City University London
MSc in Human-Centred Systems thesis report
2015

Is 5 seconds enough? An empirical investigation of the online 5 second test.

William Deng

Supervised by: Stephanie Wilson

Word count: 18,481 words

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: K Deng

Abstract

The online 5 second test evaluates the clarity of content on digital products such as websites. In this test, participants view a screenshot of a website for 5 seconds and then answer questions about what they have seen.

This technique has risen in popularity as it is low cost, easy to set up and can collect large amounts of data. This is the first project to use empirical evidence to investigate the effectiveness of this technique.

A testing platform was created, and two online studies were carried out. The first study explored how changing the exposure time affects participants' responses. The second study investigated the effects on user's responses when they saw the questions from the test before they viewed the screenshot.

The data collected from 346 participants show that 5 seconds may not be enough. The exposure time depends on the level of detail that is required to answer a question correctly. Participants in the 5 second test will write short responses, typically two to three words, regardless of exposure time or whether they have seen the question beforehand. Although the number of words in the responses remains relatively constant, the level of detail does increase with exposure time.

Acknowledgements

I would like to thank my advisor, Stephanie, for her time and advice. The knowledge that I gained whilst completing this research is a valuable addition to my education and could not have been possible without her support.

Secondly, I would like to thank Sophie for believing in me and for her moral support.

1. INTRODUCTION AND BACKGROUND 8

1.1 Research questions 9
1.1 Objectives 10

1.3 Beneficiaries 10

1.4 Report structure 10

2. LITERATURE REVIEW 12

2.1 Background to the evaluation of systems 12

2.2 The importance of usability evaluation 12

2.3 An overview of remote usability evaluations 13
2.3.1 Synchronous remote evaluations 14
2.3.2 Asynchronous remote evaluations 15
2.3.3 Remote evaluations in practice 16

2.4 The evolution of the 5 second test 16
2.4.1 The original 5 second test 16
2.4.2 The online 5 second test 18
2.4.3 Research into the online 5 second test 19

2.5 Review of literature concerning first impression judgements on aesthetics 21

2.6 A model for information processing 22

3. CREATING THE PLATFORM TO BE USED FOR THE STUDIES 24

3.1 Requirements of the platform 24

3.2 Components of the platform 25

3.3 Technical details 29

3.4 Designing the user interface 30

4. STUDY 1: INVESTIGATING EXPOSURE TIMES IN THE 5 SECOND TEST 31

4.1 Methodology for study 1 31
4.1.1 Variables 31
4.1.2 Choice of stimuli 32
4.1.3 Description of the stimuli 33
4.1.4 Producing a scoring system to measure correctness 35
4.1.5 Analysing the expert answers 35
4.1.6 Procedure for the study 37
4.1.7 Participant recruitment process 38

4.2 Data preparation 38
4.2.1 Procedure for data analysis 38
4.2.2 Removing data from unsuitable participants 39

4.3 Findings from study 1 40
4.3.1 Participants in this study 40

4.3.2 Exposure time and correctness scores 41
4.3.3 Statistical analysis on the correctness scores 42
4.3.4 Qualitative findings 42
4.3.5 Exposure time and word count 44
4.3.6 Statistical analysis 45
4.3.7 Exposure time and visual appeal ratings 46
4.3.8 Statistical analysis 47

5. STUDY 2: QUESTION ORDER IN THE 5 SECOND TEST 49

5.1 Methodology for study 2 49
5.1.1 Variables 49
5.1.2 Procedure 50

5.2 Findings for study 1 51
5.2.1 Participants 51
5.2.2 Question order and score awarded for correctness 51
5.2.3 Statistical tests 53
5.2.4 Qualitative findings 53
5.2.5 Question order and word count 54
5.2.6 Statistical tests 55
5.2.7 Question order and visual appeal ratings 55
5.2.8 Statistical tests 56

6. DISCUSSION 57

6. 1 How does exposure time to a website affect participants' responses? 58
6.1.1 Participants with more time read more words 58
6.1.2 The questions in the test matter 59
6.1.3 The expected answers to the questions matter 59
6.1.4 Exposure time does not affect the length of responses 60

6.2 How does exposure time to a website affect participants' perception of its visual appeal? 60

6.2.1 Pictures on a website have a large impact 61

6.3 What difference does it make if the test questions are shown to the participant before they begin the experiment compared to when they see it afterwards? 61

6.5 Additional analysis 62

6.6 Practical implications 62
6.6.1 Exposure time depends on the expected answers 62
6.6.2 Using tag clouds for analysing 5 second test data 63
6.6.3 Determining whether to show the questions or not 63
6.6.4 Collecting more data to inform improvements 64
6.6.5 Areas for further research 64

6.4 Feedback about the studies conducted in the project: 65

7. CONCLUSION 66

8. REFLECTION 67

REFERENCES/BIBLIOGRAPHY 69

APPENDICES

Appendix A – Project proposal
Appendix B - Link to the experimental platform built for this project.
Appendix C – Brief for the analyses with the experts
Appendix D – Results from the expert analyses (expert answers)
Appendix E – List of synonyms used in the data analysis
Appendix F – Responses, correctness scores, word counts and visual appeal ratings for each participant in all of the groups
Appendix G – List of advertisement locations
Appendix H – Output of statistical tests
Appendix I – Descriptive statistics
Appendix J – List of participants by country
Appendix K – Breakdown of completed tests
Appendix L – Feedback about the studies

1. Introduction and background

We naturally judge whatever we see.

Our time is precious, so we make quick judgements about everything so that we can make rapid decisions. In our daily lives, within the blink of an eye, we have already formed a judgement about another person (Gladwell, 2006).

With the explosion of digital content, first impressions of digital products matter even more as users can form first impressions of a website in as little as 50 milliseconds (Lindgaard et al. 2006). When we look for information, before we invest time and effort into reading it, we judge whether the content is relevant to our needs. We think about whether a website looks good and easy to use to make a judgement about its other characteristics like its trustworthiness or security (Cyr and Head 2013).

More and more Human Computer Interaction researchers and UX practitioners have begun using remote evaluation methods to assess their Digital products and services (Albert et al. 2010; Molich et al. 2010; Reinecke and Gajos, 2014). Collecting data over the internet is nothing new. Technologies such as Microsoft NetMeeting or Skype enable evaluators to test a user from the comfort of their home (Thompson et al. 2004; Petrie et al. 2006). Users set up in their natural environment thus avoiding the artificial conditions of a lab, this can avoid distorting users' natural behaviours (Jewell and Salvetti, 2012). Users can complete tasks while sharing their screen and communicating to evaluators if and when necessary. These synchronous remote testing methods are popular, but they require the evaluators to facilitate the test sessions, thus limiting the potential number of participants.

Digital products continue to expand globally, as more and more people come online. As of 2015, there are over 3 billion internet users across the world. At the same time, the practice of UX is gaining prominence within large corporations (Harvard Business Review, 2013). These trends have created a need for statistical evaluation methods involving larger sample sizes with participants from around the globe (Albert, Tullis and Tedesco, 2010). This has led to the increasingly popularity of asynchronous remote testing methods where evaluations can be conducted without the presence of a researcher (Andreasen et al. 2007; Molich et al. 2010). These techniques allow evaluators to collect large amounts of data quickly and from participants anywhere in the world (Bruun et al. 2009). To date, limited research has been conducted on the effectiveness of these newer remote evaluation methods. Examples of these methods include unmoderated remote user testing, A/B tests, card sorting, click testing and the topic of this research, the online 5 second test.

The original 5 second test was created by Christine Perfetti from User Interface Engineering. The original 5 second test used before was a traditional usability test where participants were invited into a lab to test a website (Perfetti. 2004). Prior to beginning the usability tasks, participants would be given a focused task and then asked to remember everything that they could after a web page is shown for 5 seconds. When 5 seconds is up, the moderator asks the participants to write down everything they remember followed by two questions. This 5 second test was intended for evaluating content pages and can highlight what stands out the most on a design. It achieves this by understanding what information a user has absorbed after viewing the design for a limited amount of time.

Since the creation of this technique, some online variations have arisen. These variations are provided for free or on a fee-basis by providers such as UsabilityHub or Zurb. These are the type of 5 second tests that this project investigates. The procedure for these online 5

second tests departed from the original test as participants did not have to write everything down after viewing the website and were not prescribed a focused task. In addition to these changes, designers began using the online 5 second tests to test all kinds of web pages for like homepages and landing pages.

In the UX and marketing industries, the 5 second test has been used to optimise on the understandability of content, to optimise conversions, to test calls to actions, and to reduce bounce rates (Useful Usability, 2014; Measuringu.com, 2015). Despite the widespread usage of the 5 second tests by both usability and online marketing professionals (Useful Usability, 2014; Measuringu.com, 2015), there is yet to be any convincing evidence as to why 5 seconds is chosen as the de facto exposure time necessary to communicate the value of a website to a user.

In the only study that investigated online 5 second tests, the author Doncaster (2014) stated that there are no "clear cut" answers to find out why 5 seconds was chosen. Given that several large scale empirical studies have shown that users dwell on web pages for more than 10 seconds (Liu et al. 2010; Weinreich et al. 2008), there is an obvious need to investigate whether 5 seconds is adequate. Additionally, the fact that tasks are no longer featured in many of the tests raises a question about whether it was necessary in the first place.

The rather dramatic changes to the original 5 second test seem to have been left unnoticed and rendered unimportant, yet the online 5 second test is more widely used than ever before. Doncaster (2014) who has been working closely with UsabilityHub report that more than 100,000 unique 5 seconds tests are completed on its site in a calendar year. For these reasons, the 5 second test technique deserves further investigation.

I set out to examine the technique in detail and investigate its effectiveness.

1.1 Research questions

The overall goal of this project is to investigate the effectiveness of the online 5 second test.

This goal is broken down into three research questions:

1. How does exposure time to a website affect participants' responses?

2. How does exposure time to a website affect participants' perception of its visual appeal?

3. What difference does it make if the test questions are shown to the participant before they begin the experiment compared to when they see it afterwards?

By gaining insight to these research questions, the researcher aims to provide evidence to investigate the effectiveness of the online 5 second test.

1.1 Objectives

To answer these research questions effectively, three key objectives was met:

- An online testing platform was built. This platform is similar to the existing online 5 second tests but had additional functionality which enabled the researcher to manipulate exposure times, change the type of response fields and change whether the questions are shown before viewing the screenshot or not.

- The platform was used to conduct two separate online studies. The first study manipulated the exposure time to see how user's responses would change and to explore how this would affect user's perception of visual appeal. The second study investigated the effects of showing the test questions to the participant before they see the website.

- The data that was collected was analysed which provided empirical evidence to gain insights into the research questions.

1.3 Beneficiaries

Possible beneficiaries of this project include:

- HCI researchers - Insight from this research may further their understanding of the "5 second test" technique. It also complements the work that has been conducted on first impressions of visual appeal by Lindgaard et al., (2006) and Reinecke and Gajos, 2014.

- UX practitioners - The prototype platform can be employed as a template for designing variations on the standard "5 second test" to suit the different needs of practitioners.

- The author - The knowledge and experience acquired by conducting this research have increased the author's knowledge in this field and better his research skills.

1.4 Report structure

Relevant academic literature has been reviewed, and this is presented in chapter 2 to provide context for the project, explore findings from previous studies and discover shortcomings that this project can improve upon.

Chapter 3 chapter covers the design and implementation of the prototype testing platform. Also, a description of what was created and how it was created is provided.

The findings from each study are presented and discussed in separately to maximise coherence for the reader. Chapter 4 focuses on the first study and chapter 5 focuses on the second study.

The final chapter of this report consists of a conclusion and reflection to provide readers with a summary of the key findings, discuss areas for future research and demonstrate what has been learnt from this study.

2. Literature review

In this section, academic literature that is relevant to this project is reviewed to provide context for this project.

2.1 Background to the evaluation of systems

The focus on the efficiency of systems was brought into the limelight in 1911 when Frederick Taylor published the Principles of Scientific Management. Although, this work focuses on industrial work practices, within it, there are descriptions of time and motion studies aimed at increasing the organisation and productivity of systems. Many of the principles of these evaluative studies are still applied when evaluating interactive systems. It is amazing how far science has progressed.

Since the 1940's, the study of Human Factors Engineering (also known as ergonomics) has been conducting research and implementing solutions to deal with "man-machine" problems. Alphonse Chapanis, the father of ergonomics, published the first ergonomics textbook in 1949. In this ground-breaking book, Chapanis introduces the principles of designing effective visual displays, optimising auditory communications as a means of information transfer, how to improve operational controls and how to develop better arrangements of individual workplaces and tools. In addition to founding a new field of science, Chapanis was also the first person to use statistics to analyse of errors that occurred in workplace situations. Chapanis's work and study of human factors engineering have had a significant impact on early studies into the evaluation of Interactive Systems.

In MacDonald and Atwood, (2013)'s paper, they describe the history and evolution of evaluation within the domain of Human Computer Interaction. In their paper, they point out that the "changes in technology and use contexts" influences an evaluator's choice of evaluation method. This could not be truer today.

In what they describe as the "User Experience" phase (the 2000s - present), the evolution and maturing of the internet has expanded the context of use for technology. This has led to the increased importance of evaluating the hedonic attributes of a system, as well as its pragmatic attributes (Rogers et al. 2011). MacDonald and Atwood's paper provides a useful overview to the history of HCI, but their recommendations are lacking in that they fail to mention how current software development methodologies such as AGILE have impacted the choice of system evaluation methods.

2.2 The importance of usability evaluation

The value of usability evaluation is unquestionable in the modern world of information technology (Davis, 1989; Venkatesh et al., 2003). Various empirical studies have proven that feedback from evaluations improve product quality, reduce maintenance costs and support costs (Nielsen, 1993).

The benefits of usability and usability evaluation have long been widely accepted in academic research, but it is only recently that the practice of User Experience gained

acceptance in practice (Harvard Business Review, 2013). Usability has evolved from a niche role found only in gigantic technology companies to being so widespread that most large companies have their own in-house teams.

There are three types of usability evaluations (Rogers et al, 2011):

- Those in controlled settings involving users (traditional lab testing).
- Those in natural settings involving users (remote testing).
- And those in any setting not involving users otherwise know as inspection methods.

These evaluations can be either formative or summative. Formative evaluations are used to detect and then improve on usability problems. Summative evaluations are used to produce measurements to judge systems (Nielsen, 1993; Hix and Hartson, 1993).

A wide variety of evaluation techniques exist in Human Computer Interaction literature but the most commonly used techniques are co-located synchronous methods such as user testing, heuristic evaluation and cognitive walkthroughs. These methods require that a user or expert be present at the same time as the evaluator. This means that the number of evaluations that can be performed is heavily constrained by the size of the evaluation team. The alternatives to synchronous evaluation techniques are asynchronous techniques where users can evaluate a system regardless of the presence of an evaluator.

2.3 An overview of remote usability evaluations

In recent years, there has been a trend towards more remote usability evaluations (Andreasen et al. 2007; Molich et al. 2010). This trend is fuelled by the adoption of AGILE development methodologies (Nielsen and Madsen. 2012), the increasing international focus of systems (West and Lehman, 2006), the outsourcing of system development (Spiliotopoulos, T 2010) and the need to convince stakeholders using quantitative analysis with large sample sizes (Albert, Tullis and Tedesco, 2010).

Empirical studies into remote evaluations date back to the 1990's. Hartson et al. (1996)'s report on two case studies where summative remote evaluations have shown potential for further development. The first case study compared "teleconferencing" to traditional lab tests and the second case study compared "users' self-reported critical incident reports" to traditional lab tests. In reporting their studies, they define remote usability evaluation as "evaluators are separated in space and/or time from user". Despite being early to explore the possibilities of evaluating systems remotely, the authors did not proceed to collect more data to make use of statistical analysis techniques. By only reporting on a small sample of data, they failed to explore one of the greatest strengths of evaluating remotely.

Data collection via the internet is heavily investigated in Psychology literature. Birnbaum, (2004) compares collecting research over the internet against those that are collected in the lab. The author explores the methodological problems of collecting data over the internet such as multiple submissions, dropouts, sampling bias and response bias. In conclusion, Birnbaum stated, "if web studies are properly designed, one can replicate lab results in many fields of psychology".

Birnbaum appeared to be optimistic about the future of psychology experiments performed over the internet. The author also provided recommendations on how researchers can overcome the methodological problems that he discovered. These recommendations were taken into account of when designing the studies for this project.

2.3.1 Synchronous remote evaluations

As remote testing became more developed, there has been more research to compare remote evaluations to traditional lab evaluations. Most notable are Andreasen et al (2007) and Bruun et al (2009) whom both evaluated the Mozilla Thunderbird email client version 1.5. The details of their studies are presented in Table 1.

Table 1 - Studies that compare remote evaluations

Method	Andreasen et al, (2007)	Bruun et al, (2009)
Traditional lab	• Think out loud protocol based on Rubin, (1994) with 6 participants.	• Think out loud protocol following the guidelines of a similar study by Olmsted, E. and Gill, M. (2005) with 10 participants.
Remote Synchronous	• Skype for audio communication. • Virtual Network Computing (VNC) and Microsoft Netmeeting to share desktops and enable webcams • 6 participants were tested.	• Not investigated
Remote Asynchronous	• Experts reported critical incidents that they witnessed with 6 participants.	• User reported incidents with 10 participants.
	• User reported incidents with 6 participants.	• Forum based online reporting with 10 participants.
	• Not investigated	• Diary based reporting with 10 participants.

Andreasen et al. found that the remote synchronous method was “virtually equivalent” to the traditional lab tests. Both papers conclude that although not as many usability problems are discovered in the remote asynchronous method, it still has an “appealing possibility” for usability testing and “may still be worthwhile” because these methods require significantly less time and “enable collection of user data from a large number of participants”.

Both authors appear to be sceptical about the usefulness of the qualitative data collected asynchronously. This is understandable given the lack of consistency and detail in user’s self-reported data.

2.3.2 Asynchronous remote evaluations

Asynchronous remote evaluations are useful when it is challenging to recruit users, for example, those with impairments. Petrie et al. (2006)'s study examines two case studies on asynchronous remote evaluations. One study was formative and involved users who were blind. The other study was summative and involved users with a variety disabilities. They cite that remote evaluations are useful because they allow users to use their own assistive technologies that a usability lab may not have and allow users to use their own configurations that may be challenging to replicate in the lab environment.

They find that the disadvantage of conducting remote evaluations is that it is difficult to understand how assistive technology impacts on the user's interaction with a system. They conclude that the quantitative data collected remotely is comparable to those collected in the lab. However, the qualitative data lacks the detail required for informing design improvements.

West and Lehman (2006) conducted an empirical evaluation by developing and using an automated data collection system for summative usability testing. A total of 30 internal employees were recruited as participants. These participants used the automated data collection system to complete a set of tasks. Metrics such as success rate, time on task and satisfaction rating were collected from the automated system and compared to a separate group of participants who evaluated the system in a lab setting. Key findings from their study include:

- Remote testers took significantly less time to complete the tasks and were significantly more likely to give up on a task, but there was no significant difference in how long both groups would attempt a task. The authors suggest that this may be because participants in the labs may have felt more pressure to perform and, therefore, invested more time and effort into the task than they would otherwise.

- There were no significant differences shown in the amount of written comments given between the group in the lab setting (39%) and the group in the remote setting (41%).

- The comments from the remote tests generally covered the top issues identified by the observer in the lab tests, however, the observer was able to identify 13 more issues that were related to the session as a whole and not specific to any single task.

It is clear from West and Lehman's conclusions that asynchronous remote evaluations are effective for summative evaluations. However, these evaluations lack detailed information that can allow an evaluator to capture general problems that lie within the system as a whole.

The authors provide evidence to suggest that asynchronous evaluations are more suited to evaluating specific parts of a system where less detailed information is required to identify problems. Their research indicates that remote asynchronous testing is more effective when it is conducted in short sessions rather than being extensive like traditional lab tests. However, the authors did not expand on which parts of a system and what they qualify as "less detailed information". Guidance as to which parts of a system is suitable for remote

testing would have proved useful for practitioners so that they can decide when to use remote testing.

A study conducted by Bruun and Stage, (2012) explored how task assignments and instruction types affect the number of usability problems discovered in a remote asynchronous usability test. There were seven conditions in their study:

- A lab test was used as a benchmark.
- There were 6 remote conditions that consisted of tests with predefined tasks, with user-defined tasks, with deductive instructions (the concept of a usability problem was explained), with inductive instructions (examples of usability problems are provided) and with both types of instructions.

The results of their study show that predefined tasks, as opposed to user-defined tasks, led to the discovery of more usability problems. Participants who were given inductive instructions found significantly more usability problems. Bruun and Stage, (2012)'s research highlight the importance of the instructions in a remote evaluation. In the same way, the instructions must be carefully considered when designing a 5 second test.

2.3.3 Remote evaluations in practice

In the UX industry, there has been a steady acceptance of using remote asynchronous testing as shown in Molich et al., (2010)'s CUE study. Molich's study involved 15 UX teams whom independently measured a baseline for the usability of Budget.com. Of the 15 teams, 6 chose to use asynchronous remote evaluations.

Although these teams gained from a "remarkable productivity in terms of user tasks measured with a limited effort", some of their data was unrealistic, as it had suffered contamination. This contamination resulted from undefined or loosely defined criterion for discarding data. Additionally, Molich et al. concludes "the ease of use and intrusiveness of the remote tool influences measurements". This is particularly important in 5 second tests as they are also unmoderated. The ease of use of the 5 second test must be taken into account when choosing or designing the tests.

2.4 The evolution of the 5 second test

The most common method of conducting a 5 second test is online. Providers such as Usability Hub or Clue make the 5 second test easy to set up and free to the user. However, the 5 second test was originated from practitioners who were using traditional usability tests to determine the clarity of content on a website.

2.4.1 The original 5 second test

The original 5 second test was created by Perfetti (2005). She stated that by using this technique her team at User Interface Engineering has "gathered essential [sic] for making huge improvements to our clients' sites". She reports that a 5 second test can provide a "valuable glimpse into what happens during the first moments a user sees a page." In a

Podcast with Perfetti, Jared Spool states that it is "one of the most effective techniques in our toolbox" as it is a very effective technique to understand how a user interprets a web page (Uie.com, 2015).

Essentially the 5 second test measures and assesses the understandability of content. Understandability is an important sub-characteristic of usability (Bevan, 2001) because if a system or website is not understandable then it will risk not being used by users.

It is widely accepted that users skim read web pages to get a high level understanding (Morkes, J. and Nielsen, J. 1997; Spool, J et al 1997; Liu et al, 2010) before they proceed to absorb any of the content. Therefore vital to evaluate and test whether a web page can quickly and effectively convey its purpose to its users.

Perfetti (2005) created the 5 second test to be used at the start of a traditional usability test. The moderator will begin by giving the participant a focused task and then asks them to remember everything that they can when the website is shown to them for 5 seconds. When 5 seconds pass, the moderator asks the participant to write down everything they remember followed by two questions. In Garrett (2011)'s elements of user experience, this would mean that the 5 second test evaluates the surface layer element. It must be noted that Perfetti caveats that the technique is not appropriate for pages that serve multiple purposes, like a home page.

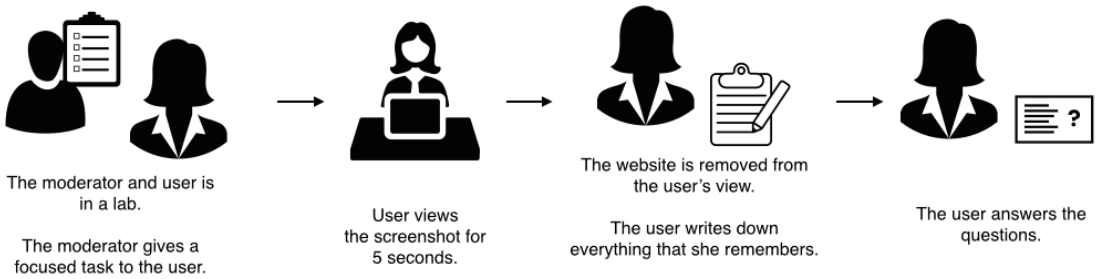


Figure 1 – Procedure for the original 5 second test

In a way, the "5 second test" is similar to the home page tour (Krug, 2006) in that it gives the evaluator an idea of what the user perceives about the content on a website before they use it. The exceptions being that Krug uses the test on homepages and asks participants to say out loud what they think the website is conveying as they are see it.

Thurow and Musica, (2009) combined web search with usability and introduced an 8 second test. This is similar to the 5 second test except that the user begins by searching for a web page through a search engine and the timer begins when the user clicks on the link in the search engine results. The extra 3 seconds is designed to take into account of the page loading time. This 8 second test provides more context to the user which may be beneficial, however, its reliability may be hindered by the page loading time and its accuracy may be biased by the other information on the search results page.

2.4.2 The online 5 second test

In recent years, several online versions of the 5 second test technique have appeared. These online 5 second tests are different to the original 5 second test as the screenshot is shown to the user first (without necessarily including a task) and then asking questions to users afterwards. In addition to these changes, designers have begun using the online 5 second tests to evaluate all kinds of web pages for example homepages and landing pages.

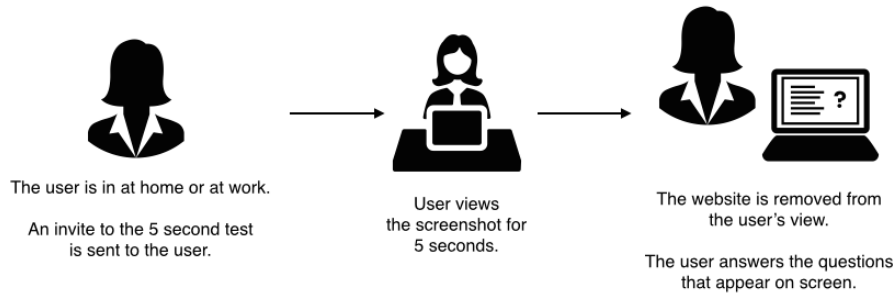


Figure 2 - Procedure for the online 5 second test

The removal of tasks from the 5 second test means that when participants take the 5 second test, they are emulating serendipity browsing (Cove and Walsh, 1988) because they do not know what will appear, and there is no structure to what they are doing. On the other hand, in Perfetti's original 5 second test, participants are performing search browsing where they know what they are looking for as their goal is known (Cove and Walsh, 1988).

By using the test on homepages rather than restricting it to content pages, users have more information to absorb in the same amount of time as homepages typically have a variety of content for example marketing content, information content, entertainment content. They are also designed to convey many different things at once unlike like content pages. These changes raise questions about whether the online 5 second test is still as effective as the original version.

The online versions of the 5 second test are much more accessible as they do not require the evaluator to set up a usability test in a lab. The online versions only require a screenshot and the use of an online service provider that is normally free. Thus, UX practitioners and online marketers have taken this opportunity to use the remote 5 second test to optimise the understandability of their content, optimise conversions, test calls to actions and reduce bounce rates (Useful Usability, 2014; Measuringu.com, 2015). Despite its simplicity and ability to reach participants across the globe, this technique has yet to be used in academic literature.

Table 2 - Overview of existing providers of 5 second tests.

	Name of provider			
	UsabilityHub	Clue	UserZoom	Verify
Ability to adjust	No	No	Yes (paid plans	No

exposure time?			only)	
Ability to show questions before the screenshot?	Questions can be inserted into the introductions page.	No	It is possible to insert questions as “tasks”.	No
Ability to change the response fields for questions?	No	No	No	No
Cost?	Free	Free	Subscription only	30 day free trial
Data analysis provided?	Raw data Tag clouds	A list of words ranked by frequency	Unknown (the author does not have the subscription)	Raw data Demographic data

2.4.3 Research into the online 5 second test

The book “The UX Five-Second Rules: Guidelines for User Experience Design's Simplest Testing Technique” is a first attempt to define and create a methodology for the online 5 second test. This book has been published by Morgan Kaufmann and is available on Elsevier's store and Amazon. The book draws on more than 300 five second tests taken on the Usability Hub platform.

The author of the book, Doncaster (2014), states that the 5 second test “is one of the most convenient rapid UX testing methods available”. He defines the 5 second test as “a type of survey methodology” that unlike most of the other surveys used in UX, can provide insights that shape a design in its early stages.

A breakdown of this book is provided as it is the first piece of research into the 5 second test. Chapter 1 of the book defines the method and how it is being used on Usability Hub. Here, Doncaster discusses how the technique has evolved and introduces some of the cognitive processes involved for participants of the 5 second test. These include concepts such as short term memory and working memory.

Chapter 2 presents the guidelines for creating effective 5 second tests. This includes how to create proper instructions, how to optimise the screenshot, how many questions to use in the test, how to order the questions, the wording of the questions and the types of questions. The main takeaways from his guidelines are:

- To ask about specific things in a question for example avoiding vague phrases such as ″prominent element″.
- To avoid giving away too much context in the questions.

- To order the questions by priority as a user's memory "fades" with each question.

- Take into account that working memory is finite therefore it is better to ask fewer questions (there is a decline in specificity after the initial two responses).

- Eliminate scrolling so that users do not spend time scrolling.

Most of Doncaster's guidelines are supported by evidence, for example, the author conducted tests to prove that there is a point where users will not be able to provide any detail in their answers due to an inability to provide more information from their memory. However, not all of the claims are supported. For example, the guideline regarding scrolling seems to have arisen as a result of the screenshots specific to his sample. During everyday usage, it is normal for users to scroll (Lukew.com, 2015). Therefore it would not be an accurate reflection of user behaviour if they were constrained to viewing only the content that is above the fold. Having said this, it is unclear whether the user was allowed to scroll or not in Perfetti's original 5 second test.

Chapter 3 and Chapter 4 of this book explores how the 5 second test can be used to evaluate attitudinal topics. This includes using the test to evaluate areas such as emotional responses, judgements on trustworthiness and credibility.

These chapters critique the design of many 5 second tests and then provides a template which the author believes will lead to more effective 5 second tests. The template for testing emotional responses requires the user to rate the website from 1 to 5 according to a specific attribute. The user then has to type in a number to respond. The attitudinal template follows a similar format where participants have to respond to a statement by typing in either agree, disagree or no opinion. These templates are useful but it is obvious that Doncaster has been constrained by the Usability Hub platform and its lack of customisation settings for the response fields. It would be more appropriate to ask the user to select a rating rather than type in a number or word.

The addition of these topics expands the uses of the 5 second test. It also makes the 5 second test an appealing option for researchers to conduct tests into how hedonic and emotional aspects of a website can affect users.

The result of this book is a set of guidelines on how to design effective 5 second tests, however, the book fails to investigate the actual effectiveness of the 5 second test itself. Doncaster stated that there is no "clear cut" answer to find out why 5 seconds was chosen (Doncaster, 2014, pp. 3).

An empirical study by Liu et al. (2010) found that most users dwell on web pages for 10-20s before they leave. This coincides with Weinreich et al. (2008)'s empirical study on web usage which found that median time spent on a website is 9.4 seconds. Cyr and Head, (2013) compared how task framing affected user's perceptions under two conditions, 5 seconds and unlimited time. Their findings show that the viewing times did affect how users make judgements about various factors such as involvement and effectiveness. When discussing the limitations to their study, Cyr and Head mention that although 5 seconds appeared to be sufficient to meet the objectives of their study, they state that "other researchers might experiment with different viewing times to determine the optimal time required for users to absorb website elements".

The evidence in the aforementioned studies indicates that 5 seconds may not be enough. They also provide support to show that it is unrealistic to assume that users only spend 5 seconds on a web page before they make a judgement on its content. The implications are that the 5 second test may be ineffective in testing the understandability of content if most users spend upwards of 10 seconds before they leave a website. Doncaster's unconvincing explanation begs the question. Is 5 seconds the optimal amount of viewing time for websites?

2.5 Review of literature concerning first impression judgements on aesthetics

A vast amount of Human Computer Interaction research has been concerned with evaluating the emotional aspects of an interactive system, namely visual appeal.

Doncaster's discussion on using the 5 second test as a technique to evaluate these emotional aspects of interactive systems is similar to how researchers have investigated into first impression judgements on aesthetics. This body of research on website aesthetics mainly focuses on how first impressions are formed and the subsequent effects on the user.

The most influential set of studies was conducted Lindgaard et al. (2006). The participants in their first study rated a set of websites two times after viewing them for 0.5 seconds (500ms). The ratings were found to be consistent which meant that perceptions of visual appeal were reliably be formed in under half of a second ($r = 0.97$).

Their follow-up study allowed participants to view the websites for as long as they liked before they gave it a rating. Afterwards, the participants viewed the websites again but for 0.5 seconds (in a randomised order) and rated them again. They found that the correction between the two sets of ratings was very high ($r = 0.98$). The final study that they conducted used only 50ms of viewing time and once again, the pair of visual appeal ratings was found to be stable ($r = 0.97$).

Since Lindgaard et al. (2006)'s study, a number of studies followed, which provided evidence to show that users form judgments about visual appeal in a very short time. Researchers have since moved onto exploring how judgements affect other factors such as trustworthiness. A summary of some of these studies is provided in Table 3.

Table 3 - A summary of studies investigating first impressions of visual appeal

Publication name	Participants	Exposure times	Main research focus	Main findings
Lindgaard et al (2006)	22, 31, 40	500ms, unlimited time, 50ms	Reliability of judgements on the attractiveness of websites.	Judgements of visual appeal are reliably formed in 50ms.
Van Schaik and Ling, (2009)	50, 115	500ms, unlimited time	Context (mode of use e.g. goal orientated or action orientated) and judgements.	Context increases the reliability of judgements on aesthetics.

Cyr and Head, (2013)	60	5000ms, unlimited time	Task framing (either for a hedonic or utilitarian interest) and viewing times on user perceptions.	Task framing does not affect perceptions but unconstrained viewing times do. For example, higher viewing times lead to higher perceived trust and enjoyment.
Reinecke and Gajos, (2014)	32,222	500ms	Visual appeal judgements and demographics in relation to colourfulness and complexity.	Judgements on visual appeal vary depending on the demographics of the user.

Many studies investigated the factors of a website lead to emotional judgements and the minimum amount of time required for these judgements. While 50ms may be the minimum amount of time required to rate aesthetics, there is no evidence for why 5000ms should be used for any other judgements.

Cyr and Head (2013) chose to use 5000 ms (5 seconds) and unlimited time for comparison as it is the average time used in previous studies. These are 3s in Lindgaard et al., (2006); 4s in Kaiser, (2001); 5s in Perfetti, (2005) and 7s in Ramsey, (2004) in human-to-human interaction. However, their findings find that participants with unlimited viewing time are more satisfied and exhibit "higher perceived involvement, enjoyment, trust and effectiveness".

Within the Human Comptuter Interaction literature, there are very few studies that examine how these judgements change over time as users absorb more content. This is an important topic because it is inevitable that as a user views more content on a website, it affect their judgments about its visual appeal the website. It is expected that the flexibility of the platform created in this study will enable the other researchers and practitioners to explore further this topic.

Since Lindgaard et al. (2006)'s study, a number of similar studies followed. These studies provided evidence to show that users form judgments about visual appeal in a very short time. Researchers have since moved onto exploring how judgements affect other factors such as trustworthiness.

2.6 A model for information processing

Most of the previous research on visual appeal can be described in terms of the information processing stage model devised Leder et al. (2004). In this research, the authors explain, from a psychological perspective, why people are attracted to art and then propose a model that represents the different stages of visual processing.

In their study, the authors' highlight 5 main stages that contribute (see Figure 3) to the evaluation of art which they state can be generalised to the processing of other aesthetic

stimuli. The perceptual analysis and implicit memory integration are the basic subconscious cognitive processes that happen spontaneously. The other stages are higher-level processes that are different for each individual and are influenced by their knowledge and experiences. Although the model does not explicitly state that the flow of information in each stage is sequential, the fact that the later stages are higher-level processes suggests that the processing of visual stimuli is time sensitive.

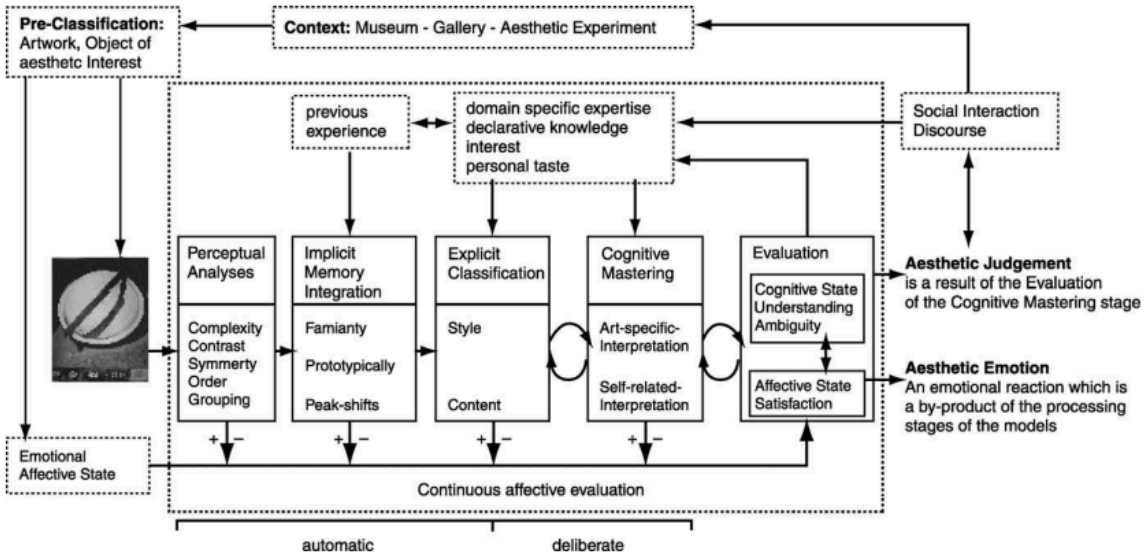


Figure 3 - Information processing stage model devised Leder et al. (2004)

Previous Human Computer Interaction research on aesthetic appraisal has concerned itself with investigating the lower levels of processing, particularly those in the perceptual analyses stage. Their focus is to determine the minimal time that is required for a user to use their perceptual sense to form a first impression. This means that there is a lack of research into how aesthetic judgements change beyond these early stages.

3. Creating the platform to be used for the studies

The first step of this project required building an experimental platform to enable the researcher to adjust different variables within the test. Following the completion of the platform, two experimental studies were conducted. The source code for the testing platform is host on Github, the web address for the repository is shown in Appendix B.

3.1 Requirements of the platform

The platform was inspired by the existing 5 second test provided by UsabilityHub.com. The platform's requirements are based on what is currently available and what is required to explore the research questions.

The functional requirements describe the fundamental properties that the platform must have to allow the researcher to create tests to collect data. The non-functional requirements are necessary to ensure that participants invest maximum effort into taking the test and so that a multitude of participants will take part in the studies.

Table 4 - Requirements for the platform

Functional requirements	**Non-functional requirements**
The platform shall enable the researcher to manipulate the exposure time for an individual screenshot.	The platform shall look aesthetically pleasing.
The platform shall enable the researcher to display a screenshot for the predetermined exposure time.	The platform shall be available for 24 hours in a day during the time that the tests are in operation.
The platform shall enable the researcher to customise the instructions.	
The platform shall enable the researcher to capture participant responses.	
The platform shall enable the researcher to export the data collected onto a spread sheet.	
The platform shall enable the researcher to disable mobile or tablet access to the studies.	
The platform shall enable the researcher to ask participants more than 1 question per individual test.	

3.1.2 Technology

The platform was built using technology that can be accessed over the internet, these were:

- User Interface - The Bootstrap framework (Html5 and CSS3) was used to format the pages and forms on the platform. Javascript and Jquery were used to

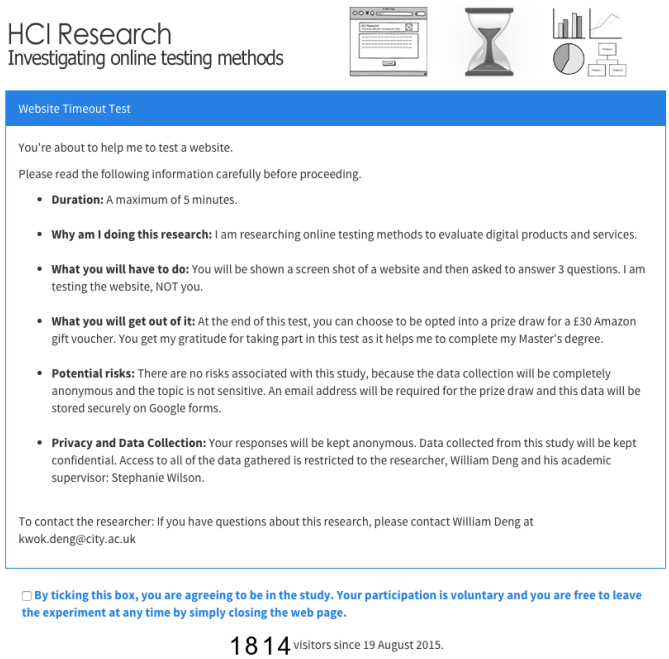
manipulate how long each screenshot is visible for and to redirect participants on small screens (mobile/tablet) to a page that does not allow them to take part.

- Data management - PHP 5.5 was used to handle data from the browser to the SQL database.

3.2 Components of the platform

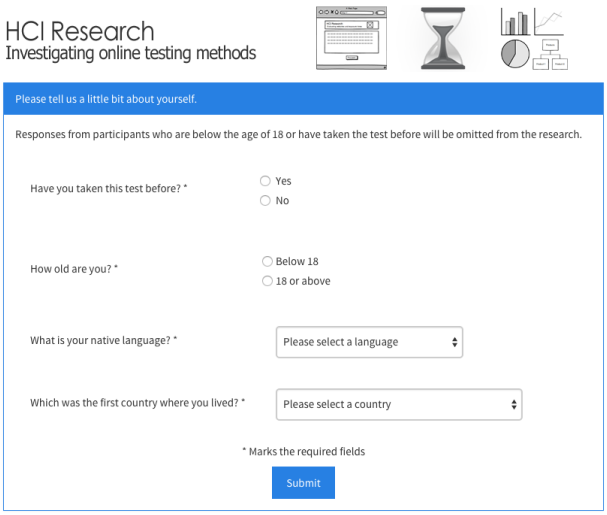
The platform was hosted on a standard Apache server and was accessible publically. There were six unique component pages. Each page required the use of different technologies and these are shown in Table 5.

Table 5 - Components of the platform

1. Upon clicking a link to the study, the first page a participant will see is the introduction to the study. This page informs the participant about what the test is about and what is required of them. The pages are formatted using the Bootstrap Cosmo theme. This is available from Bootswatch.com and distributed for free on the MIT License. Javascript is used to reveal the "Continue" button after the checkbox has been ticked. The visitor tracker was created by webestools.com which is free to use.

2. A page to capture information about the participant. This is form uses HTML5 for validation. Participants can only continue if all of the response fields are filled.

PHP5.5 is used to process this form. The input is saved in an SQL database.

It is on this page that participants are assigned to the various different groups for the studies. A simple PHP function is used to randomly distribute participants depending on the amount of responses for each group.



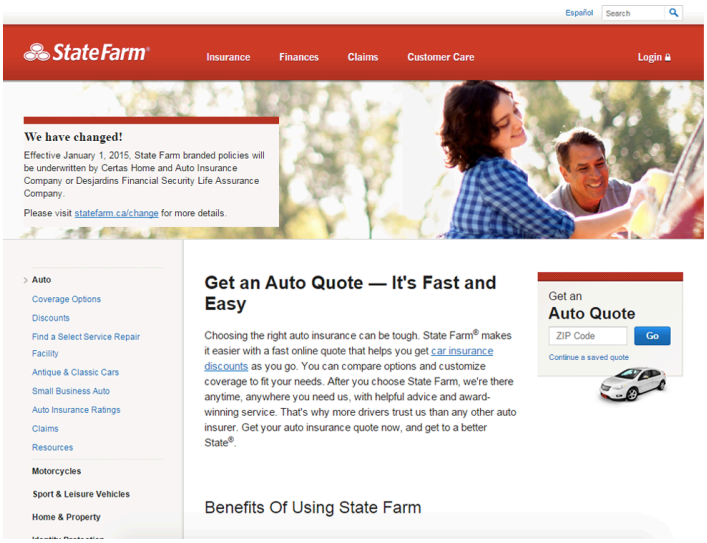
3. A page to display the instructional video. This video was created and hosted on moovly.com.

This video explains the procedure for the test. For the second study where the participant sees the questions before they see the screenshot, the instructional video was modified to reflect this.

4. A page to display the screenshot for a limited amount of time.

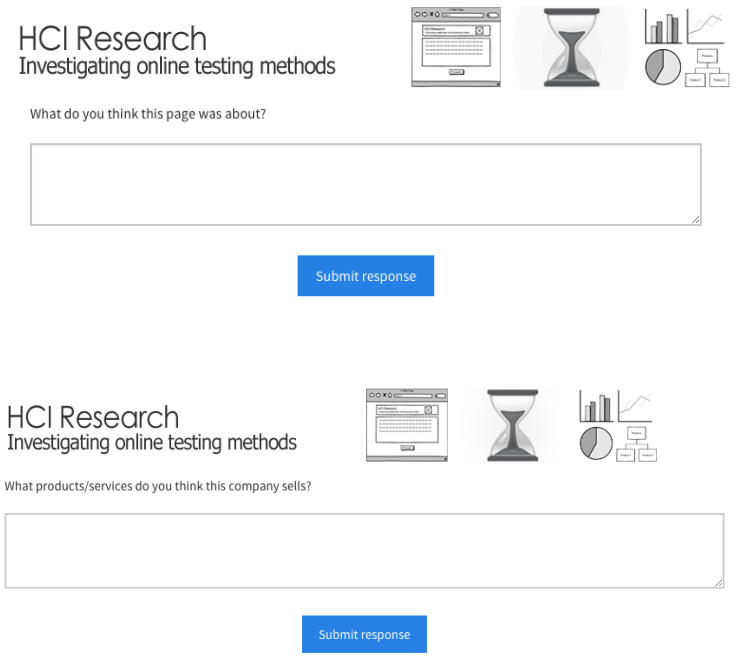
To show the screenshot of the website for a limited amount of time, the fadeOut() function in Jquery was implemented. This function hides the screenshot after a set amount of time.

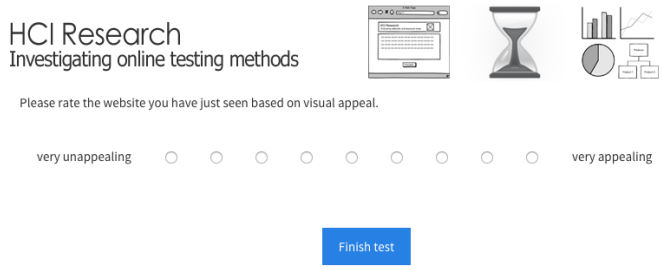
Once the time is up, the fadeIn(0) function will show the first question. These questions are processed in PHP and upon submitting, the user will go to the next page of questions.



5. Three web pages to capture participant responses.

These pages contain HTML forms are processed using PHP5.5 and stores information in the same SQL database.



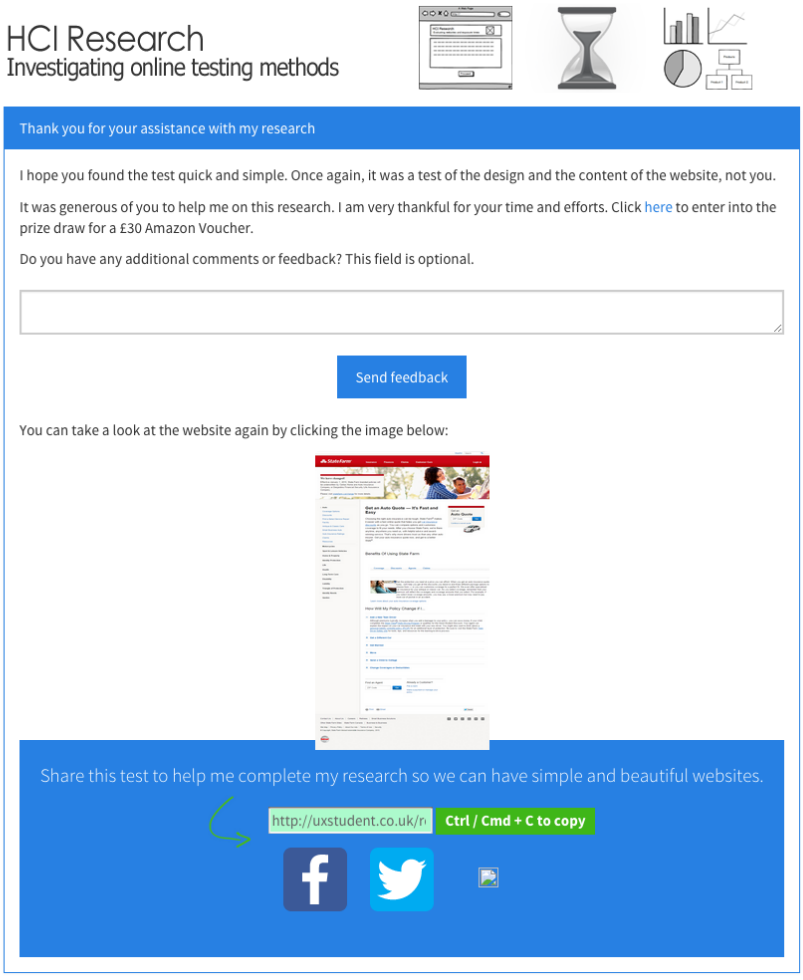


6. A completion page which allows participants to give feedback, see the screenshot again, enter a prize draw (incentive) and share the study.

Upon submitting a response to question 3, the user will arrive at this page where they may fill in a form to a prize draw, submit feedback about the study, view the screenshot and use social media to share the study.

The prize draw form is hosted on Google Forms. The feedback form is processed with PHP and stored on the SQL database.

Clicking on the screenshot will bring the user to a page with the screenshot.



Specific details about each page are provided in the procedure sections for each study (see 4.1.4 and 5.1.2).

3.3 Technical details

The screenshot for these studies was designed to be viewed on laptops and computers. Therefore, if a person were to take part in the study using a mobile phone or tablet, the data would be distorted and therefore be biased. To prevent this, a simple HTML meta refresh was implemented (this code is available in the GitHub repository in Appendix B, the code is located on the index.php page). This meta refresh code created an instant client-side redirect that led mobile and tablet users to a page that advised them to return to the experiment when they were on a laptop.

Figure 4 shows the structure of the database used to store the data collected. A simple structure was for which one table was used to store user information, one for feedback and one table for each question in each group for every study. The table structures for each question in each group remained the same as the one shown in Figure 4.

As there were two studies with four groups, of which each had three questions, this meant that a total of 12 separate tables were used to store data. These tables share the exact same structure as Group1_q1 and Group2_q1 except that they were named according to which group and which question that it stored data for.

Table name: User_Info

#	Name	Type	Collation	Attributes	Null	Default	Extra
1	**userid**	int(11)			No	*None*	AUTO_INCREMENT
2	**taken**	varchar(355)	latin1_spanish_ci		No	*None*	
3	**age**	varchar(355)	latin1_spanish_ci		No	*None*	
4	**language**	varchar(500)	latin1_spanish_ci		No	*None*	
5	**country**	varchar(500)	latin1_spanish_ci		No	*None*	
6	**unique_id**	int(55)			No	*None*	

Table name: User_Feedback

#	Name	Type	Collation	Attributes	Null	Default	Extra
1	**userid**	int			No	*None*	
2	**feedback**	longtext	latin1_swedish_ci		No	*None*	

Table name: Group1_q1

#	Name	Type	Collation	Attributes	Null	Default	Extra
1	**userid**	int(11)			No	*None*	
2	**que1**	longtext	latin1_spanish_ci		No	*None*	

Table name: Group2_q1

#	Name	Type	Collation	Attributes	Null	Default	Extra
1	**userid**	int(11)			No	*None*	
2	**que1**	longtext	latin1_spanish_ci		No	*None*	

Figure 4 – Structure of the tables in the database

3.4 Designing the user interface

Molich et al. (2010) stated that the ease of use of the remote tool and the clarity of its instructions has a considerable impact on the performance of the unmoderated participant. Therefore, the user interface of the platform was designed to be straightforward and easy to use so that participants are motivated to complete the test with maximum effort. This was achieved by:

- Ensuring that textual information on the site is written in plain English and is in a legible font size.
- Presenting the instructions to the test in a clear and an engaging way via the use of animated videos. These videos were less than one minute in length and explained what will happen in the test and what was required of the participant. This ensured that instructions are concise which has been shown to be vital in remote unmoderated tests.
- Implementing a modern look and feel by employing a flat design style.

3.5 Pilot testing the platform

Pilot tests were conducted to test the platform for stability and to examine any usability problems. These tests were also intended to verify and validate whether the requirements were met. Hyperlinks to the platform were sent to eight participants where they would take part in the tests. In each experimental group, there were two pilot test participants. These participants were not re-invited to take part in the study.

After each participant had reported that they had finished the study, they were interviewed over the phone or over the internet to uncover whether they experienced any technical or usability problems. The participants in the group with 0.5 seconds exposure time reported that they found it “too quick” and thought that they had an error. To ensure that participant feedback is captured, an open-ended feedback form was entered into the completion page. One participant also found that the link to the website screenshot on the results page was broken. This issue was fixed.

Minor usability improvements were also made such as changing the wording of the header texts as a few participants found that there was too much text in the information and completion pages. Misspellings were also corrected.

The researcher confirmed the responses captured by the platform with the participant to see if the responses were the same. All of the responses were found to be accurately captured.

4. Study 1: Investigating exposure times in the 5 second test

This study investigated how changing the exposure time in the 5 second test affected participants' responses and their ratings of visual appeal.

The hypotheses were:

1. Increased exposure time leads to responses with higher correctness scores.
2. Increased exposure time leads to responses with higher word count.
3. There will be a difference in visual appeal ratings when exposure times are increased.

The method used to define that correctness scores is presented in 4.1.2.

4.1 Methodology for study 1

Three different tests were created using the platform outlined in Chapter 3. Participants were invited to take part in these tests. The outcomes of the tests were the participant's responses to the questions in the test. Both quantitative and qualitative data were collected. Statistical tests were performed to test the hypotheses.

4.1.1 Variables

The testing platform randomly divided participants into three separate groups. By randomly assigning participants, systematic variances were avoided. Each participant only took part in the study once.

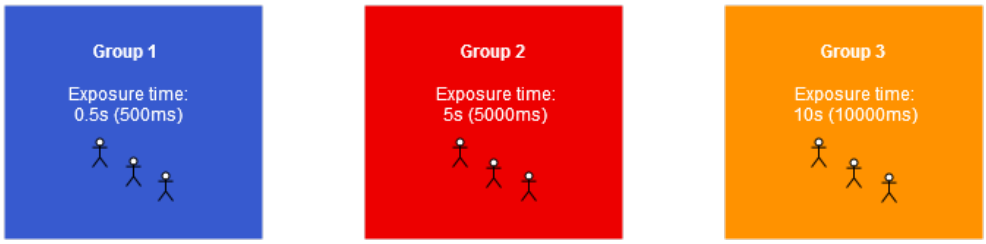


Figure 5 – The groups of participants in study 1.

The responses between participants in each group were compared therefore making this a between subjects study.

The independent variable was:

- Exposure time.

These times were chosen because:

Table 6 - Justification for the choice of exposure times

Time	Justification
0.5 seconds (500ms)	This is the time used by the majority of studies on first impressions as it is the amount of time that a person needs to be able to judge reliably the visual appeal of a website. Therefore allows for comparison between the studies. In addition to this, this is less time than what is used in a 5 second test which enables an exploration in what happens at such a low amount of exposure time.
5 seconds (5000ms)	This is exposure time is used as a benchmark to the standard 5 second test.
10 seconds (10000ms)	This is a conservative estimate on the average dwell time as reported in both Liu et al. (2010) and Weinreich et al. (2010)'s studies.

To answer research question 1, about the effect on participant responses, each response was assigned a score for its correctness using the scoring system in 4.1.4.

Participants were asked to answer three questions in this study, these were:

- Question 1. What do you think this website is about?
- Question 2. What products/services do you think this company sells?
- Question 3. Please rate the website you have just seen based on visual appeal.

Question 1 and question 2 in the test evaluates the participant's understanding of the website's content after viewing it for a limited amount of time. The visual appeal rating of the website captures participant perceptions about the website's aesthetics.

The participants replied to the first and second questions in text input boxes where they were free to write as many or as little words as they wished. Responses had to be at least one word.

For the third question, participants input their answers by selecting one of 9 radio buttons. This Likert rating scale labels the selections on the far left as very unappealing and the selections on the far right as very appealing.

The dependent variables were:

- The correctness and word count of each response for question 1 and question 2.
- The visual appeal ratings for question 3.

4.1.2 Choice of stimuli

A screenshot of the State Farm Auto Insurance page was used as the screenshot. State Farm is an American insurance company that also operates in Canada. According to their website (www.statefarm.com) they do not operate in any other country.

This page can be described as a landing page as it contains a the local navigation to access other content within the auto insurance category. This was accessed by going to the State Farm homepage and then selecting "Auto" under "Insurance" in the main navigation.

The screenshot was captured on a Chrome browser on a monitor with 2304x1440 pixels (px) on the 29th of May 2015. The display size of the browser window was resized to 1280x768 pixels to capture the screenshot.

This size was chosen because it can be comfortably viewed by the majority of internet participants (as of May 2015, according to Gs.statcounter.com, (2015), 29.88% of users had 1366x768px, 11.98% of users had 1920x1090px and 8.55% had 1024x768px). Any uncovered screen area was given a white background.

State Farm has a low exposure to people in the countries outside of America and Canada. This is where most of the participants came from (see 4.3.1).

The State Farm website was chosen for the following reasons:

- It was designed for Western cultures, which is where participants for the studies came from.

- This website was chosen as Statefarm does not operate outside of America and Canada and therefore it avoids any sort of marketing interactions such that their company's branding may affect a participant's perceptions of it.

The topic of car insurance was chosen because:

- It is a concept that is familiar with most of the people in Western cultures.

- It is a gender-neutral topic therefore allowing the researcher to test with both males and females.

An informal survey was conducted with 20 randomly selected students from the City University campus library. None of these students had heard of the State Farm brand before. This verified that most UK residents do not know of the State Farm brand.

4.1.3 Description of the stimuli

The selected screenshot can be described as both a homepage for the Auto insurance subcategory and a content page for the company's auto insurance products. The global navigation is on the top of each page and displays links to the company's products. The local navigation is on the left and shows links to sub-categories within the "Auto Insurance" category.

The screenshot used in the study is shown in Figure 6. The black box shows what appears above the fold for a laptop/PC user with a screen size of 1290 x 768 pixels.

Areas that highlighted in yellow represent informational content and calls to action. Areas that highlighted in green represent navigation elements.

Figure 6 – Screenshot used as the stimuli for the studies

The main body of the page consists of 6 sections. The header of the body is bold and catchy and reads "Get an Auto Quote - It's Fast and Easy".

On the top right of the page, participants can fill in their zip code and to get a quote. This is the call to action on the page.

Beneath this content is a section which provides information about the benefits of buying an insurance policy with State Farm. After this section, the page provides details about the insurance policies.

At the bottom of the page, there is a form to find an agent by entering in a zip code and two links that provide access to current customers. The word count including the words in the navigation areas totalled to 593 words.

4.1.4 Producing a scoring system to measure correctness

To gauge the correctness of participant responses, it was required to determine what an ideal response would be for Question 1 and Question 2. Six experts with experience in Human Computer Interaction analysed the stimuli.

These experts had not previously taken part in any of the studies, heard of the State Farm brand nor have they seen the State Farm website before. Each expert had at least a year's experience in the study of Human Computer Interaction.

Table 7 – Details of the experts employed for analysis

Expert	Years of experience	Time spent (Rounded to the nearest minute)
A	1	3
B	1	7
C	1	5
D	2	10
E	10	7
F	5	6

Experts conducted the analyses on the stimuli in a co-located setting with the researcher. The experts were briefed beforehand (see Appendix C) and then asked to view the State Farm screenshot with unlimited time.

When they were finished viewing the screenshot, the informed the researcher that they were ready for the questions. The screenshot was removed from view and then the questions were presented to them. They typed their answers onto a Word Document. These answers were subsequently decomposed into individual clauses.

These sessions were timed, and the expert answers were recorded in Appendix D.

4.1.5 Analysing the expert answers

The experts provided a total of 23 clauses to question 1 and 14 clauses to question 2. Thematic analysis was conducted where each clause was evaluated and grouped together with clauses with a similar theme. Clauses needed to contain at least one noun or adjective that conveyed the same meaning to be grouped together.

Each category label was repeatedly updated as newer clauses were analysed. This iteration was performed until natural and appropriate categories emerged for each of the two questions.

A list of synonyms was used and these words were treated equally throughout the analysis (see Appendix E). This resulted in 8 categories of responses for question 1 and 6 categories for question 2.

Table 8 - Categories for each question

Question 1	
Category	**Definition**
Q	Related to getting a quote
I	Related to information gathering for car insurance
C	Related to car insurance and synonyms for car i.e. auto, motor, vehicle
An	Related to answering frequently asked questions (FAQs)
Ins	Related to insurance
Fin	Related to finance company

Question 2	
Category	**Definition**
I	Related to information gathering for car insurance
C	Related to car insurance and synonyms for car i.e. auto, motor, vehicle
T	Related to home, health, family insurance
Ins	Related to insurance
Fin	Related to finance company

The amount of clauses in each category was counted. Where a category contained more than 0% but less than 10% of the total clauses for that question, it was decided that responses that contained themes in this category should have a score of 1. For a category that contained between 10 and 19% of the total clauses, it was given a score of 2. This criteria was applied in the same way to arrive at the scoring system for each question.

Table 9 – Scoring system for each question

Question 1		
Category	Number of clauses (percentage)	Score
Q	5 (20%)	3
I	5 (20%)	3
C	6 (24%)	3
An	2 (8%)	1
Ins	1 (20%)	1
B	5 (4%)	3
Fin	1 (4%)	1
NA	-	0

Question 2		
Category	Number of clauses (percentage)	Score
I	1 (7.14%)	1
C	7 (50%)	5
T	4 (28.57%)	3
Ins	1 (7.14%)	1
Fin	1 (7.14%)	1
NA	-	0

Total	25 (100%)	-

Total	14 (100%)	-

The correctness of the participant responses was measured by using the scoring system. These scores for each of the participant responses can be seen in Appendix F.

4.1.6 Procedure for the study

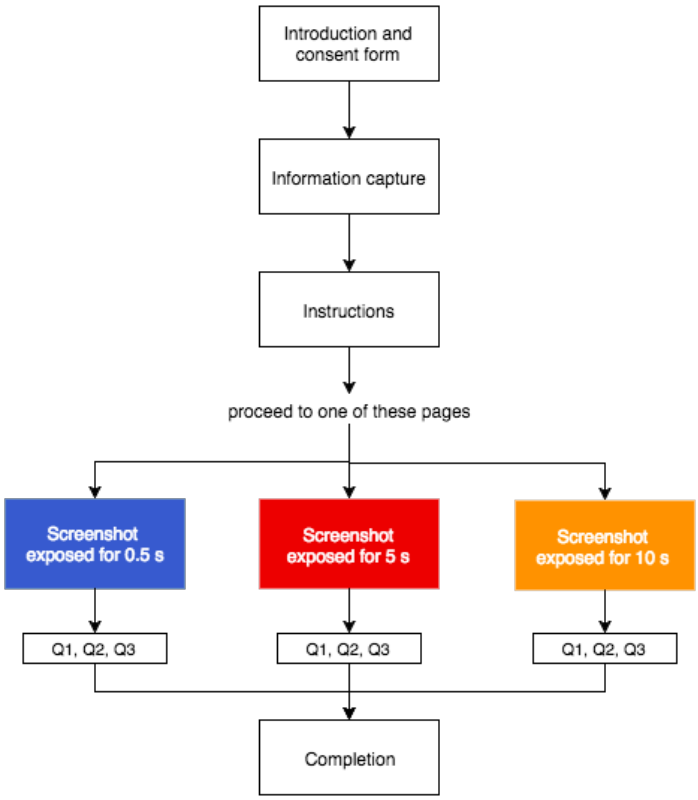


Figure 7 - Procedure for study 1

The procedure for this study is outlined below:

1. Upon clicking the link to access the experiments, participants begin on the information page. This page informs participants about the study by providing information about possible risks, the treatment of data and how to contact the researcher. Participants can only continue after they have ticked the checkbox to agree that they understand and give consent to take part.

2. The next page captures information about the participant, such as location and age. This enabled the researcher to filter out which participant responses.

3. Next, participants see an Instructions page. This page contains an animated video that explains to participants about what will happen in the study and what the participant must do.

4. After viewing the instructions, participants continue to the next page which contains a button. Upon clicking this button, the screenshot is displayed for 0.5, 5

or 10 seconds.

5. As soon as the screenshot disappears, the first question appears. On these pages, three questions followed by the appropriate response fields are shown, being either a text input box or a set of multiple-choice options.

6. The last page is the completion page. As soon as participants have completed the questions, they will arrive at this page where they can submit additional feedback, enter the prize draw or share the study to their friends.

4.1.7 Participant recruitment process

Recruitment for both studies began in August 2015 and lasted 7 days. The study was advertised on online communities and social networks such as thestudentroom, Mumsnet and Linkedin. A full list of advertisement locations is provided in G. Participants were incentivised by a prize draw to win a £30 Amazon Gift Voucher.

4.2 Data preparation

In this section, the data preparation process is presented. This includes a discussion about the datasets that was collected, which datasets were discarded and which datasets were included in the analyses.

4.2.1 Procedure for data analysis

The data collected from this study was exported out of a SQL database and imported into several Google Sheets. The output of the statistical tests is presented in Appendix H.

The amount of words in each participant's response to question 1 and question 2 were individually counted using the =countA(split()) function in Google Sheets. A sense check was performed on 10% of the randomly selected records to ensure that there were no errors.

Descriptive statistics and charts were generated in Google Sheets. The analysed data is presented in Appendix I. Inferential statistics were applied using SPSS.

The scoring system devised was applied onto each participant's responses. Question 1 and question 2 were scored individually. Longer responses may fit into more than one category code in the scoring system and therefore could score as many points as it warranted.

No responses belonged to more than two categories and the score was 6 in question 1 and 8 in question 2. Responses that contained inappropriate language were censored.

4.2.2 Removing data from unsuitable participants

This section provides a description of how the data that from the SQL database was filtered and organised.

The original dataset contained 880 participants. These are the people who filled in the information capture form (page 2 in Table 5). Of these, 54 participants reported that they had either taken the test before or were under the age of 18 and were therefore omitted.

Subsequently, 826 participants remained. These participants came from 65 different countries (see Appendix J).

The nature of the stimuli required that participants were familiar with the concept of car insurance and were not disadvantaged by any language barriers. This resulted in the omission of participants who were not from a First World Country or did not report that their first language was English were omitted.

From the list of 65 countries, only six countries (New Zealand, Ireland, United Kingdom, Australia, United States of America and Canada) qualified for this criterion. To avoid marketing interactions (see 4.1.2) participants from the United States of America and Canda were omitted.

This means that there were 460 participants remaining. By matching the user ids for these users, it was found that there were only 346 fully complete tests. This represents a 24.6% drop out rate.

The data for study 1 and study 2 are drawn from this pool of data. This data is presented in Table 7.

Table 10 – Qualifying countries and participants

Country	**Participants**
New Zealand	5
Ireland	40
United Kingdom	365
Australia	50
Total	460

Group – (experimental variable)	**Responses**
Group 1 - (0.5s)	70
Group 2 - (5s)	90
Group 3 - (10s)	88
Group 4 – Saw the questions before the screenshot	98
Total	346

The overall dropout rate (people who filled in the information capture form but did not complete the whole test) was 21.42% because there were 649 fully completed responses (a breakdown is provided in Appendix K.

Given that participants were randomly allocated to the four different groups, it is expected that the amount of tests taken will be approximately equal. However, the group where the

exposure time was 0.5 seconds has the lowest amount of complete responses (it, therefore, had the highest dropout rate). This is expected given that participants in this group are the least likely to be able to answer the questions. This is supported by the feedback that participants provided (see section 6.4.).

After taking a 0.5 second test, one participant commented on LinkedIn that they had used the back button to view the screenshot again. This was not anticipated, and as soon as it was found that this was possible, the back button was disabled to avoid any further abuses. This happened mid-way through this study.

The statistics from Google Analytics were examined to determine how many participants used this method to see the screenshot more than once. Google analytics shows that the overall bounce rate for the website was 61.14%, and there were over 500 sessions at the time of the incident. The user flow shows that only approximately 4 participants had seen the screenshot page more than once. This provided some confidence that the data had not been contaminated.

4.3 Findings from study 1

This section presents the findings from the analysis of the data gathered from the first study. Where statistical tests were used, the 95% confidence interval is used to determine the outcome of the tests, which mean that null hypothesis is only rejected when $p < 0.05$.

4.3.1 Participants in this study

This study investigated the data in group 1, 2, 3 where the exposure times were 0.5, 5, 10 seconds respectively. The majority of the participants in each group came from the United Kingdom. The demographics for each group are shown in Figure 8.

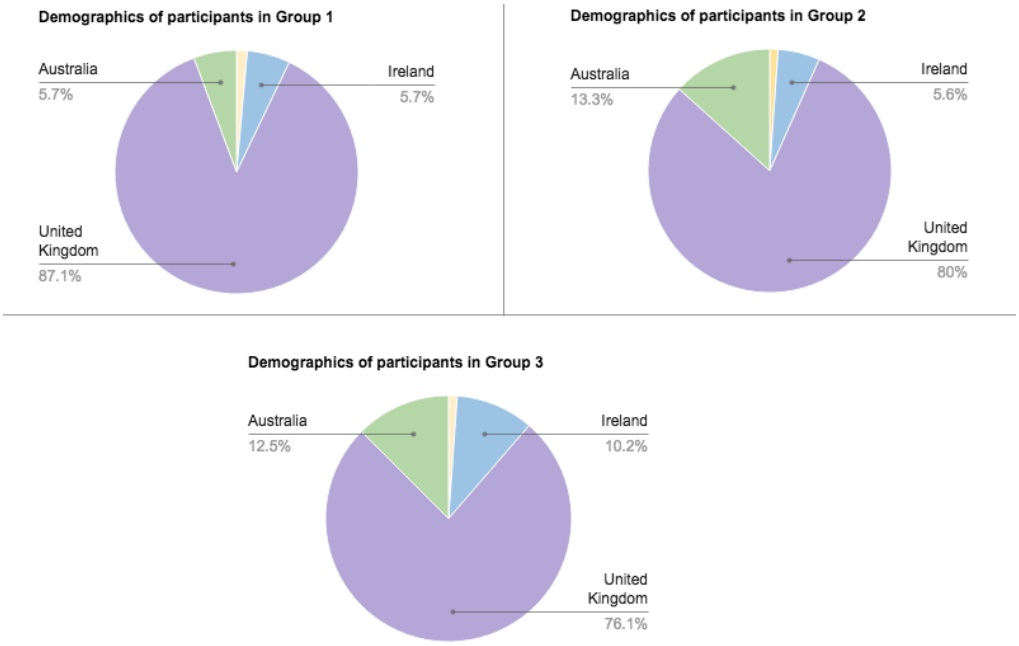


Figure 8 – Demographics for the participants in study 1

4.3.2 Exposure time and correctness scores

By applying the scoring system to each response, the qualitative data was transformed into quantitative measurements. These scores described how correct the participant responses were in relation to the expert responses. A response with a higher score was deemed to be more correct.

Table 11 – Descriptive statistics to show exposure time and correctness scores

	Group 1 (n = 70)	Group 2 (n = 90)	Group 3 (n = 88)
Mean score for Question 1	0.71	1.52	2.02
Mean score for Question 2	0.97	1.87	3.03

The mean for each question within these three groups show that as the amount of time increases, participants will score higher. Participants appear to score higher in Question 2 than in Question 1.

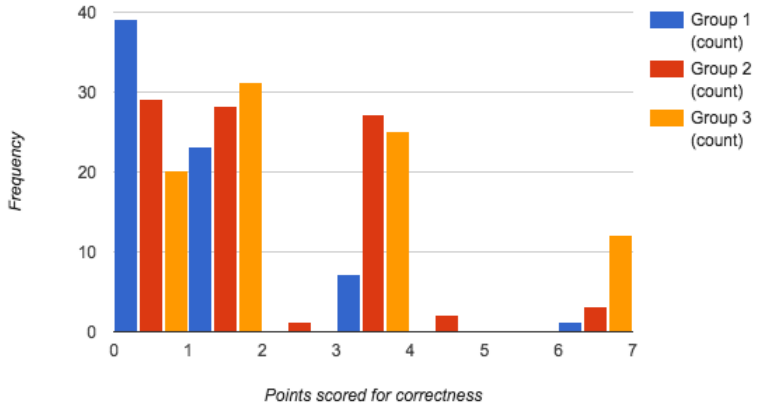


Figure 9 – Histogram to show frequency and correctness scores for question 1

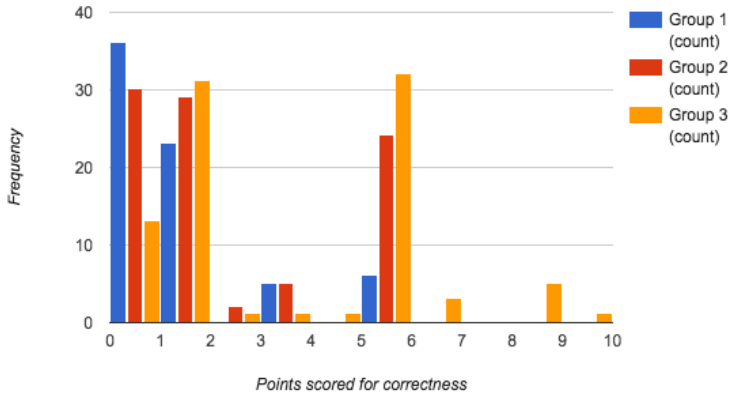


Figure 10 - Histogram to show frequency and correctness scores for question 2

The histograms show that the number of participants with the highest scores was achieved by those with the most exposure time (group 3). On the other hand, participants with the least time (group 1) received the highest amount of 0 scores.

4.3.3 Statistical analysis on the correctness scores

A One-Way ANOVA was conducted to determine whether the differences between the means of the groups were significant. The variances for both questions did not meet the assumption of homogeneity of variances (Levene Statistic = 18.335 and 38.307 for question 1 and 2 respectively, both have $p < 0.01$). Therefore the Welch's correction was applied.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 < \mu 2 < \mu 3$ (where μ is the mean and the number represents the group)

The test results show that the differences between the groups was significant ($F(2, 161.83 = 16.364)$, $p < 0$) for both question 1 and question 2($F(2, 162.023 = 20.831)$, $p < 0$). This suggests that as exposure time to the website increases, responses are more correct. To discover which group differences were significant, the Games-Howell post hoc tests was performed.

The results for question 1 show that the mean difference when increasing exposure time from 0.5 seconds to 5 seconds is significant ($p < 0.05$) but increasing the exposure time from 5 seconds to 10 seconds is not ($p = 0.139$). Therefore, the null hypothesis can be rejected for the differences between group 1 and group 2, but accepted for group 2 and group 3.

The Games-Howell post hoc tests show that, the differences between all of the groups are significant ($p<0.0$) for question 2.

4.3.4 Qualitative findings

The most popular provider of 5 second tests, UsabilityHub, uses word clouds as their default analysis technique. Word clouds, otherwise known as tag clouds, are a visual representation of text data which are used to display the frequency of keywords.

Figure 11 shows the responses for each question in each group after they transformed into tag clouds using Tagcrowd.com. The number in brackets are the frequencies which represent how many times that word appeared in the responses.

It should be noted that TagCrowd uses language-specific lists of common words to keep word clouds relevant, therefore, words such as "it" or "the" are removed.

Figure 11 – Tag clouds to show the key words to the questions in study 1

The tag clouds support the findings from the quantitative analysis as we can see that participants offered a much larger variety of answers to question 1 regardless of exposure time. For both questions, the number of unique keywords in the responses is the when participants were given 10 seconds of viewing time. Additionally, this group also contains the highest frequencies for "car" and "insurance". This shows that responses in group 3 are more reliable than those in the two other groups.

4.3.5 Exposure time and word count

The length of participant responses was analysed because longer responses could potentially provide an evaluator with more informative insights. The length of responses was measured by the amount of words in the response.

Table 12 – Descriptive statistics to show exposure time and word count for question 1

	Question 1		
	Group 1 (n = 70)	Group 2 (n = 90)	Group 3 (n = 88)
Mean words	6	4.41	3.34
Median words	2	2	2
Mode words	1	1	2

Table 13 - Descriptive statistics to show exposure time and word count for question 2

	Question 2		
	Group 1 (n = 70)	Group 2 (n = 90)	Group 3 (n = 88)
Mean words	3.11	2	2
Median words	4.51	2	2
Mode words	3.39	2	2

The word count for each question showed that, in general, the length of responses was similar regardless of the amount of viewing time. In each group, the median and mode is in the range of one to two words with the most common responses being either "insurance" or "car insurance".

The highest word counts have been observed in group 1, where participants had the least exposure time. This is contrary to the hypotheses that participants will write more when they are given more time to absorb the content.

The responses from participants in group 1 contained many guesses to the questions and explanations why they did not know the answers. This contrasted to the responses in group 3 where participants had more time to absorb the content and, therefore, provided concise answers.

It must be noted that there was a rather unusual participant (user id 605) who wrote in full sentences, described the sequence of elements that he/she saw and provided his/her justifications for his/her answer. This participant's word counts for question 1, and question 2 was 63 and 60 respectively. This response contributed significantly to the higher average amount of words in group 2. These responses were considered as outliers as the majority of word counts amongst all three groups is below 9. Therefore, this result was omitted from the statistical analysis for word counts.

4.3.6 Statistical analysis

A One-way ANOVA test was conducted to examine whether the differences in the word counts for the groups were significant. The word counts for question 1 failed to meet the assumption of homogeneity of variances (Levene statistic being $F = 8.295$, $p < 0.01$), therefore the Welch's correction was applied.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 < \mu 2 < \mu 3$ (where μ is the mean and the number represents the group)

The results show that the differences between the three groups was insignificant ($F(2, 127.117 = 2{,}413)$, $p = 0.094$) therefore the null hypothesis is accepted.

Since the Levene Test is not significant for question 2($F = 2.671$, $p = 0.071$), we assume that the variances are approximately equal, thus a one-way ANOVA (without Welch's correction) was performed. The results show that the differences in the length of responses for question 2 were also insignificant and once again the null hypothesis is accepted.

4.3.7 Exposure time and visual appeal ratings

Table 14 - Descriptive statistics to show exposure time visual appeal ratings

	Group 1 (n = 70)	Group 2 (n = 90)	Group 3 (n = 88)
Mean rating	4.43	4.62	4.75
Median rating	5	5	5
Mode rating	5	5	4

The visual appeal ratings appear to be fairly similar between each group and relatively close to the midpoint on the rating scale. This is consistent to the range of the paired ratings of Reinecke and Gajos, (2014) for respondents in the United Kingdom.

The data shows that, although mean visual appeal ratings increase slightly as a exposure times increased, the median rating is still 5 in all of the groups. The most popular rating for the group with the most time to view the screenshot is lower than the other two groups. The standard deviation for Group 1, 2 and 3 are 1.53, 1.61 and 1.95 respectively, which shows that participants with more viewing time seem to vary more in the way that they rate visual appeal. The histogram and boxplot in Figure 12 and Figure 13 also illustrates how the range of ratings differs in each group.

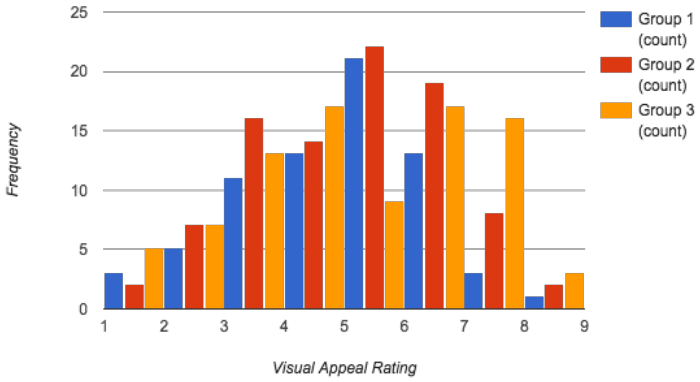


Figure 12 – Histogram to show frequency and visual appeal ratings

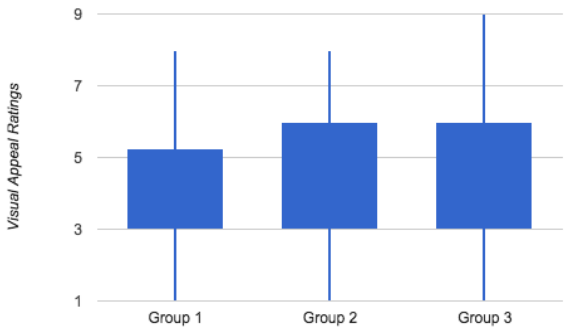


Figure 13 – Boxplot to how visual appeal ratings are distributed

The histograms show that the distribution of the results from each group is approximately normal and that participants who are given more time to view the screenshot will provide a more diverse set of ratings. The skewness for Group 1, 2 and 3 are -0.312, -0.16 and -0.144 respectively. These fall in the region -1 and 1 suggesting that outliers do not exist.

4.3.8 Statistical analysis

The sample sizes were sufficiently large enough that the distribution of sample means should be approximately normal according to the Central Limit Theory. The Levene test for homogeneity of variances is not significant ($p>0.05$) therefore the assumption for the homogeneity of variances is met.

The above suggests that a One-Way ANOVA test can be employed to test for the significance of differences. However, there is ambiguity as to whether the Likert rating scale is interval data or purely ordered categories. Although the scale used was symmetric, equidistant and contains at least 5 scores, the subject of beauty is subjective to each person, and, therefore, the intervals between scores may not be equal. For these reasons, both a parametric One-Way ANOVA test and a non-parametric Kruskal-Wallis test were conducted so as to be cautious in the treatment of the data.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 \neq \mu 2 \neq \mu 3$

The One-Way ANOVA shows that the means of group 1, group 2 and group 3 were 4.43 ± 1.52, 4.62 ± 1.61, 4.75 ± 1.94 respectively. Based on these results, it was revealed that group 3 insignificantly increased as compared to Group 2 and Group 1 ($F = 0.68$, $p=0.504$).

The Kruskal Wallis test also show that the mean rank of Group 1, Group 2 and Group 3 were 116.84, 124.62 and 130.47 respectively. This also meant that Group 3's mean rank is insignificantly increased as compared to group 2 and group 1 ($\chi 2 = 1.45$-, $df = 2$, $p = 0.484$).

Both of the statistical tests prove that the differences in visual appeal ratings are not significant. These findings demonstrate that although participant ratings of visual appeal

are more varied when they are given more viewing time, on the whole, the mean ratings counterbalance each other which result in similar mean ratings.

5. Study 2: Question order in the 5 second test

This study investigated how user's responses and ratings of the website's visual appeal were affected by whether or not the test questions were shown to them before they began the experiment. The methodology for this study will be presented followed by the findings.

This study investigated the effects on user's responses when they saw the questions to the test before they were exposed to the screenshot.

The hypotheses were:

1. Showing the questions to the participants before they view the screenshot will lead to responses with higher correctness scores.

2. Showing the questions to the participants before they view the screenshot will lead to responses with higher word count.

3. There will be a difference in visual appeal ratings when the questions are shown to participants before they view the screenshot.

The method used to define that correctness scores is presented in 4.1.2.

5.1 Methodology for study 2

This study used the same stimuli (see 4.1.2), the same questions (see 4.1.1) and followed the same recruitment process (see 4.1.7) as the first study. The methodology section here only presents the differences between the two studies.

5.1.1 Variables

The findings for Group 2 in the first study were used to compare with the group of responses resulting from this study. The exposure time for both groups was 5 seconds.

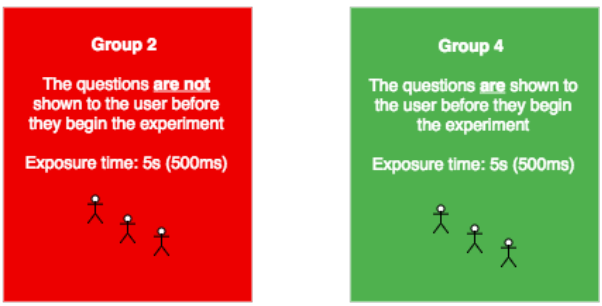


Figure 14 – The groups of participants in study 2

The responses between participants in both Group 2 and Group 4 were compared, this is therefore a between subjects study. The independent variable was:

- Whether or not the participant sees the preview of questions prior to seeing the screenshot

The dependent variables were:

- For Question 1 and 2: The correctness of responses and the word count in each of the responses.

- For Question 3: The visual appeal ratings.

The correctness of the responses was determined by using the same scoring system as study 1 (see 4.1.4).

5.1.2 Procedure

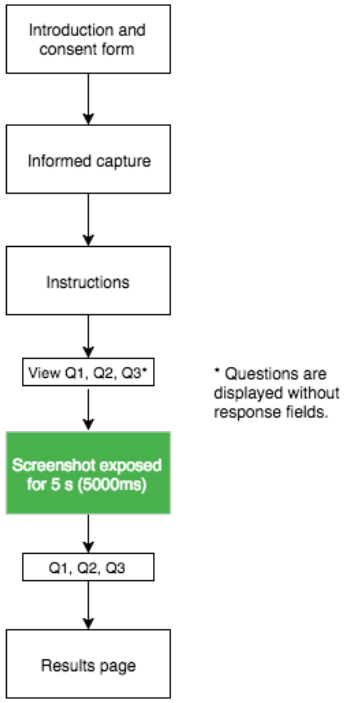


Figure 15 – Procedure for study 2

The procedure for the question order study is outlined below:

1. Participants begin by landing on an information page. This page is exactly the same as in the exposure time study.

2. The next page captures information about the participant.

3. Depending on which group the participant was allocated to, their instructions differed. For those in Group 4, they will still see an extended instructions video.

   This video is similar to the one in the first study except that it notifies the

participant that the next page will present them with the questions that they will answer after viewing the screenshot (rather than going straight to the page that shows them the screenshot)

4. After viewing these instructions, participants will continue to the next page which displays the 3 questions. To continue the participant needs to click a button which reveals the screenshot for a limited amount of time.

5. Once the screenshot disappears, the question pages will appear. These pages are the same as in the first study.

6. The final completion page is also the same as in the first study.

5.2 Findings for study 1

This section presents the findings from the analysis of the data gathered from the second study. Where statistical tests were used, the null hypothesis (H0) is only rejected when $p < 0.05$.

5.2.1 Participants

To conduct this study, the data from group 2 in the first experiment was re-used and compared to a new set of data where participants viewed the questions before they saw the screenshot. This is group 4. It consisted of 98 participants. Note that these participants were chosen from a larger pool of participants as they satisfied the criteria described in the data preparation chapter (see 4.2).

The pie chart in Figure 16 illustrates where the participants were from. All of these participants spoke English as their first language and were aged over 18. Participants in group 4 received a preview of the questions before they saw the screenshot.

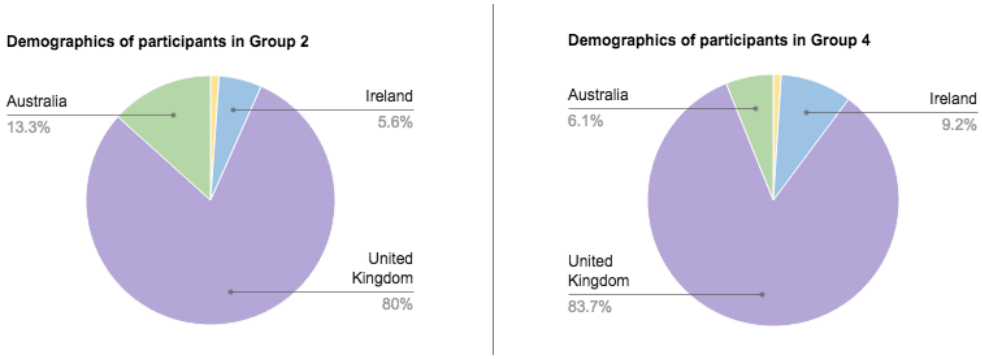


Figure 16 – Demographics for the participants in study 2

5.2.2 Question order and score awarded for correctness

The mean scores for each question increased when participants were aware of the questions they were required to answer after viewing the stimuli. This suggests that the priming of participants had a positive effect on their ability to understand the screenshot. In other words, participants who knew the questions also knew what to look for and then retained that knowledge to answer the questions.

Table 15 – Descriptive statistics to show the correctness scores in group 2 and group 4

	Group 2 (n = 90)	Group 4 (n = 98)
Mean score for Question 1	1.52	2.01
Mean score for Question 2	1.87	2.65

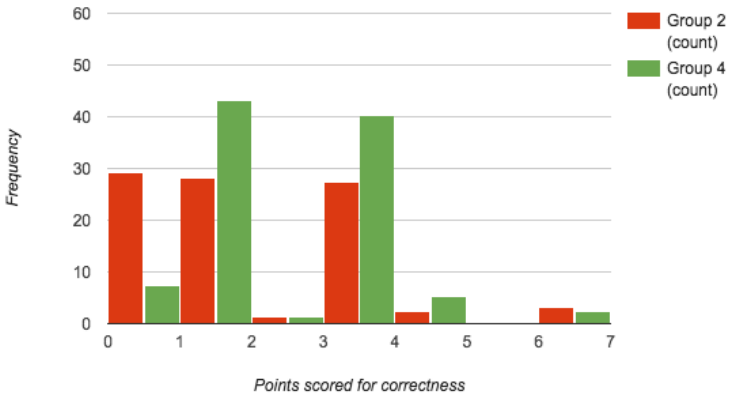


Figure 17 – Histogram to show the correctneness scores between the group that saw the questions before the test began and the group that did not see it for question 1

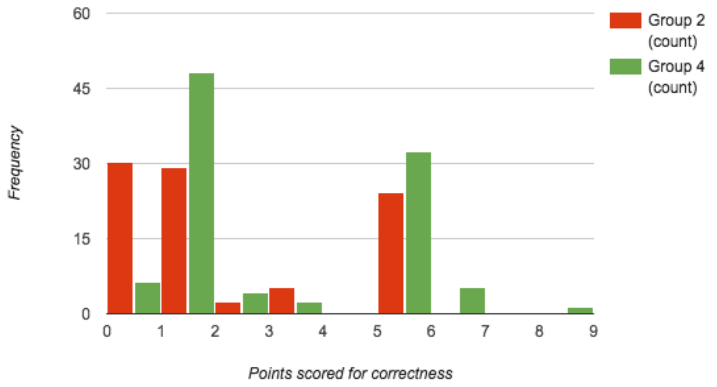


Figure 18 - Histogram to show the correctneness scores between the group that saw the questions before the test began and the group that did not see it for question 2

By visually inspecting the histograms alone, it is apparent that participants in group 4 received higher overall scores. A deeper look into the distribution of each code assigned to participant responses showed that when participants see the questions beforehand they are less likely to provide an irrelevant answer (NA). This is particularly true for question 2.

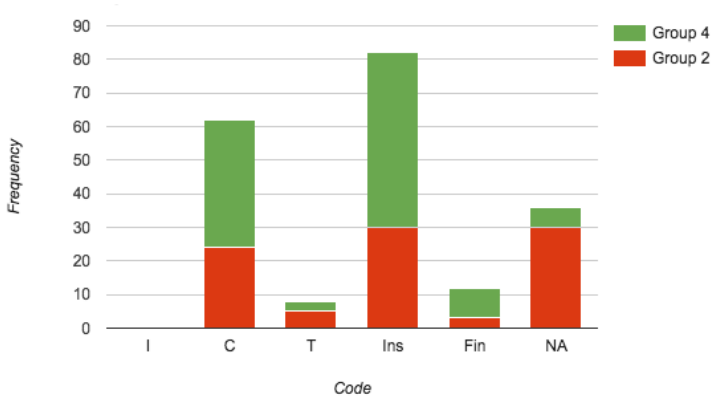


Figure 19 – Histogram to show the frequency of responses in each category in question 2 for both group 2 and group 4

5.2.3 Statistical tests

The Levene test for equality of variances for question 1 and question 2 in this study were 0.172 and 0.1 respectively, therefore this data satisfies the homogeneity of variances assumption. Subsequently, t-tests were performed.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 < \mu 2 < \mu 3$

The t tests conclude that the differences in group means for question 1 are significant as $t(186)=1.880$, $p = 0.019$. They were also significant for question 2; $t(186)=2.729$, $p = 0.01$. Therefore, by viewing the questions before the study, participants responses are significantly more correct and receive higher scores.

5.2.4 Qualitative findings

The tag clouds support the findings from the quantitative analysis show that when participants see the questions beforehand, they provide more concise responses; this is true for both of the questions.

Figure 20 – Tag cloud to show the key words to the questions in study 2

5.2.5 Question order and word count

Table 16 - Descriptive statistics to show the word count in each group

	Question 1		Question 2	
	Group 2 (n=90)	Group 4 (n=98)	Group 2 (n=90)	Group 4 (n=98)
Mean	4.41	3.45	4.51	2.94
Median	2	2	2	2
Mode	1	2	2	1
Range	62	37	59	19

In general, the length of responses provided in group 2 and group 4 are similar. There is a slight increase in the precision of the responses in group 4 as seen by the lower range of values and higher average correctness.

5.2.6 Statistical tests

Both questions passed the Levene's test for equal variances ($F = 0.267$, $p = 0.606$ for question 1 and $F = 1.928$, $p = 0.167$ for question 2). The sample sizes warranted the use of an independent samples t test.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 < \mu 2 < \mu 3$

The t-test results show that the differences in mean word counts for both question 1 and question 2 are insignificant; $t(185) = -0.127$, $p = 0.899$ and $t(185) = 0.899$, $p = 0.370$ respectively. Consequently, the null hypothesis was accepted. Note that the outlier response from participant 605 was removed in these tests.

5.2.7 Question order and visual appeal ratings

Table 17 - Descriptive statistics to show the visual appeal ratings in each group

	Group 2 (n = 90)	**Group 4 (n = 98)**
Mean rating	4.62	4.39
Median rating	5	5
Mode rating	5	3

The results show that both groups of participants are just as likely to give a "moderate" score for visual appeal regardless of whether they saw the question first or afterwards.

Figure 21 – Histogram to show the distribution of visual appeal ratings

The histogram shows that the mode for group 2 is 5 and the mode answer for group 4 is 3. Whilst the mean has remained the same in both groups, this difference could suggest that when people are asked to rate the website's visual appeal before seeing it, as in group 4, they may scrutinise the design more than if they were just browsing the website casually.

5.2.8 Statistical tests

The ratings for both groups were similar in their skewness (-0.16 and -0.08 respectively) and their standard errors differed marginally (1.605 and 1.797 respectively). In the same way that these data meet the assumptions required for parametric analysis, as did experiment 1, both the parametric t-test and the nonparametric Mann-Whitney test were performed.

The hypotheses were:

H_0: $\mu 1 = \mu 2 = \mu 3$

H_1: $\mu 1 \neq \mu 2 \neq \mu 3$

Unsurprisingly the mean differences were insignificant. The t-test shows that the mean difference of group 4 insignificantly differed (4.39 ± 1.7) as compared to group 2 (4.62 ± 1.61), $t = 0.90$, $p = 0.38$.

The Mann-Whitney U test indicated that group 3 (mean rank = 91.27) was insignificantly different from group 2 (mean rank = 98.02) showed that group 2 was insignificantly different as compared to group 4 ($U = 4093$, $p = 0.387$). Therefore the null hypothesis was accepted.

6. Discussion

A summary of the key findings is provided in Table 18 and Table 19.

Table 18 – Summary of key findings from study 1

Hypothesis	Findings
Increased exposure time leads to responses with higher correctness scores.	Increased exposure time from 0.5 seconds to 5 seconds led to responses with significantly higher correctness scores. When exposure time increased from 5 seconds to 10 seconds for the first question in our study, this did not lead to significantly higher correctness scores. When exposure time increased from 5 seconds to 10 seconds for the second question in our study, this did lead to significantly higher correctness scores.
Increased exposure time leads to responses with higher word count.	Increased exposure time (in all of the groups) did not lead to responses with significantly higher word counts.
There will be a difference in visual appeal ratings when exposure times are increased	Increased exposure time (in all of the groups) did not lead to responses with significantly different ratings for visual appeal.

Table 19 - Summary of key findings from study 1

Hypothesis	Findings
Showing the questions to the participants before they view the screenshot will lead to responses with higher correctness scores.	Showing the questions to the participants before they viewed the screenshot did lead to significantly higher correctness scores. This applied to both questions in the test.
Showing the questions to the participants before they view the screenshot will lead to responses with higher word count.	Showing the questions to the participants before they viewed the screenshot did not lead to responses with higher word count.
There will be a difference in visual appeal ratings when the questions are shown to participants before they view the screenshot.	Showing the questions to the participants did not lead to any significant difference in visual appeal ratings.

6. 1 How does exposure time to a website affect participants' responses?

Weinreich et al. (2008) and Lie et al. (2010) both found that users stay on a website for more than 5 seconds before they leave. This indicates that users are likely to stay on a website for more than 5 seconds before they can determine whether the website has the information or has the functionality that the user desires. This is supported by our findings.

6.1.1 Participants with more time read more words.

We found that responses were significantly more correct in the group that was given 10 seconds to view the screenshot when compared to those that had 0.5 seconds for all of the questions. For question 2, participants in the group with 10 seconds provided significantly more correct responses than those who had 5 seconds of exposure time. This result is not surprising and is in accordance with our research hypothesis.

An average person can read 250-300 words per minute on paper and approximately 180 words per minute on a screen (Muter and Maurutto, 1991; Ziefle, 1998). This means that participants would be able to read 1.5, 15 or 30 words when given 0.5 seconds, 5 seconds or 10 seconds respectively. For viewing websites, this is an overestimate as participants are likely to spend some time looking at the pictures. However, increasing the exposure time from 0.5 to 5 seconds and then from 5 to 10 seconds still represents a 1900% and 100% increase of words read.

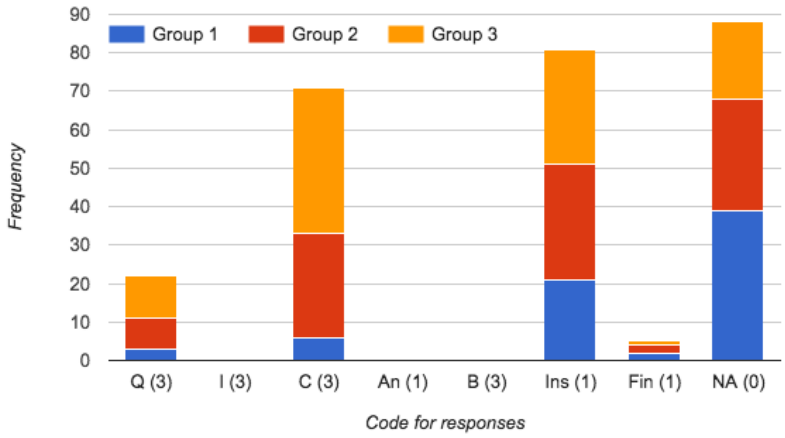


Figure 22 – Histogram to show the distribution of responses in each category for question 1 (for groups 1, 2 and 3)

Given the same conditions, when a participant has more time to read the words on a website, they will have a higher chance of understanding more of the content and subsequently be more able to recall information. The significant increase in correctness between 0.5 seconds and the other groups is evident when we look at the distribution of the codes assigned to the responses Figure 22. We can see that group 2 responses were more explicit and detailed as it contains more "Car insurance" (C) in relation to the more general responses such as "Insurances" (Ins) code in group 1.

6.1.2 The questions in the test matter

Several of the experts were able to recall that the website was about "gathering information related to car insurance" (code I) or "benefits of using Statefarm" (code B) but in our analysis of the responses collected from participants, we found that none of these themes were raised.

Given the informative nature of the website, we found this to be surprising. We suspect that this is the result of the nature of conducting tests online where participants input the least amount of information possible so that they can complete the test as quickly as possible (West and Lehman, 2006).

The anonymity aspect of the test may add to this. Bargh et al., (2004)'s conducted an investigation into the psychology of internet users and how this impacts the way users communicate online. They find that anonymity encourages self-expression, but the lack of physical or non-verbal interaction cues means that "a lot tends to be left unsaid and open to inference and interpretation".

In the case of our studies, it is rather unlikely that participants were oblivious to the fact that there was information about car insurance. They probably knew the website was about insurance and contained information about the policies, but they did not respond with this extra information. By only entering a phrase like "car insurance", in the mind of a participant, they have already implied that it would have information about the topic. To elicit responses about what information was on the website, we would have to be more direct and include another question. This shows that a successful exposure test must not rely on participants to volunteer information. By asking short and direct questions, we are likely to get more correct responses.

6.1.3 The expected answers to the questions matter

For question 1 in the test, we found that the difference in correctness between the responses in the group with 5 seconds and 10 seconds were not significant. In contrast to this, responses to question 2 were found to be significantly more correct when participants had 10 seconds of viewing time compared to 5 seconds. This raises a question about why the responses were significantly more correct for the second question but not the first one.

We can rule out the possibility that participants responded with longer responses for question 2 because the word count for question 1 and question 2 are similar. The reason for the differences lies with the fact that the experts gave a wider variety of responses for question 1.

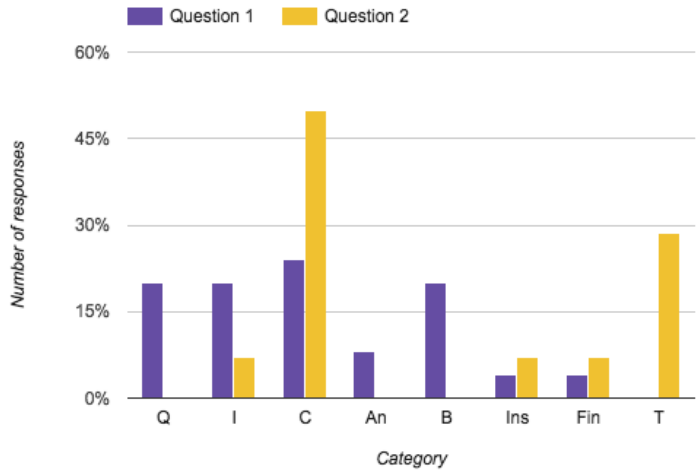


Figure 23 - Histogram to show the percentage of expert answers in each category

Table 9 in section 4.2.6 shows how the scoring system was devised by looking at the number of themes in the answers provided by the experts. Figure 23 shows that the answers provided by the expert to question 1 were more varied and stretched across a range of themes. In contrast to this, in response to question 2, half of the expert's responses were just related to car insurance. This meant that the experts expected a wider range of possible answers to be correct for question 1 and a more narrow range of answers to question 2.

As a result of deriving the scoring system from the expert's answers, answers that related to "Car Insurance" (code C) were awarded a score of 3 for question 1. On the other hand, the same theme (category) was awarded a score of 5 for question 2. The implication of this scoring system meant that for questions that are more open (experts will expect a variety of "correct" answers), when participants are given more exposure time, they can only score high if they write more in their responses. Given that the word count of responses are similar regardless of time, we know that the test is unlikely to be able to motivate participants to give longer responses that can capture the other responses provided by the experts.

6.1.4 Exposure time does not affect the length of responses

The word count remained around two words in all of three groups in this study. This is most likely due to participants wanting to complete the test correctly with minimal effort, and they did not know that they would get better scores if they provided a variety of more detailed responses. The instructions in the 5 second test do not instruct nor motivate participants to provide more varied and more detailed questions. This is an area that could be investigated in future research studies.

6.2 How does exposure time to a website affect participants' perception of its visual appeal?

Visual appeal ratings they were found to be insignificantly different between the groups with different expose times. This is consistent the findings by Lindgaard et al. (2006) and Cyr and Head (2013).

6.2.1 Pictures on a website have a large impact

In Cyr and Head (2013)'s study, it was found that regardless of how instructions were framed, 51/60 (85%) of the participants in the group with 5 seconds said that they looked at the picture because of its prominence and visual attraction. In contrast to this, in the group without the time constraint, only 18/60 (13%) said that they saw the picture. When these participants did comment about the pictures, they focused on style and emotion caused by the pictures rather than its prominence or attention grabbing properties.

Their insights help us to explain the impact of pictures on a website – especially when the viewing time is brief. They emphasise how pictures with facial features attract attention and the importance of using pictures to engage and inform users. In our studies, the picture of the people washing the car on the State Farm website did not help the participants in our study to understand the purpose of the page. Instead, it led participants in group 1 and group 2 to think that the page was about "healthcare", "parenting", "car wash service", "teaching", "health" or "hobbies". This may also explain why there is a higher standard deviation (more varied ratings) in the visual appeal ratings as the exposure time increased.

Our findings show that when people are given more time to judge the website, aside from looking at the pictures, they take other factors into consideration. In terms of Leder et al. (2004)'s model of visual processing, when participants have more time, they move beyond the perceptual analyses and implicit memory integration processes and proceed to the explicit classification stage. In this stage and beyond, the viewer will take into account of the style and the content. These factors may also include content that is below the fold. In addition to this, it is possible that participants with more time took into account the content below the fold when making their judgements on visual appeal.

It could also be that that people are giving a more thought into what they judge to be beautiful as opposed to using heuristics to make judgements. For the stimuli used, it can be seen that in general the ratings from those in group 3 were more ″generous″ than those in the other groups with the most frequent scores being 4, 6 and 7 compared to group 1′s 4, 5 and 6 and group 2′s 3,5 and 6. This could suggest that the longer the participant has to view the website, the more time they have to appreciate some of the finer visual details of the website than those with less time to view it.

6.3 What difference does it make if the test questions are shown to the participant before they begin the experiment compared to when they see it afterwards?

Unsurprisingly, priming the participants with the questions led participants to selectivity attend to certain aspects of the screenshot. Priming is an implicit memory effect, where the exposure to some stimuli influences a person′s subsequent responses (Meyer and Schvaneveldt, 1971). The priming effect allowed them to selectively process the information on the screenshot, thus allowing the participant to prioritise some aspects of information while ignoring others.

This positive priming appears to have sped up the participants' processing of the website. Rather than looking around at the whole website, group 4 participants skimmed through the information quickly in an attempt to find out how to respond to the questions that they saw beforehand. Once they found the answers, they may have tried to consciously remember them. This explains why they were able to achieve significantly more correct responses than those in group 2.

These results bear some resemblance to Bruun and Stage, (2012)'s study where predefined tasks led to the discovery of more usability problems when compared to user-defined tasks. Attention is a selective process which turns looking into seeing (Carrasco, 2011). By showing the questions to the participant, it allowed them to focus on certain locations on the website. The implications of this are discussed further in section 6.6.5.

6.5 Additional analysis

We saw that as exposure time increased, both the visual appeal ratings and correctness increased and therefore Spearman's Rho test was calculated to determine whether there was statistical dependence between the two variables. This test was chosen as it can be applied to non-parametric data, as is the case for the visual appeal ratings. The results from the test show that there is no dependence as $r=0.017$, $p=0.758$.

6.6 Practical implications

When taking into consideration of all these findings, we can see that there are practical implications for evaluators who use the 5 second test. These are explained in the following section.

6.6.1 Exposure time depends on the expected answers

We have seen that the exposure time required for an effective 5 second test depends on the level of detail that the evaluator wants in the responses. This means that 5 seconds may be sufficient for simple questions that require little detail in the responses i.e. generic one-word answers. In contrast, if an evaluator wanted to test for more details about what users understand, then more time is required.

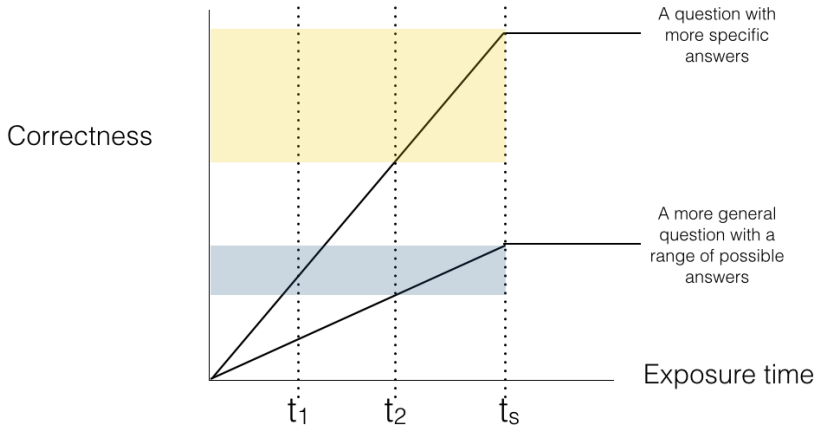


Figure 24 – The relationship between correctness scores, exposure time and the question asked

Figure 24 shows the relationship between the variables found in our study. As we can see, the difference in correctness depends on how specific the question is, which in turn is dependent on what kind of answers the evaluator expects or defines as correct (see 6.1.3).

Note the line becomes flat here because beyond a certain time, more exposure time will not lead to significantly more correct responses (as evident in question 1 of our study).

6.6.2 Using tag clouds for analysing 5 second test data

Tag clouds are the most commonly used method for analysing data collected from 5 second tests. We have seen from the results section that this is a weak method of analysis and can, in fact, mislead evaluators. Although these tag clouds are useful, they do not show how the words were combined. For example, in our study, the word "car" itself will not receive a score for correctness unless it is written as "car insurance". This means that the tag cloud method of analysis is more suitable for questions with one-word answers. More importantly, the size of the text in a tag cloud is only proportional to the total amount of words in that group which means that just by inspecting the diagrams visually, it is not possible to accurately compare between the groups.

The method of analysis used in this project, offers a more powerful solution which results in better interpretations in two ways. First, a predefined criteria for scoring means that evaluators can objectively measure whether their design meets their expectations. Secondly, the depth of analysis can draw out many more themes than simple tag cloud analysis as we go beyond counting words and examine responses as a whole.

6.6.3 Determining whether to show the questions or not

Showing the questions before the test has a noticeable effect on the results of the test. This means that an evaluator should consider how users will arrive at their website before they set up a 5 second test.

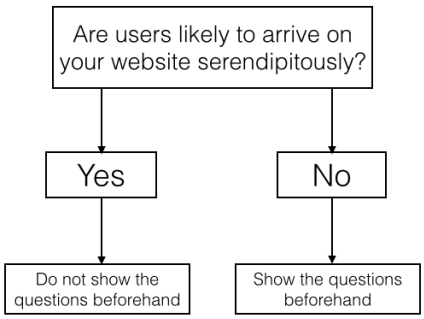


Figure 25 – How to determine whether to show the questions or not

If the website is expected to be arrived on serendipitously, it would be sufficient to test the

website without showing the questions beforehand. If the website is intended to be used when users are performing "search browsing" then the evaluator should show the questions to the participants first, to reflect how the website would be used in the real world when they have certain goals they wish to achieve in their search. Otherwise, the test results would be significantly different as shown in our second study.

6.6.4 Collecting more data to inform improvements

Despite having evidence of what the test takers understand and recall from our studies, we do not know which elements of the website caused them to think in this way. From our results, we can infer which elements the participants saw but we cannot be certain, nor do we know in which order they saw the content.

As an example, we can see that the phrase "auto quote" appears in two prominent positions above the fold as shown in Figure 26. Without asking the participants or using eye tracking devices, it is not possible to determine which element attracts more attention. This could mean that the test could be improved by having a separate group of participants describe what they saw and then combining the datasets.

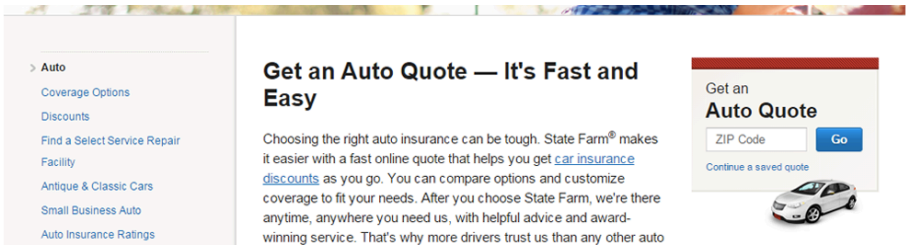


Figure 26 – The placement of the phrase "auto quote" in the screenshot

6.6.5 Areas for further research

The findings on the word counts indicate that the 5 second test is not effective at eliciting more details about what participants understand, even if they have that information. This may be improved if we were to change the way that participants respond to the questions. This suggests that allowing participants to respond by typing in answers in a text box may not be optimal. For example, instead of using text boxes, checkboxes with possible answers could be used (see Figure 27). This reduces the effort required from participants and encourages them to provide more than one answer. It also changes the type of memory required, from recall to recognition, which may be more appropriate given that we are interested in gathering data on what they saw and understand rather than what has been stored in their short-term memory.

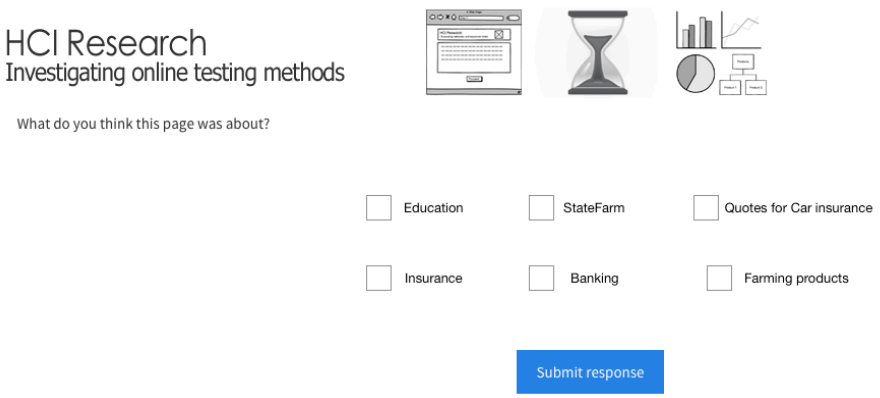


Figure 27 - Example of how response types can be changed

This study only investigated three different exposure times, and we have seen hints that the variances for judgements of visual appeal destabilise after the initial first impressions (0.5 seconds). This means that there may be an exposure time beyond the initial impressions (50ms as determined by Lindgaard et al., 2006) at which the visual appeal judgements stabilize. This is the exposure time at which users have made a confident judgement about the aesthetics of the website as a whole. This is important for websites that are frequently visited and, therefore, is concerned not only about the first impressions for a first-time visitor but is interested in maximising its visual appeal for all of its users. Future research into the 5 second test may investigate the relationship between the exposure time and the type of website to be tested.

This study only used one website as stimuli and, therefore, provides no insight into this area. It could be possible that less time is required for simple content pages with only informational content and more time is required for landing pages with navigation options and other calls to action.

6.4 Feedback about the studies conducted in the project:

We received 93 comments as feedback on the studies. These can be seen in Appendix L. Most of the comments were positive and praised the design of the tests.

Of the 93 comments, 14 were participants who were included in our analysis. As expected there were many responses from people who found that the screenshot timed out too quickly. Of the 14 in our analysis, 5 of them had 0.5 seconds of exposure time and of these 4 complained that the screenshot only appeared for an extremely brief period and they "thought it was a mistake". 4 of the 14 in our analysis were given 5 seconds. Of the four people that had 5 seconds of exposure time, 2 of them commented that the viewing time was too short.

7. Conclusion

This project has made a useful contribution to the understanding and knowledge about the 5 second test. By recreating a 5 second test platform that enables the exposure time, response field type and whether or not the questions can be viewed before hand, two online studies were conducted which provide insights into the effectiveness of the 5 second test.

The findings provide evidence to support that 5 seconds may not be enough to evaluate the clarity of content. The exposure time depends on the level of detail that is required to answer the question correctly. This means that if an evaluator wanted generic one word answers, then 5 seconds may be sufficient. However, if explicit detail is desired in the responses, then more time is required.

The results show that regardless of the exposure time and whether the question is seen beforehand, participants in the 5 second test will write short responses, typically around two to three words. Although the amount of words in the responses remain relatively constant, the level of detail does vary depending on exposure time. The findings also support the previous research that visual appeal ratings can be reliably formed within 0.5 seconds.

This study also demonstrates a new technique which can transform a large amount of qualitative data collected in a 5 second test into quantitative data that allows for more detailed insights to be revealed.

8. Reflection

The overall objective of this project was to investigate the 5 second test. The studies were planned, designed, carried out and analysed successfully. In this final chapter, I discuss what I have learned by completing this project.

Originally, I had planned the project to include a much larger scope. I had intended to test many different websites under different exposure times to determine the relationship between the attributes of a website and the exposure time.

Testing many different websites with many different exposure times would have meant collecting a much larger set of data. I did not proceed with this because I thought that the time cost for data collection would be too large given the time constraints of the project. With hindsight, given the success of my participant recruitment strategy, this initial plan may have worked. However, if each participant were to test many different websites in each test, then the test would have taken longer to complete which may have affected the effectiveness of the recruitment.

My approach to conducting the literature review was very comprehensive but not efficient. Since I knew there would be a limited amount of research that was specific to the 5 second test, I read extensively about remote evaluation and first impressions in various different domains. Although this provided a formidable understanding in these areas, a large amount of literature that I thought might have been useful turned out to be irrelevant. For example, reading about how humans process information from a cognitive psychology point of view was too complicated to be incorporated into this topic. It would have been more effective if I had planned how much time I should allocate to each topic in the literature review rather than aiming to read as much as possible hoping that it may "fit in" somewhere.

Nonetheless, the literature review did reveal various insights that I took into consideration for my project, for example, the findings from Weinreich et al. (2008) and Lie et al. (2010) helped me to determine which exposure times to choose in the study.

As I have not programmed the "backend" of a website before, and I have not coded anything for a long time, I found it challenging to build the testing platform. I taught myself how to make the forms work and make the system collect data reliably. Of course, I would not have been able to do this without asking for programming advice on online communities such as Stackoverflow.com so I am really thankful for the existence of such communities.

The build of the platform was successful but in retrospect, the pilot testing that was conducted was not sufficient. I tested the platform with too few people and, therefore, did not figure out all of the ways in which a participant could hack the system. I only found out that a participant could press "back" on their browser and view the screenshot again when someone commented on LinkedIn. Although this did not affect the results in a major way (as responses in the 0.5 second group were significantly less correct), I felt that this flaw could have caused the data to be useless.

I enjoyed analysing the data but initially, I was overwhelmed by all of the data that I had collected. I had to plan how I would organise, sort and filter the data that was exported from the SQL database as the platform was not designed to facilitate this type of analysis. As soon as the data was organised into groups, I was highly motivated to find out whether my hypotheses were true or not.

I overcame the issue of measuring correctness by having experts review the stimuli. This idea was not mine, and it was only possible by using what I had learned from the Human Centred Systems course and using the advice of my supervisor. From this, I have seen that reading academic literature is not just about discovering what insights a fellow researcher has uncovered, there is also much to be learnt about the way that the researcher has created an experiment to elicit these findings.

The timing for the project felt limited but I felt that I did not struggle as I had imagined. This was because I was able to combine the two studies so that they ran at the same time and the plan that I created for recruitment turned out to be successful. In terms of planning, if I were to do the project again, I would allocate more time to focus on creating measurable research questions rather than refining them as the project progressed.

Overall, I am pleased with my achievements and what I have learnt through the process of completing this dissertation project.

References/bibliography

Albert, B., Tullis, T. and Tedesco, D. (2010). Beyond the usability lab. San Francisco, Calif.: Morgan Kaufmann.

Dijksterhuis and L. Nordgren, "A theory of unconscious thought," Perspectives on Psychological science, vol. 1, no. 2, p. 95, 2006.

Garrett, J. (2011). The elements of user experience. Berkeley, CA: New Riders.

Gladwell, M. (2006). Blink. London: Penguin.

Gs.statcounter.com, (2015). StatCounter Global Stats - Browser, OS, Search Engine including Mobile Usage Share. [online] Available at: http://gs.statcounter.com [Accessed 22 Sep. 2015].

Hartson, H., Castillo, J., Kelso, J. and Neale, W. (1996). Remote evaluation. Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96.

Harvard Business Review, (2013). The Rise of UX Leadership. [online] Available at: https://hbr.org/2013/07/the-rise-of-ux-leadership/ [Accessed 16 Sep. 2015].

Hix, D. and Hartson, R. (1993). Ensuring Usability Through Product and Process. Wiley.

Jewell, C. and Salvetti, F. (2012). Towards a combined method of web usability testing. Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA '12.

Leder, H., Belke, B., Oeberst, A. and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology, 95(4), pp.489-508.

Lukew.com, (2015). LukeW | There Is No Fold. [online] Available at: http://www.lukew.com/ff/entry.asp?1946 [Accessed 22 Sep. 2015].

Muter, P. and Maurutto, P. (1991). Reading and skimming from computer screens and books: the paperless office revisited?. Behaviour & Information Technology, 10(4), pp.257-266.

Nngroup.com, (2015). F-Shaped Pattern For Reading Web Content. [online] Available at: http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/ [Accessed 18 Sep. 2015].

Robins, D. and Holmes, J. (2008). Aesthetics and credibility in web site design. Information Processing & Management, 44(1), pp.386-399.

Roth, S., Tuch, A., Mekler, E., Bargas-Avila, J. and Opwis, K. (2013). Location matters, especially for non-salient features–An eye-tracking study on the effects of web object placement on different types of websites. International Journal of Human-Computer Studies, 71(3), pp.228-235.

Spiliotopoulos, T. (2010). Integrating usability engineering for designing the web experience. Hershey, PA: Information Science Reference.

Stephanidis, C. (2009). The universal access handbook. Boca Raton: CRC Press.

Tractinsky, N., Katz, A. and Ikar, D. (2000). What is beautiful is usable. Interacting with Computers, 13(2), pp.127-145.

Tuch, A., Presslaber, E., Stöcklin, M., Opwis, K. and Bargas-Avila, J. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. International Journal of Human-Computer Studies, 70(11), pp.794-811.

Uie.com, (2015). Usability Tools Podcast: 5-Second Usability Tests » UIE Brain Sparks. [online] Available at: http://www.uie.com/brainsparks/2007/09/10/usability-tools-podcast-5-second-usability-tests/ [Accessed 18 Sep. 2015].

Van Schaik, P. and Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. International Journal of Human-Computer Studies, 67(1), pp.79-89.

Ziefle, M. (1998). Effects of Display Resolution on Visual Performance. Human factors, 40(4), pp.554-568.