limitation could be a conflicting learning style [57]—if participants in this study were predominantly verbal learners, then the textual (i.e. visual) WOM may not have been effective. Without prolonged exposure, text may not have been enough to fully understand the impact of textual WOM on visual appeal and usability.

It may have been more effective to give the WOM's message subliminally or at least less overtly, but this will be subject to similar issues as written WOM (e.g. they may not be understood). For example, implementing biased WOM 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 texts in fact do not influence or do not significantly influence ratings of

visual appeal and usability. These issues were addressed in Study 2 by the addition of a confederate who verbally reinforced the textual WOM.

5 Study 2

5.1 The impact of *textual and verbal* WOM on congruent visual appeal and usability levels

Textual communication did affect visual appeal and objective usability in some circumstances, as seen in Study 1. However, the effect of them on visual appeal and usability was smaller than anticipated. Therefore, the purpose of Study 2 was to reinforce texts by implementing them verbally as well. Study 2 only used the good website, with the good and bad WOM to re-examine the effect of WOM on visual appeal and usability, to see if more significant results could be obtained. The nuanced texts were reinforced by a confederate who acted like a participant finishing the study and gave the real participant their "opinion" (i.e. verbally) after "having completed the study themselves". This opinion was in fact a similar speech to the task description.

5.2 Method 2

5.2.1 Participants 2

A sample of 20 (16 males, 4 females; 16 aged 18–30 years, 4 aged 31+) different participants were recruited in Study 2. Participants were 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 and seven were only somewhat familiar. Participants were randomly assigned and individually tested, approximately 1 h 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 Study 1.

5.2.2 Apparatus, materials, and location 2

All apparatus and materials pertaining to this study are the same as in Study 1. The confederate used a standard memorized script, mimicking the message in the task descriptions. The usability laboratory used in the previous study was also used here.

5.2.3 Design and the confederate 2

This study adopted a one-by-two (one website, two different messages via WOM: either completely positive or completely negative) between-group design. One con-

federate was used in this study: a native English speaker, PhD student. The role was to act like a participant finishing the study and then tell her opinion of the website to the real participant via standard script. The confederate was added at the very beginning, as seen in Fig. 6. Then, identically to the design of Study 1, 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.

5.2.4 Procedure and data analysis 2

The procedure is identical to Study 1, with the exception of the very beginning in which the confederate was added. As seen in Fig. 6, the study started with the experimenter greeting the participant and asking the participant to wait while the computers were reset from the previous participant. A confederate was in the experiment room, picking up their things and acting like they were getting ready to leave after having participated in the study themselves, as the participant entered the room. The experimenter then asked the confederate if they were all done and the confederate would respond that they were just leaving. The experimenter thanked the confederate for participating, and then the experimenter left to "set up the computers". This left the real participant and the confederate alone in the room. The confederate then told them about the usability and visual appeal in the form of their experience with the website (since they were acting like they had just done the same experiment themselves) and then left. The level of usability and visual appeal (high or low) varied depending on what condition the participant was placed in, as seen in Fig. 6. The experimenter then came back into the room and then started with the introduction and rest of the procedure from Study 1. The data were analysed in the same way as in Study 1.

5.3 Results 2

5.3.1 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 (Fig. 7) and pre- and post-use perceived usability (Fig. 8). In both of these figures, the first column on the left represents the good website's control condition, the middle one is the good website's good WOM condition, and the one on the far right is the same website's bad WOM condition. The grey beanplots are the pre-use measures, and the white ones are the post-use measures. As mentioned earlier, the control condition was not redone in Study 2, but the data was taken from Study 1 for the purposes of comparison.

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 neutral and bad conditions. Pre-use and between conditions, the good WOM condition was perceived to be slightly prettier than the control condition, which was perceived to be slightly prettier than the bad WOM condition. Post-use, the participants in the neutral WOM condition seem to have lowered their ratings of visual appeal while the ratings stayed identical to what they were pre-use in the good WOM condition. Post-use for the bad WOM condition,

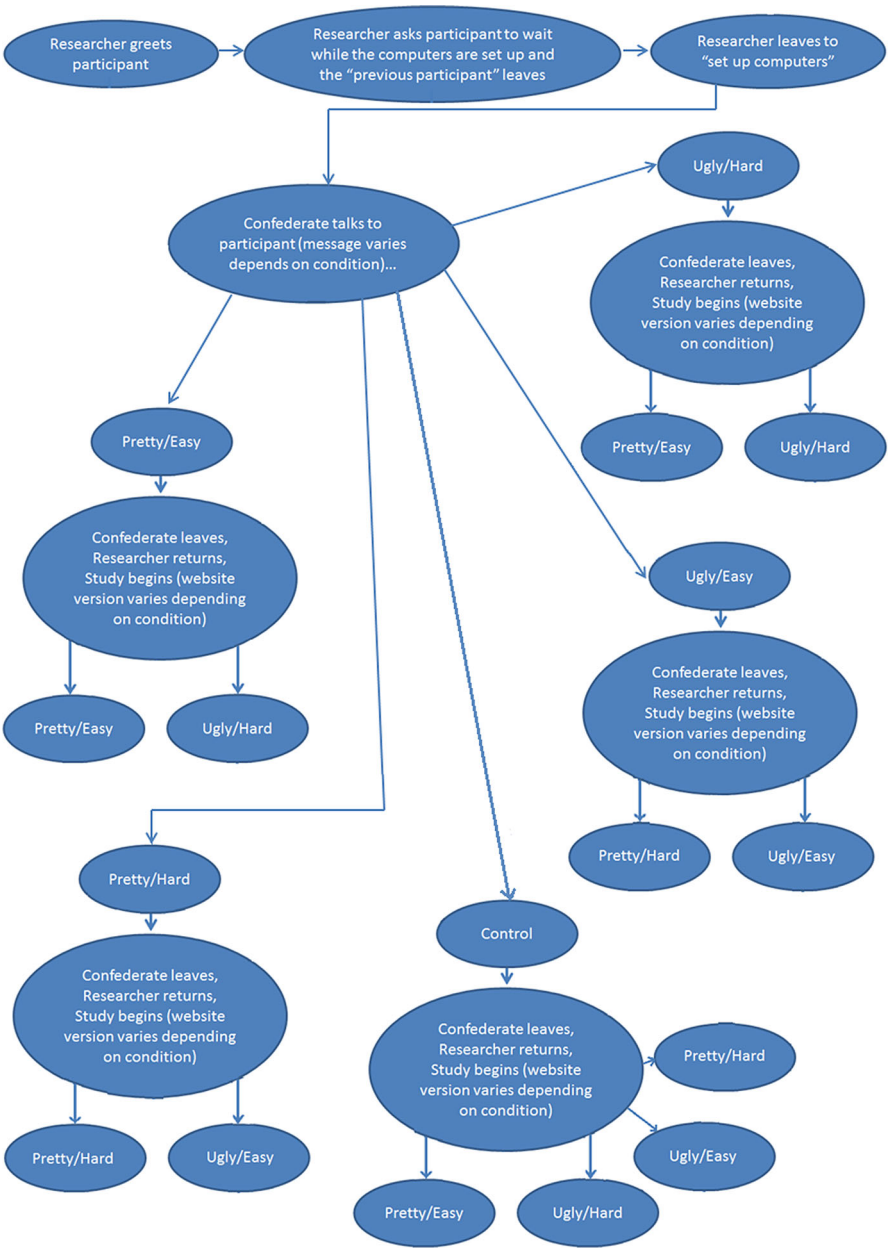

Fig. 6 The procedure for Studies 2 and 3

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 addition via confederate seems to affect participants' perceptions of visual appeal more drastically post-use.

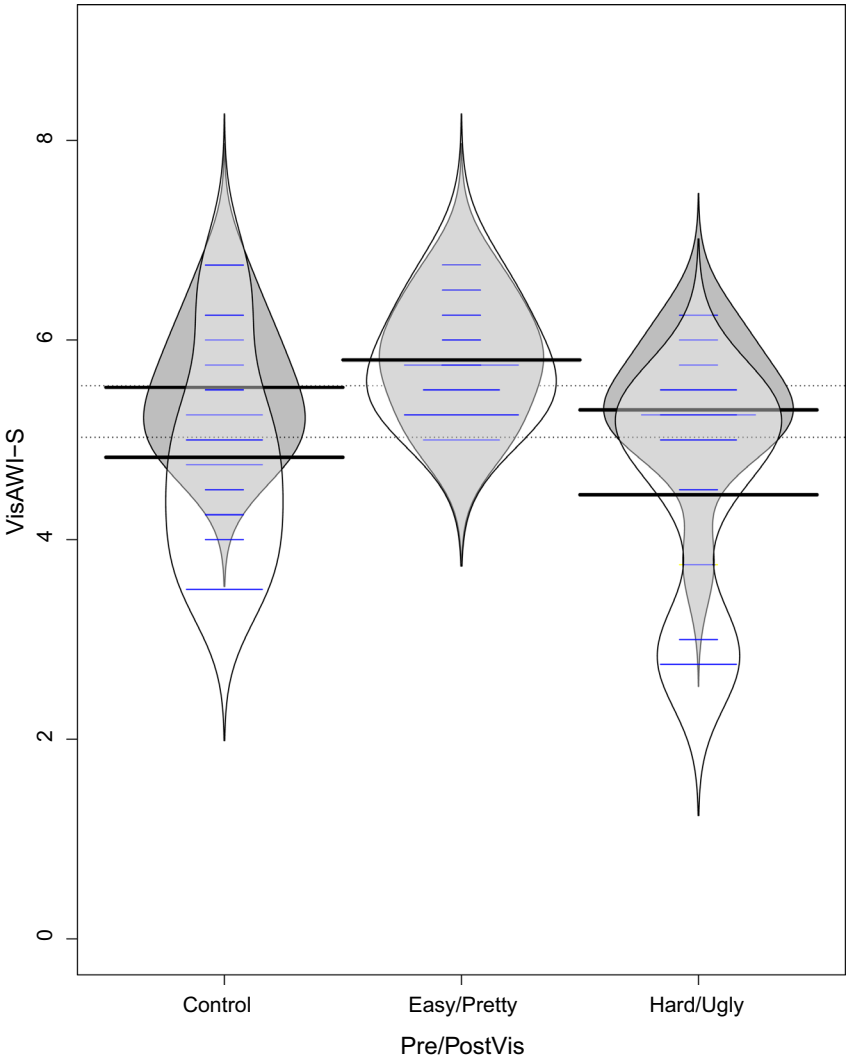

Fig. 7 Beanplot of the pre- and post-use visual appeal results

Altogether, visual appeal does differ between conditions in the same website, pre- and post-use.

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 Figs. 7, 8). 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 bad WOM condition was perceived to be the hardest, given the height of the distributions and the red lines which depict their means. The addition of verbal communication via

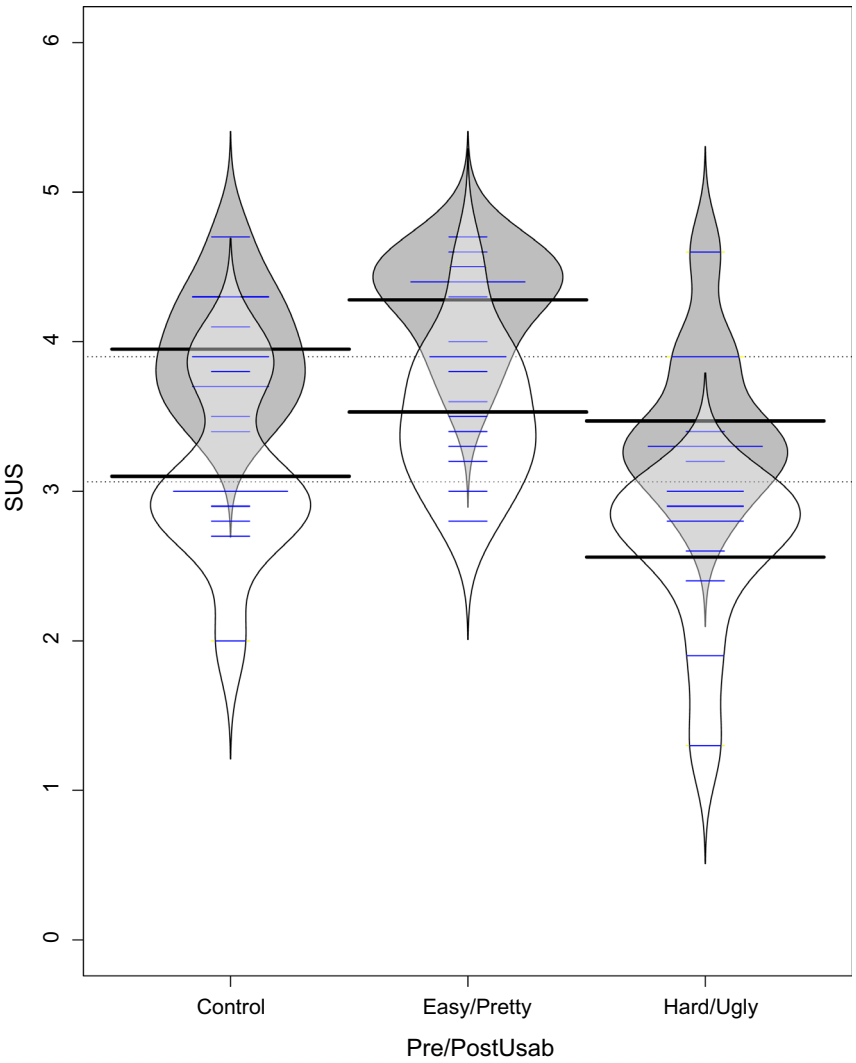

Fig. 8 Beanplot of the pre- and post-use perceived usability results

confederate does seem to affect usability perceptions, and the impact lasts post-use as well.

Therefore, perceived usability differs between conditions in the good 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.

5.3.2 Statistical hypothesis testing 2

Main effects were found in pre-use usability ($H = 11.553$, $p < .001$), post-use visual appeal ($H = 9.296$, $p < .01$), and post-use usability ($H = 11.853$, $p < .001$). Paired

comparisons showed that good website with good and bad WOM differed in pre-use usability ($H = 88.5$, $p < .01$), post-use visual appeal ($H = 89$, $p < .05$), and post-use usability ($H = 94$, $p < .01$). Post-use visual appeal differed between the control and good WOM conditions ($H = 21$, $p < .05$). Moreover, pre-use usability ($H = 19.5$, $p < .05$) differed between the bad WOM and control websites.

The average number of clicks per task ($p < .05$) differed within the good website conditions. Pairwise comparisons showed that the number of clicks were different ($p < .01$) between the good website with good and bad WOM conditions. Specifically, participants in the bad WOM condition, on average, clicked more often per task (3.82 clicks) than those in the good 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 good website with good and bad WOM conditions. Participants took over half a minute longer to complete a task in the bad WOM condition (i.e. 108.13 s in bad versus 70.67 s in good). The significance of the comparison of the average of passed tasks was confirmed ($p < .01$, one-tailed) with a Fisher's Exact test, in which the good website with good WOM had a larger success rate (0.83) than bad WOM condition (0.58).

Therefore, participants rated the same website as prettier and easier to use when they were told that it was going to be 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. WOM influenced both how participants viewed and interacted with the website.

5.3.3 Correlations study 2

The Spearman correlations can be seen in Table 4. Contrary to the results in Main Study 1 of the good/good condition, where most of the variables were correlated, here the correlations disappear for the most part. Pre-use visual appeal was highly and positively correlated ($r = .725$, $p < .01$) with post-use visual appeal, only in the good/good condition.

Participants seem to evaluate usability and visual appeal separately, when they are equally high in the good website. In fact, given that the correlation between pre- and post-usability does not exist here, one might conclude that the good WOM set the

Table 4 Spearman correlations between visual appeal and perceived usability for good/good and good/bad conditions, respectively

Good/good				Good/bad			
	PostUsab	PreVis	PostVis		PostUsab	PreVis	PostVis
PreUsab	.098	.611	.349	PreUsab	.444	−.263	.268
PostUsab	–	.330	.519	PostUsab	–	.034	.409
PreVis		–	.725**	PreVis		–	.125

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

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 the good/bad condition on the other hand are equally as insignificant as they were in Study 1. This would suggest that participants initially had hope for the website but later agreed with the WOM, and lowered their ratings for it even more so than initially (given the discrepancy in means between good and bad WOM).

5.4 Discussion 2

5.4.1 Results summary 2

5.4.1.1 Beanplots For visual appeal, highly positive communication 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 bad WOM 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 affected participants' perceptions of visual appeal pre- and post-use.

For usability, the good WOM condition was rated as the easiest to use and the bad WOM condition rated as the hardest, both pre- and post-use. The additional verbal implementation via confederate did affect usability perceptions, and the impact lasted post-use as well.

5.4.1.2 Statistics The good website with good WOM and the good website with bad WOM 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 they read and 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 good website with good WOM and the good website with bad WOM conditions. More precisely, the bad WOM condition made more clicks, took nearly double the time, and had a lower success rate than the good WOM condition, doing the same tasks and using the same website.

5.4.1.3 Correlations Only pre-use visual appeal was highly and positively correlated with post-use visual appeal, in the good/good condition. 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 WOM did influence ratings of visual appeal and usability, and that participants largely agreed with the WOM.

5.4.2 Limitations and future research 2

5.4.2.1 Threats to internal validity This study was done using a confederate (as described above) 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 the textual communication. However, this may not be the best way to do so given the unfamiliarity, untrustworthiness, and minimal exposure to the confederate. Yet, in this study, the results showed that the communication did influence usability and visual appeal more so than in Study 1. Therefore, a confederate was used in the next as well.

6 Study 3

6.1 The impact of *textual and verbal* WOM on *incongruent* visual appeal and usability levels

To gain a deeper understanding of what effect WOM had on usability and visual appeal, Study 3 examined the influence of it (both textual and verbal) on visual appeal and usability when they are incongruent with each other. Specifically, the easy/ugly and hard/pretty website versions and messages were examined.

Each website version was subjected to three WOMs: high usability and low visual appeal (easy/ugly; where the first descriptor is always of usability and the second is of visual appeal), low usability and high visual appeal (hard/pretty), and neutral which was the control condition.

WOM 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 versions of nuanced WOM, there were six conditions in this study: (1) easy/ugly website with good WOM for usability but bad WOM about visual appeal (easy/ugly) WOM, (2) easy/ugly website with bad WOM for usability but good WOM in the text and verbal script for visual appeal (hard/pretty), (3) easy/ugly website with no WOM, (4) hard/pretty website with easy/ugly WOM, (5) easy/ugly website with hard/pretty WOM, and (6) the hard/pretty website with no biased WOM.

Similarly to the previous two studies, it was hypothesized that when the WOM's message for usability are set to be high but for visual appeal they are low, then participants will rate them accordingly. 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 information in the communication. Conversely, when usability is said to be low and visual appeal to be high, then participants should internalize this information and rate the website accordingly. These are anticipated because accepting the information and adjusting their perception of the website will help them achieve consonance, according to the cognitive dissonance theory.

Consequently, it was also hypothesised that WOM would also affect participant performance (in the form of the classical objective usability measures). Specifically, in the easy/ugly website, participants should find the easiest to use to be the easy/ugly WOM group, followed by the control group, and then should struggle the most in the hard/pretty WOM conditions (i.e. incongruent). In the hard/pretty website, again, participants should find the easy/ugly. WOM as the easiest, followed by the control group, and the hardest should be the hard/pretty WOM condition. With similar reasoning as in the previous studies, this is hypothesised because participants would try

to match what was told to them to the website as a means to diminish any possible cognitive dissonance.

6.2 Method 3

6.2.1 Participants 3

A sample of 60 (38 males, 22 females; 49 aged 18–30 years, 11 aged 31+) 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 1 h per session, ten participants per condition.

6.2.2 Apparatus, materials, location, procedure, design, and data analysis 3

The same laboratory and all apparatus and materials pertaining to this study were the same as in Studies 1 and 2. As mentioned earlier, two versions of the website were used: easy/ugly and hard/pretty, with the according task descriptions and speeches for the confederate. Participants' audio and video were again not recorded. The procedure from Study 2 was repeated here, with the confederate.

The design was exactly the same as Study 1, which was a two-by-three (two websites, three sets of messages) 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. The data were analysed in the same way as it was in Studies 1 and 2.

6.3 Results 3

6.3.1 Preliminary beanplot results

Beanplots were created to gain a general understanding of the data, with pre- and post-use visual appeal in Figs. 9 and 10, respectively, and pre- and post-use perceived usability in Figs. 11 and 12, respectively. In all of the figures, the grey beanplots are the easy/pretty website measures and the white ones are the hard/pretty website measures. The first columns on the left represents the control conditions, the middle columns are the easy/ugly conditions, and the ones on the far right are the hard/pretty conditions. The thick black lines indicate each condition's mean.

Pre-use, the distributions in the beanplots in Fig. 9 show that the visual appeal was generally rated higher (i.e. prettier) in the white beans which were the hard/pretty website conditions. This result accurately reflects the website's actual visual appeal levels,

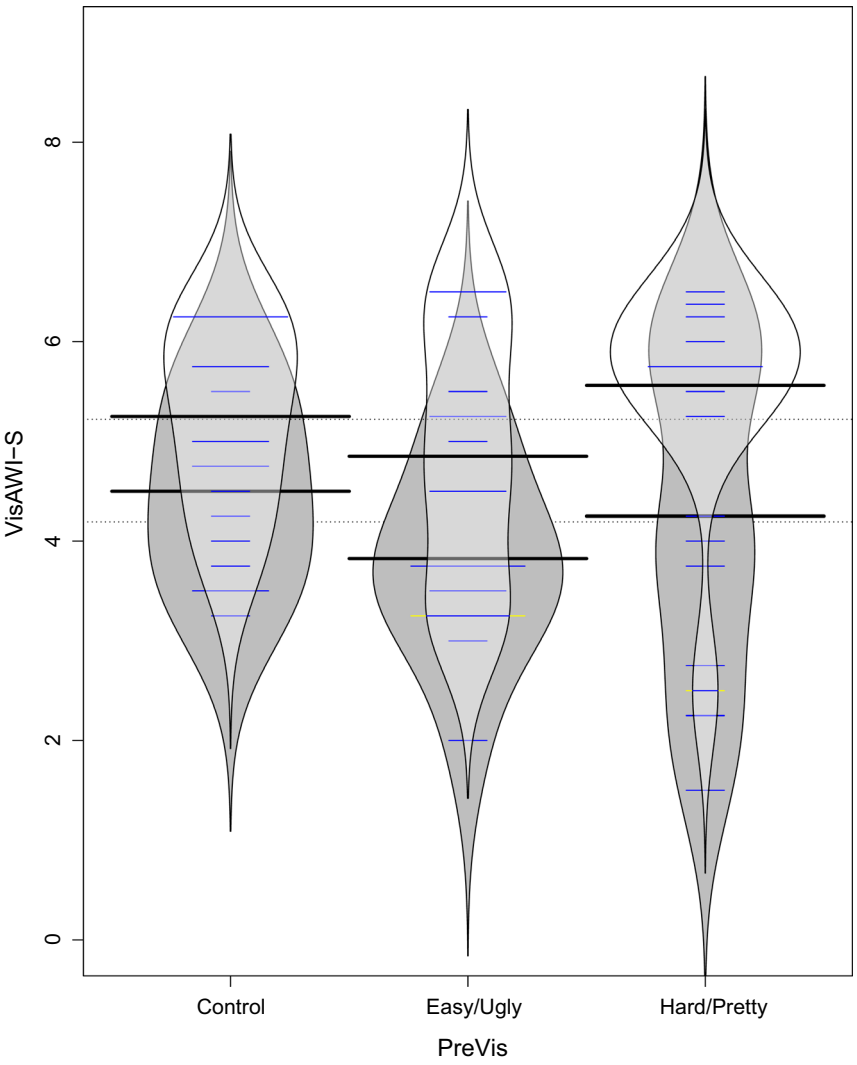

Fig. 9 Beanplot of the pre-use visual appeal results

with the prettiest condition being the hard/pretty website with hard/pretty WOM. The distribution of the easy/ugly website control group (the first grey bean) appears to be normal and has a mean that is higher than the two experimental groups of the same website. The easy/ugly website has the lowest mean, suggesting that the lower-quality information on the WOM did impact the visual appeal rating to be lower than the other two conditions. Also in the easy/ugly website, the hard/pretty WOM condition has a higher mean than easy/ugly. This suggests that the positive WOM marginally increased and the bad WOM marginally lowered the rating of visual appeal in the respective conditions, pre-use.

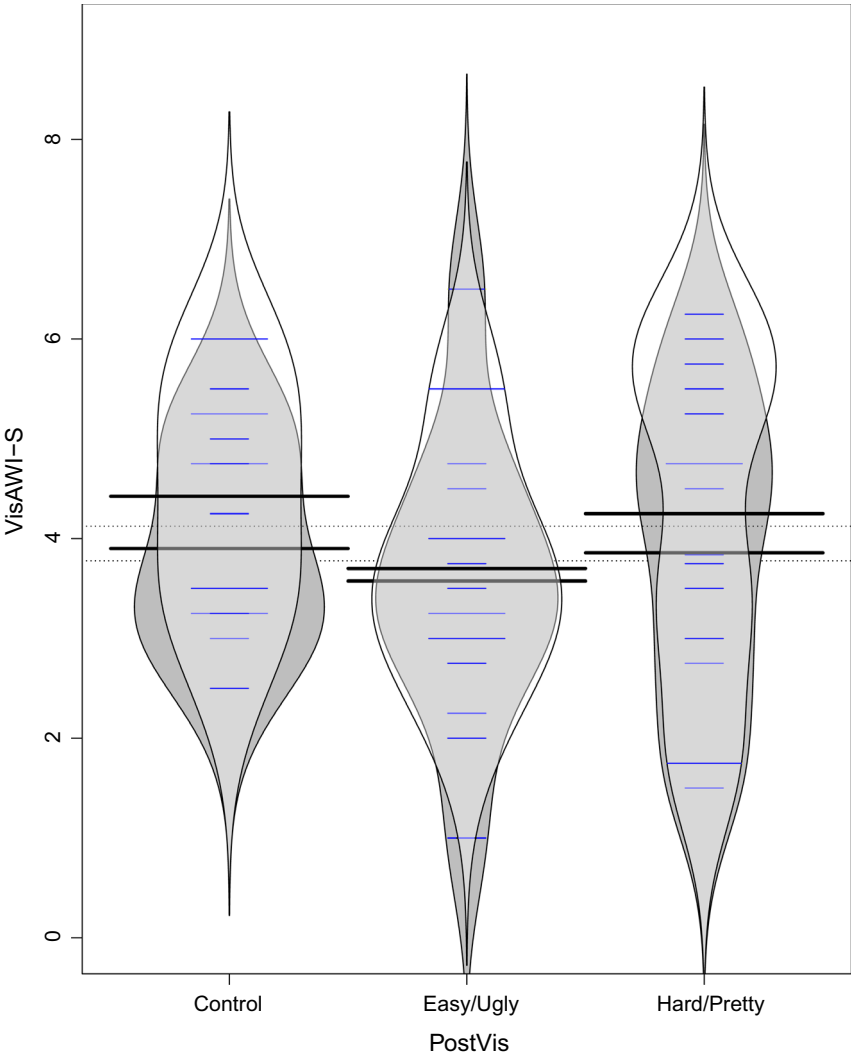

Fig. 10 Beanplot of the post-use visual appeal results

Also in Fig. 9, for the hard/pretty website, the visual appeal was rated highest when the WOM was positive for visual appeal, and lowest in the bad visual appeal WOM condition. However, the hard/pretty WOM condition (which corresponds to the actual website level) appears to be slightly bimodal, with a very small number of participants disagreeing with the high praises of visual appeal, and rate it as uglier than the bad visual appeal WOM category. For these couple of participants, the WOM worked inversely.

Post-use, visual appeal equalizes throughout all six conditions, as seen in Fig. 10. All of the ratings are lower post-use than they were pre-use. This suggests that website

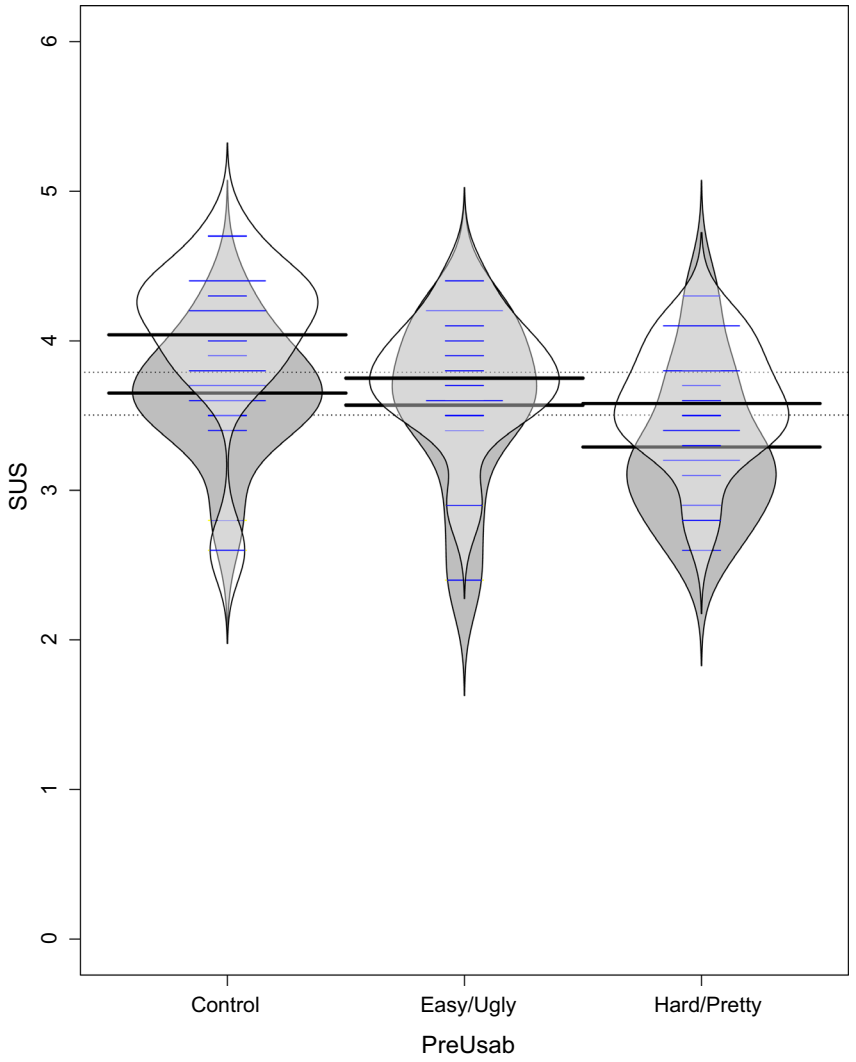

Fig. 11 Beanplot of the pre-use perceived usability results

use impacted the ratings of visual appeal. Specifically, low usability may have lowered the visual appeal ratings in the hard/pretty website.

Out of the slight variations between the means within a website, the group with the communication conveying a low visual appeal message has the lowest post-use visual appeal means, for both website versions. This does suggest that the low WOM impacted the perception of the website's appeal. The right- and leftmost columns have similar means, with the high visual appeal condition having a slightly higher distribution than the control group. The hard/pretty (white beans) website is still rated as slightly prettier than the low visual appeal website (grey beans), but the averages

largely do not portray the actual difference in visual appeal between the two website versions. For the easy/ugly website, the ratings slightly dropped for visual appeal, between pre- and post-use. In addition, post-use, as seen in Fig. 10, for the easy/ugly website, the ugliest rated was the bad WOM visual appeal condition, while the highest was the control condition but the positive WOM condition for visual appeal was very close second. A similar result occurred in the hard/pretty website. Statistical tests were done to determine if the difference was significant.

Pre-use usability ratings across all six conditions can be seen in Fig. 11. The control condition for the hard/pretty website (first column, purple bean) was rated as easiest to use, thanks to the high visual appeal of the website and no biased communication 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 positive message for usability is rated as second highest in usability pre-use, followed by the condition with bad WOM for usability, in the hard/pretty website. The bad WOM seems to show more variance compared to the control condition, suggesting the impact of low WOM was stronger than the positive. This result is consistent with Study 2's pre-use usability ratings, in which judgments for pre-use usability were strongly based on pre-use visual appeal and on the WOM. For the easy/ugly website, the results follow the same trend as well.

The usability ratings completely change post-use (as seen in Fig. 12 when compared to the pre-use ratings in Fig. 11). Post-use ratings normalize across all conditions. Post-use usability ratings dropped for the hard/pretty website (white beans), to better reflect the website's poor usability. Within the easy/ugly website, the highest rated usability levels came from the positive usability WOM group, while the lowest are from the bad WOM condition. Again, showing evidence that WOM impacts user perception of usability. Statistics were done to determine the significance of these findings.

6.3.2 Assumptions testing 3

Shapiro–Wilk tests showed that pre-use visual appeal at hard/pretty website with hard/pretty WOM ($p = .001$) and pre-use usability for the hard/pretty 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 hard/pretty website with hard/pretty WOM was − 2.498 (SE = 0.687), the kurtosis measure of 7.038 (SE = 1.334). Moreover, pre-use usability in the hard/pretty website 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 nonparametric 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.

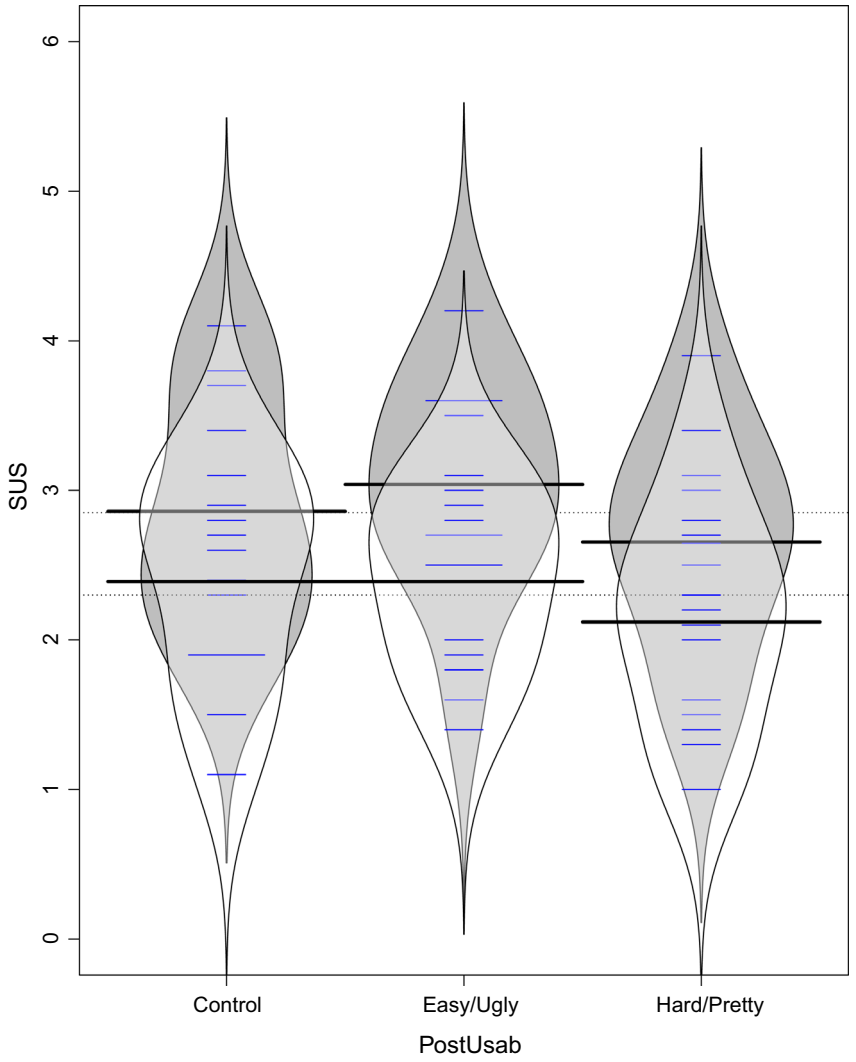

Fig. 12 Beanplot of the post-use perceived usability results

6.3.3 Statistical hypothesis testing 3

Only pre-use perceived usability ($p < .05$) was found to vary in the hard/pretty website conditions. Paired comparisons showed that hard/pretty website with hard/pretty. WOM and the control condition differed in pre-use usability ($p < .05$). Therefore, partial statistical evidence exists that usability was rated lower when WOM was set to be low, especially between the control group and the low WOM group, which rated the website as harder to use. This difference was only found pre-use, suggesting that the impact was only strong enough to influence ratings before having been exposed

Table 5 Correlations between visual appeal and perceived usability for easy/ugly control condition and the hard/pretty control condition, respectively

Easy/ugly control				Hard/pretty control			
	PreUsab	PostVis	PostUsab		PreUsab	PostVis	PostUsab
PreVis	.532	.443	.273	PreVis	.290	.466	.254
PreUsab	–	.577	.661*	PreUsab	–	−.092	.202
PostVis		–	.545	PostVis		–	.541

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

to the website for roughly an hour. However, this difference was neither found in the easy/ugly 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. WOM did not seem to significantly impact use. Thus, there is insufficient statistical evidence to conclude that participants struggled more when WOMs were negative about a variable or that they did better when WOMs were positive about a variable.

6.3.4 Correlations study 3

6.3.4.1 Correlations in the control conditions In this section, the results for the Spearman correlations, ρ, between visual appeal and perceived usability within each control condition (easy/ugly control condition and hard/pretty control condition; $n = 10$ in each) are seen in Table 5.

In the easy/ugly control condition, pre- and post-use usability ($\rho = .661$, $p < .05$) was highly and positively correlated. This suggests that their opinions on usability did not change much with use, or that they changed in the same direction. Visual appeal and usability were neither significantly correlated in easy/ugly control condition nor in hard/pretty control condition. This suggests that participants' options on visual appeal changed after use and that they were separately judging usability and visual appeal.

In the easy/ugly control condition, none of the objective usability measures were correlated. However, in hard/pretty control condition, hovers were correlated with time per task ($\rho = .891$, $p < .01$) and passes ($\rho = -.920$, $p < .01$). The success rate (passes) and time were also correlated ($\rho = -.963$, $p < .01$). These results confirm that they measure the same construct (i.e. objective usability).

6.3.4.2 Correlations when WOM and website usability and visual appeal levels are congruent All Spearman correlations between visual appeal and perceived usability with conditions where WOM levels were congruent with the actual website levels can be seen in Table 6. In the easy/ugly website with easy/ugly WOM, pre- and post-use perceived usability were highly and positively correlated ($\rho = .657$, $p < .05$). Pre-use perceived usability was also correlated highly and positively with pre-use visual appeal

Table 6 Correlations between visual appeal and perceived usability for easy/ugly website with easy/ugly WOM and hard/pretty website with hard/pretty WOM, respectively

Easy/ugly website, easy/ugly WOM				Hard/pretty website, hard pretty WOM			
	PreUsab	PostVis	PostUsab		PreUsab	PostVis	PostUsab
PreVis	.827**	.904**	.870**	PreVis	.421	.302	.325
PreUsab	–	.848**	.657*	PreUsab	–	−.040	.073
PostVis		–	.835**	PostVis		–	.888**

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

($\rho = .827$, $p < .01$). In addition, post-use perceived usability was correlated with post-use visual appeal ($\rho = .835$, $p < .01$).

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.

In the hard/pretty website with hard/pretty WOM, 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.

In addition, for the easy/ugly website with easy/ugly WOM, post-use usability was correlated with passes ($\rho = .652$, $p < .05$), passes was also correlated with clicks, ($\rho = -.656$, $p < .05$), and clicks was correlated with time ($\rho = .669$, $p < .05$). In the hard/pretty website with hard/pretty WOM, hovers and clicks were correlated ($\rho = .731, p < .05$). 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 WOM 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.

6.3.4.3 Correlations when WOM and website levels are incongruent All Spearman correlations between visual appeal and perceived usability in conditions where WOM of these were incongruent with the actual website levels can be seen in Table 7. In the easy/ugly website with hard/pretty WOM condition, pre- and post-use visual appeal was positively and significantly correlated ($\rho = .896$, $p < .05$). This suggests that their first impressions of visual appeal did not change after use. This was not the case with

Table 7 Correlations between visual appeal and perceived usability for easy/ugly website with hard/pretty WOM and hard/pretty website with easy/ugly WOM, respectively

Easy/ugly website, hard/pretty WOM				Hard/pretty website, easy/ugly WOM			
	PreUsab	PostVis	PostUsab		PreUsab	PostVis	PostUsab
PreVis	.329	.896**	.067	PreVis	.641*	.446	−.071
PreUsab	–	.494	.511	PreUsab	–	.537	−.189
PostVis		–	.438	PostVis		–	.319

*Significant at 0.05 (two-tailed)
**Significant at 0.01 (two-tailed)

the hard/pretty website with easy/ugly WOM, where only pre-use visual appeal and pre-use usability were moderately and positively correlated ($\rho = .641$, $p < .05$).

This shows that pre-use, participants judged these two similarly even though the WOM 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.

In easy/ugly website with hard/pretty WOM condition, time and passes were correlated ($\rho = -.916^{**}$, $p < .01$), evidence that there were fewer passed tasks with longer times (as per the definition of passes). In hard/pretty website with easy/ugly WOM, pre-use visual appeal was correlated with passes ($\rho = .719$, $p < .05$), the number of clicks was correlated with time ($\rho = .697$, $p < .05$), and time was correlated with passes ($\rho = -.755^{*}$, $p < .05$).

6.4 Discussion 3

6.4.1 Results summary 3

6.4.1.1 Beanplots High visual appeal in both the website and WOM 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 WOM were low. Thus, WOM did seem to influence the perception of visual appeal and usability, more vividly pre-use, supporting the hypotheses.

6.4.1.2 Statistics Hard/pretty website with hard/pretty WOM ($\bar{x} = 3.58$) was rated lower than the control group for the hard/pretty website ($\bar{x} = 4.04$) in pre-use usability. Visual appeal was not influenced by WOM. No significant results were found for objective usability. Thus, insufficient evidence exists determine the impact of WOM when visual appeal and usability were incongruently levelled.

6.4.1.3 Correlations Based on the correlations, in the easy but ugly website (no WOM), 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

WOM), 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 WOM, 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 WOM (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 WOM, where participants judged visual appeal and usability similarly, but only pre-use.

6.4.2 Limitations and future research 3

6.4.2.1 Threats to statistical validity The relatively small sample sizes may have been the largest factor in the lack of significance in the statistical testing, especially for Study 3. 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.

6.4.2.2 Threats to internal validity One possible issue with the implementation of WOM, as was found before, would be unfamiliarity of the location and experimenter, which could have influenced WOM 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 WOM (i.e. one that is high and one that is low) was a bit confusing for participants. Future studies should strengthen the wording from the confederate to see if that would have a greater impact on participants.

7 General discussion

7.1 Result summary

7.1.1 Study 1 beanplots

A small trend did emerge post-use: participants in the high WOM group rating it as easiest and low WOM group rating it as hardest.

7.1.2 Study 2 beanplots

The verbal reinforcement of the textual implementation of WOM affected participants' perceptions of visual appeal and usability, pre- and post-use.

7.1.3 Study 3 beanplots

WOM did seem to impact the perception of visual appeal and usability, more vividly pre-use.

7.1.4 Statistical result summary

WOM influences the perception of visual appeal and usability, more so when visual appeal and usability levels are congruent and when textual WOM are reinforced verbally, irrespective of use. Congruency thus allowed for an easier transmission of information with little to no confusion.

In the case when WOMs and the website had incongruent usability and visual appeal levels, then it seems that only low verbal and textual WOM of usability influence the perception of usability, pre-use. In addition, the ugly website seemed to lower the usability rating as the colours distracted 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. 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. More work is needed to gain more insight into the nuances of the impact.

7.1.5 Correlations summary

Figures 13, 14, and 15 are created in R. For Study 1, Spearman correlations were examined between the pre- and post-use visual appeal and usability, without taking the conditions into account, seen in Fig. 13. If there were no conditions, then visual appeal and perceived usability pre-use were positively and strongly correlated ($\rho = .657$, $p < .001$). Post-use, the correlation between visual appeal and perceived usability was slightly weaker but still positive and significant ($\rho = .585$, $p < .001$).

For Study 2, to get an understanding of the data as a whole, Spearman correlations were examined between visual appeal and perceived usability, pre- and post-use without taking into consideration the different conditions. They can be seen in Figure 14. Across the good website, pre-use perceived usability was positively and significantly correlated to both post-use usability ($\rho = .669$, $p < .001$) and pre-use visual appeal ($\rho = .419$, $p < .05$). Similarly, post-use visual appeal was highly and positively correlated with both pre- ($\rho = .564$, $p < .01$) and post-use ($\rho = .543$, $p < .01$) perceived usability. In addition, pre- and post-use visual appeal ($\rho = .524$, $p < .01$) were all also positively and moderately correlated. As with the findings of the good website correlations in Study 1, these correlations also agree with the literature and show a relationship between usability and visual appeal pre-and post-use.

In Study 3, Spearman correlations were first examined between all the variables, without taking the conditions into account ($n = 60$), seen in Fig. 15. If there were no conditions, then visual appeal and perceived usability pre-use were positively and moderately correlated ($\rho = .479$, $p < .01$). Post-use, the correlation between visual appeal and perceived usability was slightly weaker but still positive and significant ($\rho = .426$, $p < .01$).

Therefore, as WOM became more apparent (i.e. verbal and textual) and more complicated with different levels of usability and visual appeal, the correlation strength became weaker.

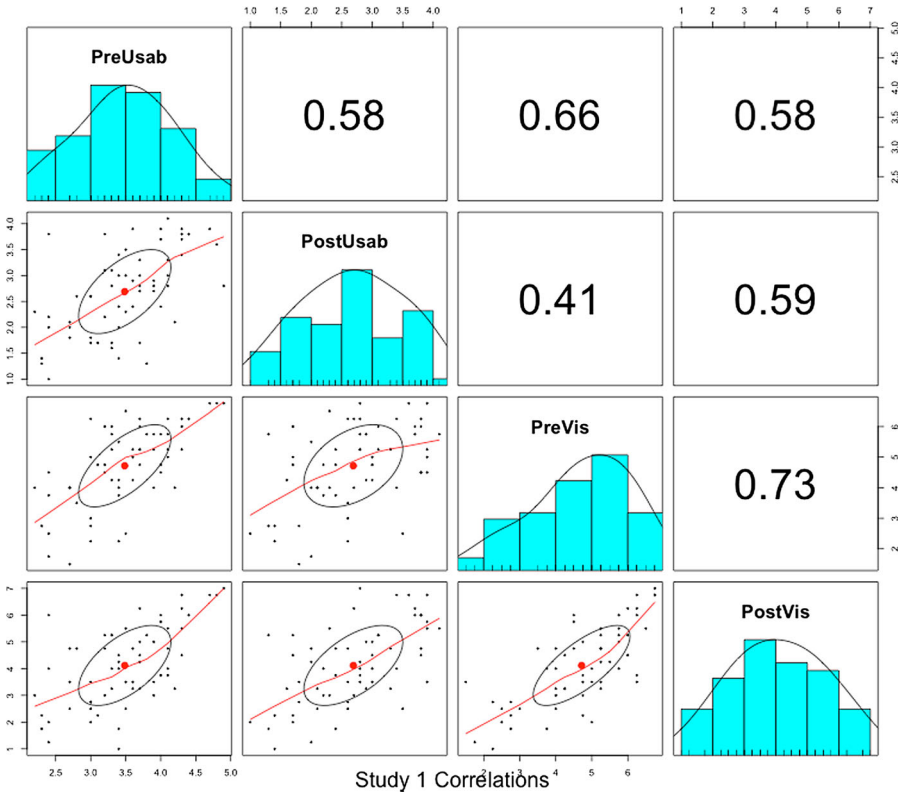

Fig. 13 Correlations across all conditions for Study 1

7.2 Implications for theory

The findings in this study may still support the cognitive dissonance theory, in that participants all internalized the WOMs 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 WOMs, others may have gone with the other approach and disagreed with them. 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 WOMs (positive levels of one variable and negative levels of the other) could have been confusing; this is further discussed in the next section.

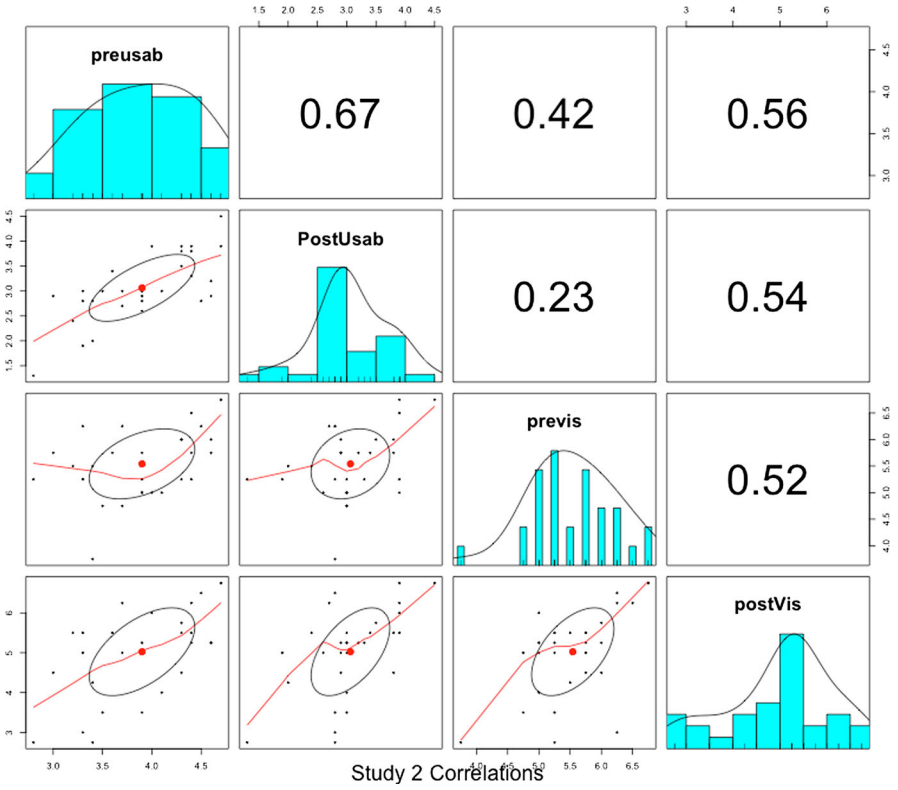

Fig. 14 Correlations across all conditions for Study 2

7.3 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 using WOM facilities such as 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 use WOM facilities, 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.

8 Conclusion

Overall, this research contributes to an improved understanding of the relationship of usability and visual appeal by added understanding of the effects WOM have on these

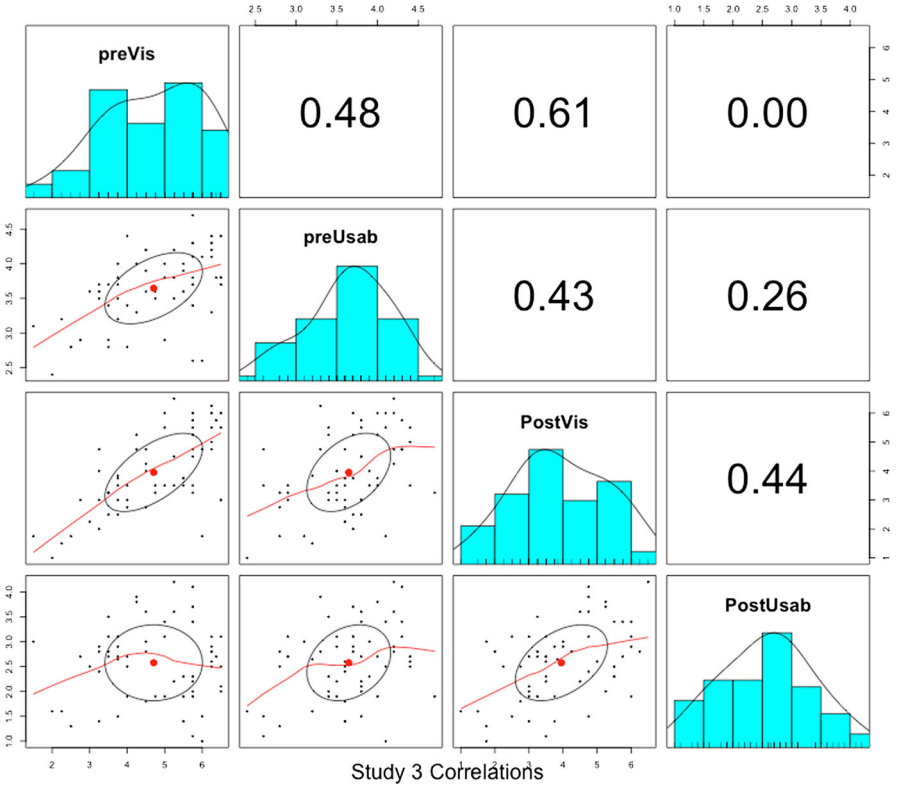

Fig. 15 Correlations across all conditions for Study 3

variables, in a web environment. WOM influence the perception of visual appeal and usability, more so when visual appeal and usability levels are congruent and when textual WOM are reinforced verbally, irrespective of use. The practical application of the findings here is that talking up your web service via social networks and media really do help with how easily people use it and how visually appealing they think it is—especially if it actually is well made. So, investing some time and money into the marketing of it through users and YouTubers, for example, may impact your likability as well.

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.

The value of this research is important to any new website but it can also be applied to established websites. For example, the results of this research suggest that government websites can overcome their negative bricks and mortar reputations by replacing them with a more usable, pleasing online alternative. Adding strategic WOM advertising would aid their reputations even more.

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.

Acknowledgements Thanks to Milos Stojmenovic for creating the website data sample and thanks to Gitte Lindgaard for advice on previous related work.

References

1. Heidmann F (2009) Human–computer cooperation. In: Bullinger HJ (ed) Technology guide. Springer, Heidelberg
2. Litvin S, Goldsmith R, Pan B (2008) Electronic word-of-mouth in hospitality and tourism management. Tour Manag 29(3):458–468
3. Forrester Research: The State of Retailing Online (2006) The 9th annual shop.org study. www.clickz.com/3611181
4. Burtuskova A, Krejcar O Evaluation framework for user preference research implemented as web application. In: Badica C, Nguyen NT, Brezovan M (eds) ICCCI, 201, LNCS, vol 8083, Heidelberg, pp 537–548
5. ISO 9241/11 (1996) International organization for standardization. Retrieved June 2012
6. Feagin SF (1995) Beauty. In: Audi R (ed) The Cambridge dictionary of philosophy Cambridge. Cambridge University Press, Cambridge, p 66
7. Blijlevens J (2011) Typically the best? Perceived typicality and aesthetic appraisal of product appearances. PhD thesis, Delft University of Technology, The Netherlands
8. Gefen D, Karahanna E, Straub D (2003) Trust and TAM in online shopping: an integreated model. MIS Quart 27(1):51–90
9. Stojmenović M, Pilgrim C, Lindgaard G (2014) Perceived and objective usability and visual appeal in a website domain with a less developed mental model. In: The 26th ACM Australian Computer–Human Interaction Conference OZCHI'14, December 2–5, 2014, Sydney, NSW, Australia, pp 316–323
10. Stojmenović M, Grundy J, Farrell V, Biddle R, Hoon L (2016) Does textual word-of-mouth affect look and feel? In: The 28th ACM Australian Computer–Human Interaction Conference OZCHI'16, November 29–December 2, Launceston, Tasmania, Australia, pp 257–265
11. Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford
12. Harmon-Jones E, Amodio DM, Harmon-Jones C (2009) Action-based model of dissonance: a review, integration, and expansion of conceptions of cognitive conflict. In: Zanna MP (ed) Advances in experimental social psychology, vol 41. Academic Press, Burlington, pp 119–166
13. Tractinsky N, Katz AS, Ikar D (2000) What is beautiful is usable. Interact Comput 13(2):127–145
14. Dion KK, Berscheid E, Walster E (1972) What is beautiful is good. J Personal Social Psychol 24:285–290
15. Thorndike EL (1920) A constant error in psychological ratings. J Appl Psychol 4(1):25–29
16. Hall RH, Hanna P (2004) The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. Behav Inf Technol 23(3):183–195
17. 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, p 396
18. Tractinsky N (1997) Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: Proceedings of ACM SIGCHI, New York, ACM, pp 115–122
19. Norman D (2004) Emotional design: Why we love (or hate) everyday things. Basic Books, New York
20. Katz A (2010) Aesthetics, usefulness and performance in user-search-engine. Int J Appl Quant Methods 5(3):424–445
21. Tuch AN, Roth SP, Hornbæk K, Opwis K, Bargas-Avila JA (2012) Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. Comput Hum Behav 28(5):1596–1607
22. De Angeli A, Sutcliffe A, Hartmann J (2006) Interaction, usability and aesthetics: What influences users' preferences?. University Park, College Town, pp 271–280
23. Lee S, Koubek RJ (2010) Understanding user preferences based on usability and aesthetics before and after actual use. Interact Comput 22(6):530–543

24. Sangwon L, Koubek RJ (2010) Understanding user preferences based on usability and aesthetics before and after actual use. Interact Comput 22(6):530–543
25. 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 and Information Technology, vol 117, EcomIT & GBM, Sri Lanka
26. Quinn JM, Tran TQ (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, New York, pp 353–362
27. McLellan S, Muddimer A, Peres SC (2012) The effect of experience on system usability scale ratings. J Usability Stud 7(2):56–67
28. Ip C, Law R, Lee H (2011) A review of website evaluation studies in the tourism and hospitality fields from 1996 to 2009. Int J Tour Res 13(3):234–265
29. Sokkar A, Law E (2013) Validating an episodic UX model on online shopping decision making: a survey study with B2C e-commerce. IN: Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp 297–306
30. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
31. Herr PM, Kardes FR, Kim J (1991) Effects of word-of-mouth and product-attribute information on persuasion: an accessibility-diagnosticity perspective. J Consum Res 17(4):454–462
32. Smith D, Menon S, Sivakumar K (2005) Online peer and editorial recommendations, trust, and choice in virtual markets. J Interact Market 19(3):15–37
33. Pavlou PA, Dimoka A (2006) The nature and role of feedback text comments in online marketplaces: implications for trust building, price premiums, and seller differentiation. Inf Syst Res 17(4):392–414
34. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340
35. Ba S, Pavlou PA (2002) Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. MIS Q 26(3):243–268
36. Chevalier J (2006) The effect of word of mouth on sales: online book reviews. J Market Res 43(3):9
37. Vermeulen IE, Seegers D (2009) Tried and tested: the impact of online hotel reviews on consumer consideration. Tour Manag 30(1):123–127
38. Voss B (1984) Slips of the ear: Investigations into the speech perception behaviour of German speakers of English. G. Narr, Tübingen
39. Chandy R, Gu H (2012) Identifying spam in the iOS app store. Paper Presented, pp 56–59. https://doi.org/10.1145/2184305.2184317
40. Ellison G, Fudenberg D (1995) Word-of-mouth communication and social learning. Q J Econ 110(1):93–125. https://doi.org/10.2307/2118512
41. Asch SE (1951) Effects of group pressure on the modification and distortion of judgments. In: Guetzkow H (ed) Groups, leadership and men. Carnegie Press, Pittsburgh, pp 177–190
42. Asch SE (1955) Opinions and social pressure. Sci Am 193:31–35
43. Asch SE (1956) Studies of independence and conformity. A minority of one against a unanimous majority. Psychol Monogr 70(9):1–70
44. 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
45. Druckman JN, Kam CD (2009) Students as experimental participants: a defense of the "narrow data base"*. SSRN: http://ssrn.com/abstract=1498843
46. Brooke J (1986) System usability scale (SUS). © Digital Equipment Corporation, UK
47. Moshagen M, Thielsch M (2012) A short version of the visual aesthetics of websites inventory. Behav Inf Technol 32(12):1305–1311
48. Kruskal W (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621
49. Kampstra P (2008) Beanplot: a boxplot alternative for visual comparison of distributions. J Stat Softw 28(1):1–9
50. Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J Stat Model Anal 2(1):21–33
51. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611
52. Cramer D, Howitt D (2004) The SAGE dictionary of statistics. SAGE, London
53. Doane DP, Seward LE (2011) Measuring skewness. J Stat Educ 19(2):1–18

54. Martin WE, Bridgmon K (2012) Quantitative and statistical research methods: from hypothesis to results. Wiley, Somerset
55. Nordstokke DW, Zumbo BD, Cairns SL, Saklofske DH (2011) The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. Pract Assess Res Eval 1(5):1–8
56. Rettig M (1991) Nobody reads documentation. Commun ACM 34(7):19–24
57. Felder RM (1993) Reaching the second tier: learning and teaching styles in college science education. J Coll Sci Teach 23:286–290