Improving user interfaces through heuristics analysis, think alouds, and eye gaze tracking

A comparative analysis

Research Report
Abstract

Usability testing has a demonstrated success in improving the quality of user experience for a product or service. Discount usability testing is the recommended usability testing method in industry. Heuristics analysis and think alouds are the most prominently used discount usability testing methods. The nature of feedback from these methods is typically not benchmarked, which makes it difficult to compare the results of usability tests with each other from different products or services.

This project aims to create a structured framework with which heuristics analysis and think aloud can be performed, and achieve indexed results which are comparable for multiple products and services. The frameworks relies on abstracting and encapsulating principles and theories in human-computer interaction, and enables the development of context independent user experience analysis.

It was observed that structured heuristics analyses can help develop a whole-of-system health perspective about the interface and user experience design. Think alouds, on the other hand, allow pinpointing specific elements of the user experience in addition to enabling assumption testing and behavioral analysis.

A metacognitive analysis lens towards the think alouds rendered the insight that an average of 74 verbalisations could be tallied in a 40-minute think aloud session, of which 18.3% indicate clarifying responses, 16.6% synthesis, and 15.6% personal responses. These findings add value to understanding the nature and efficacy of Nielsen’s usability heuristics.

Finally, it was observed that only 25% of the participants could use second-order deductive reasoning for task completion, which helps develop further insights into how usability tasks should be constructed.

Keywords:
Human Computer Interaction, User Experience, Usability Testing, Heuristics Analysis, Think Aloud, Eye Tracking, Metacognition, Design Principles
1. Introduction

Inspired by Asimov’s laws of robotics, Jef Raskin proposed two laws for interface design [1], they read as follows:

“First Law: A computer shall not harm your work, or, through inactivity, allow your work to come to harm. Second Law: A computer shall not waste your time or require you to do more work than is strictly necessary.”

What Raskin was trying to say, and what is now at the heart of what we now call design thinking, is that a good system is a system which dances around the user, and doesn't make the user dance around it, and the beauty of good design lays in the subtlety and grace of this dance.

In the last 10 years or so, the world has seen a shift in its approach towards creating products, services, and capabilities that are becoming progressively user and human centered. This is evident in government led initiatives of digital transformation across a range of Australian government organisations, and in the industry. This shift in focus from iterative improvements to technology as opposed to disruptive focus on user was Microsoft’s corporate strategy well until recently, when under their new CEO, Satya Nadella, they have now begun to create disruptive and innovative cross-platform solutions with the user in mind. And, the results have been phenomenal.

There are many dimensions to human centered design, human centered computing, and human computer interaction. A vital component of this larger picture is usability testing. Usability testing is a method to test whether or not a given piece of technology is something it's intended user would like to use, would want to use -- and there are numerous ways to do that. Some of these methods of usability testing are more rigorous and expensive than the others, and each one serves a purpose and meets specific outcomes relevant to the objective of their use.

In the user experience design community, however, preference is generally towards inexpensive methods of usability testing. There are numerous reasons for this, which will be explored as part of this research report. In general, the preferred methods are think alouds, and heuristics analysis. This project aims to evaluate the goodness or lack thereof of both these methods, and how a structured and methodological approach can help achieve new definitions of success with these usability testing methods.

By testing assumptions, behaviours, and the nature of interaction early enough, great products and services can be designed with minimal wastage of unnecessary resources, and drive the development of incredible things. If we can save 5 minutes of a user’s time or energy through good design, that is time they can spend with their family, friends, or doing other fun and important things. This is why we put users at the heart of design, and this is why the right kind of usability testing is important.
2. Background

2.1. Motivation for this Project

Too often products and services are designed without fully considering the user in mind. With services this typically translates to unhappy citizens, with products this translates to unhappy end-users.

One could question the success of a well-engineered service or product if the people for whom they are intended are not pleased with it. In computer science, the field of human-computer interaction (HCI) aims to address is this, by putting human at the center of design.

One of the key drivers in HCI as a discipline is usability testing, and the industry standard best practice to ensure quality usability tests is to do them early in the development lifecycle, in an inexpensive manner. This can be achieved through heuristics analysis, and think alouds -- both of which allow rapid generation of valuable user feedback at low resource expenditure.

However, given the low-fidelity nature of these methods, the results from studies are difficult to compare and index. The aim of this project to understand and evaluate how a cross-platform benchmark, and a structured framework to implement heuristics analysis, and think aloud can help enhance the quality of usability testing, in addition to making the results context independent.

To achieve this, this project undertook studies with expert and novice users to compare and contrast their experience in finding usability issues as a result of the different methods used, and their experience with these methods as a whole.

2.2. The Test Object: HCC Workshop

The HCC Workshop is a collection of hands-on activities that demonstrate the research that the Human-Centred Computing Research group carries out at ANU. Human-Centred Computing uses the latest technology to explore computer-human interactions to make computers 'think ahead', and be responsive in their interactions with humans. We want computers to understand what users’ want when they perform specific tasks, and to enhance their user experience.

In the HCC Workshop, students are able to play with some of the latest technology that we use in our research, such as having their eye gaze tracked. Students will have the opportunity to play games and have fun learning about leading research in an exciting blend of different fields such as computing, art, science, music, security, psychology, social network analysis, digital humanities and biologically inspired computing.
Table of Contents

Abstract 2

1. Introduction 3

2. Background 4
   2.1. Motivation for this Project 4
   2.2. The Test Object: HCC Workshop 4

3. Table of Contents 5

3. Heuristic Analysis 8
   3.1. Overview 8
   3.2. Why use Heuristics Analysis: An analytical model for understanding 8
   3.3. Nielsen’s Heuristics: Understanding and explaining usability problems 9

4. Design Principles 11
   4.1. iOS Human Interface Guidelines: An overview of design principles in currency [23] 12
   4.2. Materialising design concepts and principles into final products 13

4.3. Heuristics for Analysis 14
   4.3.1 Usability Criteria: Dialogue Principles (1996) 14
   4.3.2. Evaluation Criteria and Sub-Criteria 15
      4.3.2.1. Analysis of Benchmarks and Standards 15
      4.3.2.1.1. The ISO/IEC FDIS 9126-1:2001 [48] 15
      4.3.2.2. Criterion Used as Metrics 16
      4.3.2.2.1. Sub-Criterion that comprise Metrics 16
   4.3.2.2. Rating Methodology 17
      4.3.2.2.1. Quantitative 17
      4.3.2.2.2. Qualitative 17
   4.4. Application Framework: A contextual update of existing influential heuristics 18
      4.4.1. Bringing together worlds of knowledge 18
      4.4.2. Developing A Score To Determine Usability 19
      4.4.3. Future Improvements To The Score 19

5. Implementation of the Heuristics Analysis 20
   5.1. Method 20
   5.2. Alternative Frameworks Used To Draw Comparative Results 20
   5.3. Implementation on the HCC Workshop 21
5.3.1. Observations and Score from Expert #1 21
5.3.2. Observations and Score from Expert #2 21
5.3.3. Observations and Score from Expert #3 22
5.4. Key Findings 23
5.4.1. Future Studies 23

6. Think Alouds 24
   6.1. Overview 24
   6.2. Method 25
   6.3. Analysis 26
      6.3.1. Participants Selection and Representativeness 26
      6.3.2. Metacognitive processes analyses 26
      6.3.3. Task Completion 27
      6.3.4. Exit Interview Feedback 27
   6.4. Insights 28
      6.4.1. Metacognitive processes analyses 28
      6.4.2. Task Completion 29
      6.4.3. Exit Interview Feedback 31
   6.5. Comparison With Heuristics Analysis 32

7. Retrospective Think Alouds with Eye Tracking 34
   7.1 Overview 34
   7.2 Analysis 34

8. Conclusions and Future Work 35

9. Acknowledgements 36

10. References 37

11. Appendix 1 40
    11.1 Project Description 40

12. Appendix 2 41
    12.1 Study Contract 41

13. Appendix 3 43
    13.1. Participant Information Sheet 43
    13.2. Screener Questionnaire 46
    13.3. Written Consent for Participants 47
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.4. Demographic Information Questionnaire</td>
<td>48</td>
</tr>
<tr>
<td>13.5. Experimental Plan</td>
<td>49</td>
</tr>
<tr>
<td>13.6. Task/Activity Information Sheet</td>
<td>51</td>
</tr>
<tr>
<td>13.7. Surveying Document - Think Aloud</td>
<td>52</td>
</tr>
<tr>
<td>13.8. Surveying Document - Task Completion</td>
<td>53</td>
</tr>
<tr>
<td>13.9. Exit Interview</td>
<td>54</td>
</tr>
<tr>
<td>13.10. The Heuristics Analysis Framework</td>
<td>55</td>
</tr>
</tbody>
</table>
3. Heuristic Analysis

3.1. Overview

Heuristics evaluation [2] is a discount usability inspection method [3] that has seen widespread use in evaluating usability issues [4] in the user interface design of software and hardware. A typical heuristics evaluation involves a set of evaluators inspecting the user interface against a set of usability principles called ‘heuristics’ [5]. Heuristics analysis subsumes a number of principles from multiple user-centered systems design paradigms, and allows the ‘expert’ evaluator to generate insights about a system’s usability rapidly [6]. This is of particular importance in usability engineering, as usability evaluations for small or medium scale systems or interfaces -- new media content or low-risk applications -- tend to be limited by the time and budgetary constraints.

3.2. Why use Heuristics Analysis: An analytical model for understanding

The gold standard for discount usability through most of the 1990s focused heavily on comprehensive studies driven to generate quantitative data for subsequent evaluation by researchers. This form of usability evaluation and testing proved to be expansive and expensive [7]: a large amount of effort is required to set up ‘clean’ laboratory experiments; usability reports require a significant amount of effort to be generated, and thus are slow in delivery; developers and usability engineers would encounter mismatches in mental models due to the slowness of delivery of usability reports and analyses. The cost of running such thorough but resource heavy exercises may not be justified in the context of relatively simple and low-risk applications.

Nielsen [8] talks about the cost-benefit analysis of running data-driven empirical studies in various contexts, and advocates the use of simplified methods to run usability studies as it allows researchers to focus on early adoptions of changes in the research protocols, in addition to frequent usability input because of the underpinnings of early and rapid iterations -- something that the software industry has now come to love with the recent popularisation of lean and agile development. Carroll and Rosson make similar arguments, and propose approaching HCI research from the perspective of action science [9], and in doing so subsuming contextualization, mediated evaluation, participatory design, and ethnographically-informed design. Heuristics analysis and evaluation enable researchers to thus employ a broad spectrum of operational principles with relative ease and flexibility, in a time/budget constrained environment.

GOMS (Goals, Operators, Methods, and Selection Rules) is a recognised framework of systematic evaluation [10] of the goals, methods, and actions that typically constitute an interaction between users
and a software system. There have been attempts to extent models such as GOMS towards automation so as to minimise effort required by HCI researchers, these attempts have nevertheless not seen much traction due to multiple reasons, especially the complexity of incorporating sophisticated UX research methods into the development process. Expert heuristics analysis and evaluation are, primarily due to their relative simplicity, easier and cost-efficient for such development process cycles. Research has indicated that overly complex UX research protocols deters developers from addressing UX issues [3], in contrast to relatively simpler methods to which researchers seem to have a comparative affinity [11]. Thus, by the measure of the likelihood of actually being used, expert usability analysis stand out as one usability evaluation is infinitely more than zero.

3.3. Nielsen’s Heuristics: Understanding and explaining usability problems

Heuristics can be typically defined as a set of protocols or methods that can be employed in a practical manner to achieve -- perhaps imperfect -- sufficiently optimal objectives. In the context of HCI design and principles, heuristics are a set of evaluators used to inspect the interface with respect to a small pool of broad usability principles [12]. Nielsen’s heuristics [6] are widely used as a benchmark for usability inspection. The final set of published heuristics are [13]:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

These heuristics were formulated using a factor analysis to determine which factors account for the greatest amount of usability problems from a database of 249 usability problems collected by Nielsen [12]. These factors, and the resulting heuristics framework has been used heavily in interface design and analysis, and forms the veritable foundation of a large pool of work in ergonomics and interface design [14-18]. The seven factors that account for more than 3% of the variance of usability accounts, and their constituent heuristics with loading of .60 and more are detailed as follows:

1. Factor 1: Visibility of the system
   1.1. A5 Feedback: keep users informed about what goes on
   1.2. C8 Provide status information
   1.3. F7 Feedback: show that the input has been received
1.4. E13 Features change as user carries out task

2. Factor 2: Match between systems and real world
   2.1. A2 Speak the user’s language
   2.2. C7 Contains familiar terms and natural language
   2.3. G2 Speak the user’s language
   2.4. F1 Metaphors from the real world

3. Factor 3: User control and freedom
   3.1. G23 Undo and redo should be supported
   3.2. D4 Obvious way to undo actions
   3.3. F8 Forgiveness: make actions reversible
   3.4. C18 Ability to undo prior commands

4. Factor 4: Consistency and standards
   4.1. A4 Consistency: express same thing same way
   4.2. B5 Consistency
   4.3. F4 Consistency: same things look the same

5. Factor 5: Error prevention
   5.1. A9 Prevent errors from occurring in the first place
   5.2. G22 System designed to prevent errors

6. Factor 6: Recognition rather than recall
   6.1. F3 See-and-point instead of remember-and-type
   6.2. D1 Make the repertoire of available actions salient

7. Factor 7: Flexibility and efficiency of use
   7.1. G14 Accelerators should be provided
   7.2. A7 Shortcuts: accelerators to speed up dialogue
   7.3. B8 User tailorability to speed up frequent actions

Close analysis of the results of Nielsen’s study indicates that ‘Visibility of systems and status’ and ‘Match between systems and real world’ are the most important components of usability problems, and account for approximately 12% of all such issues studied. Furthermore, the leading heuristics (with the highest loading values) for each of the seven factor appear to be: A5 Feedback: keep the user informed about what goes on; A2 Speak the user’s language; G23 Undo and redo should be supported; A4 Consistency: express same thing same way; A9 Prevent errors from happening in the first place; F3 See-and-point instead of remember-and-type; G14 Accelerators should be provided.
4. Design Principles

In his book, The Design of Everyday Things, Donald Norman talks about *knowledge in the head and knowledge in the world* [19] and tries to illustrate the ease of fallacious lines of reasoning being formulated in the users’ mental models, which consequently results in limitations in their ability to achieve specific objectives as intended by the designers. “Precise behaviour can emerge from imprecise knowledge”, says Norman. As such, the design of interactive interfaces must be aware of natural and artificial constraints. In doing so, the designers can minimize the completeness, precision, accuracy, and depth of knowledge that the users must internalise to be able to accomplish a task or goal. It is important to understand that memory is fluid knowledge in the head, but it only becomes crystallized through repetition, so if the interface is not intuitive, the designers must ask themselves if their design is truly user-centric?

Modern websites, largely follow a three-partite design layout that address three main questions for users: *Where am I? What's here? Where can I go?* [20] The placement of relevant information that address these questions may vary depending upon the nature and type of interface [21], but these questions must nevertheless be answered with minimal noise, and a lot of clickability. This is to say, that each page much exhibit a simple visual display, which lots of clickability. If the users want more, they can find more, and these more features must be hidden away in more, and not distract the users from their primary goals. Nevertheless, this is not ubiquitous [22]. For instance, brands that need to maintain that brand-image as opposed to provide a service will follow inherently different approaches to achieve the same design principles: Nike website vs. google homepage, Nike needs to create a Nike experience and thus must have a multi-media rich homepage as opposed to the Google homepage which serves the one great purpose of helping users search the web for whatever they are seeking.
4.1. iOS Human Interface Guidelines: An overview of design principles in currency [23]

Aesthetic Integrity is critical to creating an immersive user experience according to Apple, this is nevertheless driving in large part by the nature and expected outcomes from the use of the interface and the application in question itself. Aesthetics are delivered through simplicity according to Apple, by assisting the users accomplish their objectives through strong yet subtle (even subliminal) modelled behaviours.

Consistency is key to allowing users to easily transfer their knowledge in the head with knowledge in the world [24] [19], from task to task, from application to application. Decades of research have allowed Apple to generate key insights into what makes users tick, and it appears consistency in product and interface design allows users to think minimally, while acting maximally. This is particularly interesting as it amalgamates multiple theories of interface consistency including those of Norman [25-26], and Grudin [27]. It appears that Apple, knowingly or unknowingly (can't say deterministically as the iOS Guidelines mention this explicitly), consolidate multiple points of view about the subject matter, in a contextual and tailored manner.

Direct Manipulation allows users to interact with the interface, and receive immediate and visible results. Being able to control objects allows users to feel more engaged with the task, and understand the impact of their actions on the interface [28]. This is to say that being able to translate, rotate, or scale central axes or objects of the interface empower the user, and enable them to modify the interface in accordance with their mental models. Adding subtle tactile, sonic or visual feedbacks in conjunction with direct manipulation solidifies the conceptual model of the interface in the users’ mental models.

Feedback help clarify the result of the actions that the users take, which helps them in understanding the progress of their tasks, and acknowledges their intent. Feedback about the result of an action must be immediate, latency has been demonstrated to irk users [29]. Moreover, such immediacy in feedback allows users to develop the cause and effect relations about the interaction with the interface, which has been a proven principle in underpinning the interface designs’ mental models in the knowledge in the heads of the users.

Metaphors allow users to draw bridges between the knowledge in their heads with their knowledge of the world by building upon familiar experiences, thereby allowing users to quickly assimilate how to use the interface. The most commonly known metaphor in computer science, is the desktop metaphor: the user can find files, folders, documents, trash cans etc on their desk. In the context of a web interface, this may translate in a number of ways: for instance, a platform that is intended to collate multiple activities can draw from the UX of a gaming station, allowing users to have a sense of familiarity.

User Control puts users in control of the interface, as opposed to getting the interface to drive the uses’
actions. Interfaces should be intuitive enough to alert users should their action have potentially undesirable consequences, but this should not limit the ability to take decisions away from the user. A good interface uses a blend of affordances, signifiers and forcing functions to make the interface behaviour and control familiar and predictable -- in other words, simple. Apple’s research indicates that users do not like an on-boarding experiences, and instead prefer to play around with the interface and learn by trial and error. This is similar to what Norman [30] says about memory about interface features and functionality being a resultant of repetition of interaction, as opposed to briefings in the form of help etc.

4.2. Materialising design concepts and principles into final products

The Apple developers guidelines indicate that the interfaces of applications running on its operating systems benefit from drawing on the external systems view, thereby minimising the cognitive loading that the users must experience in order to accomplish tasks. This builds on Norman’s take on the three aspects of mental models: design model; user’s model; system image [31]. The user’s aspect is, nevertheless, the most critical one in that it is at the heart of user-centred design, and is the driving force for the interface design and analysis.

In order to achieve such a vision-centric development approach, it is useful to have a mission or product definition statement to which the developers can reflect back upon for reference. Potential features, functionalities, and interface behaviours can thus be made coherent, and thereby simpler, by having a robust and clear central declaration of desired intention.

There are multiple ways in which a vision-centric interface design can be achieved, and progressively ICT as an industry has begun to incorporate and internalise methodologies in industrial design from other industries and disciplines. A good example of a such systems oriented engineering product development methodology [32-33] is Axiomatic Product Development Lifecycle (APDL) [34], which calls for a transdisciplinary [35-37] approach to manage the input and forcing functions for the system itself while taking into account the feedback loops using tools such as causal relationships or influence factors. The key idea is to separate concerns [36-38], while working with an underlying objective of a holistic systems perspective -- a system of systems, if you will.

The fuzziness of the early stages of the product development can be mitigated to some extent by the use of a formal method of retaining the definition-centric approach, such as APDL or one of the very many other methodologies of demonstrated success in assisting developers achieve a user-centric interface. Rapid iteration has seen heavy adoption in the software and interface development business in recent years [39]. Process and lifecycle management methodologies such as Agile, Scrum, Kanban, ScrumBan etc [40] speedfast market release through rapid prototyping, test-driven development and active stakeholder management, all of which has a proven positive impact on UX of interfaces.

There is much to be gained from the use of such process and project management methodologies, especially in the case of low-risk applications where errors or faults do not necessarily pose a serious direct or indirect to users, or their interests. This is not to say that such methodologies do not have a place
in complex systems environments [41], the question in such cases become how can such methodologies be used contextually, and whether or not the use or lack thereof of such methodologies actually benefit the development of the product and the interface. The CMMI Institute’s Guide to Scrum and CMMI: Improving Agile Performance with CMMI [42] is a great example of how comprehensive and complex process and project management paradigms fit in seamlessly with rapid-prototyping and early delivery paradigms.

4.3. Heuristics for Analysis

4.3.1 Usability Criteria: Dialogue Principles (1996)

Suitability for the task is concerned with determining that the functionality of the interface components and the interface as a whole are appropriate, that is, the interface allows the users to meet their needs while trying to carry out tasks. This is of particular importance in helping developers stay on track of a user-centric design approach, as opposed to a feature-driven or a technology-driven approach -- it may help bridging the gulfs of execution and expectation in the conceptual models of the developers.

User Control demands a check on the ability of the user to have explicit control over the interface, as much as possible. This is a critical property of any software, or control system as it is pivotal the stabilization of unstable systems, or systems of sub-optimal performance. In an attempt to achieve optimal control, and balance feedback loops, the interface gains better performance in terms of observability and controllability, thus making for a simple user experience.

Flexibility of an interface can be deemed optimal by the extent to which is allows itself to be calibrated by an individual experience. A container-esk or boxing approach to develop interfaces normalises nuances that are expected of individual users. Flexibility in the manner in which information is presented to, and digested by the users allows the interface to adapt to the individual user -- this tracks back to the golden rule of design thinking, make an interface that dances around users, not one that makes the users dance around it.

Error Management concerns itself with minimising and mitigating the possibility of user errors by making a lot of design and engineering magic happen under the hood. This is to say that error tolerance, prevention, and correction should ideally occur in the system view, and not the user view. It is evident in most good UIs, and especially in the iOS Human Interface Guidelines, that good design has in-built facility to finding and managing errors, and does not require user to be concerned with it.

Compatibility deals with hidden and perceivable affordances, signifiers, and forcing functions. The signifiers in the perceived affordances drive users, in conjunction with the forcing function, to determine and perform tasks. This can typically be achieved by conformance to commonly acknowledged conventions, expectations, and standards -- collating natural, intuitive, or expected interface behaviours.

Self-descriptiveness is one of the most fundamental dialogue principles in human-computer interaction.
and is concerned with feedback loops of information to and from the user, and support of activities and user guidance. The key idea is to enable users to understand the system image with minimal effort, this may perhaps be easiest to achieve by narrowing the gap between the system and the user view.

*User Workload* has a demonstrated correlation to task performance[43]–[45] which is can be caused minimal or excessive cognitive function loading [46]. Determining such workloads is typically a nontrivial task, and is generally characterised by the following properties [43]: diagnostic capability; selectivity; low intrusiveness; reliability; implementation ease.

Consistency is also a factor as described in ISO 9241:1996, but this heuristics analysis is not incorporating consistency a usability criteria as good interface design directs focus on users and the tasks that they must perform, consistency -- in the way it has been traditionally implemented in interface design -- shifts this focus away towards properties and behaviours external to the user-centered design. Traditionally definitions of consistency create a conformity bias in developers that has a strong correlation with developers introducing systems behaviours that are normative, without context. Grudin [27] and Nielsen [8] have made similar arguments. Proper system, or interface behaviour, is a resultant of independent judgement exercised by users -- speaking in terms of an analogy, putting the user in the driver’s seat, without the system behaviour being the backseat driver. It is for this reason, that consistency is not included as a usability criteria in this heuristics analysis, and interface design as a discipline may benefit from moving from such broad definitions of *consistency*.

### 4.3.2. Evaluation Criteria and Sub-Criteria

#### 4.3.2.1. Analysis of Benchmarks and Standards

The ISO 9241-11: 1998 *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability* [47] details the measures and metrics for use the a product by users; these can be described as a stroke of three broad parameters: effectiveness; efficiency; satisfaction. These principles and measures span across a varying range of usability parameters such as ease of control, recovery and resistance to errors and faults, bridging expectations, ease of understanding functions and functionality etc.

4.3.2.1.1. The ISO/IEC FDIS 9126-1:2001 [48]

The International Standards Organisation (ISO), and the International Electronical Commission (IEC), presented a Software Engineering Product Quality standard that in Part 1 sets the standards for the Quality Model expected from software systems and interfaces.

The Quality Model categorise the quality of a software or a software system as a function of it's functionality, reliability, usability, efficiency, and maintainability.

- Functionality of the product is a measure of its suitability, accuracy, interoperability, security and functional compliance to the project scope and specifications.
- Reliability of the system is determined by its ability to perform with maturity, fault tolerance, recoverability, and reliability compliance over a period of time stated relevant by the specifications of the product or project.
- The usability of such systems is driven by their understandability, learnability, operability, aesthetics and attractiveness, and compliance to usability guidelines of the product or project specifications.
- The efficiency of the system is product of its time behaviour, resource utilisation, and adherence to efficiency mandates according to the system specifications.
- The maintainability of the system is a factor of it's analysability, changeability, stability, testability and maintainability compliance.
- The system’s portability define if the software or software system can be transferred from one operational environment to another by measuring its adaptability, installationability, coexistence, and replaceability.

Some of these factors were replaced, reclassified, and reframed in ISO/IEC 25010:2011, the Software Product Quality Requirements and Evaluation standard -- also referred to as SQuaRE. The methods in which adherence to these standards drives development is beyond the scope of this report. For the intent and purposes of achieving the primary and secondary purposes of this research project, adequate consideration was effectuated to refer to the relevant guidelines described by both the ISO/IEC 9126-1:2001, and the subsequent changes introduced in SQuaRE.

The primary purpose of the study of these, and other benchmarks and standards was to explore and discovery the expectations of and from software and software systems in the context of industrial design and activities. This was achieved through close scrutiny of ISO/IEC 9126-1:2001, and close analysis of the nuanced differences between the definition and classification of the terms and guidelines in SQuaRE rendered the insight that the design and usability principles in both are essentially unchanged, they have simply been framed in a strategic manner to help with industrial design and production engineering.

4.3.2.2. Criterion Used as Metrics

Close readings of ISO 9241-11, ISO/IED 9216-1:2011, and SQuaRE, in addition to the limitation of scope of this research project (time), and discussions with Chris Chow helped determine that the best course of action to benchmark the evaluation criteria for this research project are: Effectiveness; Efficiency; and Satisfaction.

4.3.2.2.1. Sub-Criterion that comprise Metrics

Effectiveness is determined as a sum of two factors: Successful completion of tasks; and, Accuracy of completed tasks.

Efficiency is determined at the goodness of the Learning Time required to be able to use the system in a reasonably good manner.
Satisfaction is determined by the Overall User Experience Satisfaction.

4.3.2.2. Rating Methodology

4.3.2.2.1. Quantitative

The evaluation criteria are measured on a scale of 0 to 5. The meaning of different scores is described as follows:

<table>
<thead>
<tr>
<th>Score</th>
<th>Word Description</th>
<th>Emotive Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-0.9</td>
<td>Abysmal</td>
<td>Cannot deal, this makes me want to cry. Please change this ASAP</td>
</tr>
<tr>
<td>1-1.9</td>
<td>Awful</td>
<td>Who designed this?! Why is this such a bad design?!</td>
</tr>
<tr>
<td>2-2.9</td>
<td>Average</td>
<td>Don't really care for it, pretty ambivalent to be honest</td>
</tr>
<tr>
<td>3-3.9</td>
<td>Good</td>
<td>It is functional, not pretty, but does the job</td>
</tr>
<tr>
<td>4-4.9</td>
<td>Awesome</td>
<td>Almost perfect, could maybe make it a bit better</td>
</tr>
<tr>
<td>5</td>
<td>Perfection</td>
<td>Couldn't do this any better</td>
</tr>
</tbody>
</table>

4.3.2.2.2. Qualitative

The subjective feedback from the expert is the most valuable product of an expert heuristics analysis. Heuristics analysis are typically performed in the early stages of usability testing, to gauge the direction in which the product or project development is headed, and if or not come critical or important changes need to be made. This can be very useful especially in typical small scale, low risk applications that follow Agile or other similar development methodologies.

Numbers are usually a good indicator of a metric or measure, but they don't speak volumes without context. The ability to derive direct feedback in the form of a qualitative analysis of the interface with respect to the usability heuristic therefore, enables the developers to truly understand the problem and the problem space with regards to the interface and the interface design. This can help them derive concrete actions, changed, and steps that they can undertake in order to make the interface or interface design more usable.

The usability heuristics have been purposefully been framed in a way that they communicate certain ideas and help the usability tester develop a certain mental model when considering the specific element of the interface that they are being asked to evaluate. The intention is for the usability tester to, without having a
comprehensive knowledge of usability principles or being a subject matter expert, be able to produce insights that allow the developers to infer the goodness or lack thereof, of their core design principles.

4.4. Application Framework: A contextual update of existing influential heuristics

The heuristics analysis framework draws from principles and accepted norms of existing heuristics frameworks, and theories engendering usability engineering and human-centric design, and extends them to allow the usability of a broad spectrum of user interfaces -- software and computer hardware -- to be evaluated by an expert usability testing professional.

4.4.1. Bringing together worlds of knowledge

The key idea behind the creation was to develop a heuristics analysis framework for the Human-Centric Computing (HCC) workshop developed at the Australian National University (ANU) by Chris Chow and his colleagues. However, in the process of creation of this framework, it became evident that there are a few gaps in the knowledge that has been applied in the context of heuristics analysis in the field of user-centric design and evaluation. The heuristics analysis framework presented in this report is an attempt to narrow most of these gaps in knowledge, and bridge some fundamental gaps which had not seen much update since the original work conducted on by Nielsen in the 1990s.

The primary objective was to analyse the usability of the HCC workshop created at the ANU. In the process of developing a workshop-specific heuristics analysis framework, it became evident that creation of a framework will assist in primary usability testing of future products and systems created at the ANU, as developers and researchers have to spend time on developing a product or system specific usability test manual and framework, which adds to their workload. The use of a good framework that encoppasses usability principles and testing principles in a way that is at the same time broad enough to be applied to any product or system, and simultaneously allows the tester to be able to gain and deliver product or system specific insights without having to develop a dedicated framework would, ideally, save time and effort on such future exercises.

In order to make this heuristics framework broad enough in it's ability to alloy the tester to apply a wide range of usability, design, and systems principles, extensive background analysis was done to identify existing frameworks, and the principles that were common to all of them. Evaluation of a wide variety of expert heuristics analysis frameworks rendered the insight that they are largely based on the 10 principles of usability as described by Nielsen. The key difference in the structure and depth to which these frameworks went was largely determined by the scope of the project to which they were applied, and the nature of feedback that was expected of them.

For instance, for a mobile interface specific interface, the heuristics were fundamentally drawn from Nielsen's heuristics, and articulated in a manner that was specific to the product for which the usability test was being developed. In another similar instance, the heuristics analysis framework developed for a
web based application for a large consultancy was developed with the organisation’s design principles and brand image communication in mind. Dedicated effort on the development of individual frameworks can be reasonably justified in the context of large scale projects, where the scope of available resources -- time, cost, effort, etc -- allows for dedicated effort to yield a justified return on investment.

However, for projects typically conducted at university’s, especially as part of software systems, or components developed as part of university coursework, honours, or even small-scale research projects, the expense of such time and effort might be somewhat superfluous. The expert heuristics analysis framework developed as part of this research project aims to assist in saving this time and effort which could be exploited for other activities, especially given the limited bandwidth at which researchers and developers typically function in the context of such projects.

4.4.2. Developing A Score To Determine Usability

The framework also attempts to help develop a mixed approach -- qualitative and quantitative -- to produce a score as a result of the expert heuristics analysis. Heuristics analyses typically have a high degree of subjectivity involved in their conduct, and the results are usually not meaningfully relevant to an individual who can't be deemed to be an expert, either in relevance to the interface itself, or ni usability testing.

The ability to develop a score, therefore, allows for a relatively benchmarked analysis to be performed. The scores for different systems and interfaces would naturally in terms of their performance, and other metrics, the usability, nevertheless, should ideally be high. This is to say, that a system, no matter how complex or simple in it's systems image, should appear good to use to the user. Determination of a score, therefore, allows the usability of different systems to be contextually gauged, and help determine decision making parameters, and benchmarks.

4.4.3. Future Improvements To The Score

Determination of the performance of a system is still driven, in part, by a subjective determination of the score allocated by the expert. However, this can still be reasonably considered to be an improvement on the existing frameworks, that is, none. It can be reasonably inferred that there is perhaps still room for improvement in minimizing the amount of subjectivity to which this heuristics analysis is prone, and future work on this framework -- and the design principles upon it is based -- will aim to produces a more rigorous product, inferences from which could be even more objectively determined.

Some preliminary work was done to achieve this level of objectivity, however it became evident from discussions with some usability testing experts that the level of granularity that was being measured in the said heuristics framework was extensively time and effort intensive, which defeats the point of a heuristics analysis a bit: it should be easy and cheap to use, and should not require a lot of effort; it is supposed to be a preliminary test that the expert does to determine if or not the product is headed in the right direction. Therefore, development of the said framework was put on the backburner for future
evaluation because of time and scope constraints of this research projects. Regardless, the aforementioned framework will be revisited in the future to minimise the cognitive workload on the tester, and achieve a succinct methodology to achieve the desired level of breadth in usability principles, and depth of objectivity, to create a more rigorous methodology to allocate score to the usability of a system or an interface.

5. Implementation of the Heuristics Analysis

5.1. Method

The usability analysis was performed by the author of this report, and 3 expert users. The said expert users have backgrounds in usability testing and user experience design.

The users were given a brief overview of this research project, it's intention, and scope. The users were walked through the framework for a period of fifteen minutes. The concepts of the usability heuristics were discussed with them, and what the terms (of the evaluation criteria) effectiveness, efficiency and satisfaction implied. It was also explained to the evaluators that the differentiating factor between a typical user test, and this expert heuristics analysis was that the evaluators and the experimenter were free to communicate during the test, in matters of explanation of the interface. The evaluators were also informed that they were encouraged to think-aloud, and verbalise their thoughts as it helps the experimenter derive expert feedback in real time.

The testers were presented with the HCC workshop on [hcc-workshop.anu.edu.au](http://hcc-workshop.anu.edu.au)\(^1\), and use it for the first 15 minutes. They were then asked to evaluate the interface at the end of 15 minutes by writing their thoughts in the feedback section of the framework, and allocating scores as described in the aforementioned table against the usability criteria.

They were given one hour to carry out this exercise. They were also informed that the experimenter will try and avoid communication unless they were in absoluter trouble, and found the system completely unusable. The evaluators were advised to self pace, and use the HCC workshop in whatever order, and manner they saw fit. They were purposefully not instructed about the intention of the HCC workshop with the aim of discovering if the workshop was able to deliver on its core design principles, and was self-explanatory.

5.2. Alternative Frameworks Used To Draw Comparative Results

The expert heuristics framework developed as a result of this research project was implemented, in addition to 2 additional heuristics analyses, to draw a comparative assessment of findings, and contextualise the key findings in reference to both the HCC workshop, and the nuanced differences between the 3 frameworks used. It should be noted that, given the limited time for which the expert

\(^1\) This will only be accessible on ANU machines, or through ANU-Secure
evaluators were available, it was not possible to have them run 3 individual heuristics analysis -- the evaluators only performed the framework created as part of this research project.

The key differentiating factor between the frameworks studied, and the framework proposed as part of this research project is the inclusion of two layers of evaluation criterion, in addition to the nature of formulation of the usability heuristics. One, the usability heuristics have been drafted as a result of insights generated from data-driven work done in the past. Two, these usability heuristics have been updated with contextual and contemporary human interface guidelines as recommended by the forerunners in the field of UX design -- Apple Inc. Furthermore, each usability heuristic has been designed to comply with ISO/IEC standards and their definition of the nature and scope of usability and ergonomics. Moreover, the evaluation criterion of these usability heuristics and criterion are measured against benchmarked scores on a likert scales, as opposed to arbitrary measures of goodness observed in the frameworks studied.

The analysis, construction, and methodological structure of this framework is intended to ensure that the feedback generated as a result of implementation of framework is based by design principles, and has a higher degree of accountability and traceability to benchmarks, standards, and industry-grade practices.

5.3. Implementation on the HCC Workshop

5.3.1. Observations and Score from Expert #1

Expert #1 was a behavioral insights specialist, with extensive background in applied psychology, computer science, human centered computing and behavioural economics. The extensive background of expert #1 was useful in rapid and easy identification of usability issues with the interface, and interface design with precision. It was apparent from the verbalisations, that the subject matter expertise had deep foundations, almost second nature, to the expert, which allowed them to easily provide comprehensive feedback.

Expert #1 found the HCC workshop to be functional, and quite usable. It makes good utilisation of design principles studied as fundamental in HCI. The interface is simple, and tile based -- perhaps mimicking a windows interface to draw on knowledge in the head of the users. There are a few minor issues with the interface behaviour when subjected to change in browser window size, these have been detailed in the heuristics framework.

In essence, the workshop is usable, and accomplishes its primary objectives of initiating conversation, and being educative for students as part of various educational events and activities at the ANU.

5.3.2. Observations and Score from Expert #2

Expert #2 was a professional UX designer for a design consultancy, and specialises in gamification and UX design. The expert was firstly allowed to interact with the HCC-workshop for fifteen minutes, during
which they were asked to think aloud, which precipitated a lot of feedback about aspects of the user interface. Given the expert’s background, the key observations that they made with regards to the interface involved its interactivity with the user, and aspects of gamification.

Overall, the experts found the interface to be highly usable, with the lack of consistency of fonts being one of the only major points of concern. Specifically, the style formats differed from the landing page to the workshop pages, especially when used in different browsers -- this might need closer investigation. Another major suggestion made was the use of more iconography and illustrations to minimise textual content so as to increase the likelihood of the user assimilating the content, as opposed to being dissuaded by the longevity of certain excerpts in certain workshop activities.

Addition of badges upon level up was also a key point raised during the think aloud. According to the expert, the onboarding experience -- in a very prominent manner -- exclaims that the new user is at level-0, ‘baby’, but this does not communicate how long the user journey would be other than mentioning that there are seventeen levels. The expert recommended that a progress bar, in conjunction to the addition of iconography corresponding to the different levels -- baby, primary schooler, undergraduate, professor etc -- would help communicate intent and purpose of levels and purposes of progressing to workshops in a simpler, and clearer manner.

The expert was asked to use the heuristics framework to help us draw feedback from their experience, and to not exceed 45 minutes to do so. The expert allocated scores, and filled in the feedback sections sparingly -- that is, did not fully complete it. This can perhaps be inferred to imply that the framework might require further simplification to allow timely completion of the feedback process, as cross-referencing the heuristics with the evaluation criteria seems to be time-intensive from the expert’s feedback about the heuristics framework.

5.3.3. Observations and Score from Expert #3

Expert #3 has a background in software engineering and business analysis. Following a the research protocol, the expert’s think aloud unravelled an experience that focused on the subtleties of the web design of the HCC Workshop, with verbalisation of the underlying technologies and how the expert envisioned them working behind the user interface. This was interesting in that it illustrated that expert users perceive their experience using a given interface from the lens of their background.

Consequently, the expert began to explore different pathways and verbalising the user journeys that a non-sophisticated user would experience, and potential issues around the user experience design around the HCI Activity -- in particular the Mapping and Conceptual model activities. Interestingly, the expert user’s verbalisations illustrated the nature of discount usability tests, and how a subjective analysis -- that draws from subject matter expertise -- can aid in unravelling large amount of feedback from a relatively short input of effort.

In the exit interview, the expert gave feedback on the nature of the Framework, and that it needs further
simplification in its structure as their experience following the instruction set and recounting their own user experience to evaluate the HCC Workshop relied on recounting from memory. It can be reasonably be inferred that drawing from memory may have drawbacks in terms of the cognitive workload that may be required to provide feedback.

5.4. Key Findings

5.4.1. Future Studies

Nielsen and Landauer [49] have indicated that there is proportional relationship between the number of evaluators who conduct usability tests, and the number of usability issues discovered. The number of usability issues discovered increases sharply with over five evaluators performing the usability tests. To derive more detailed insights, future studies will attempt to have a larger number of evaluators perform the expert heuristics analysis.

The extra issues discovered from larger sample sizes converge quickly, unless the sample size reaches 40 participants. Having 5-10 participants is determined to be optimal for heuristics analysis, beyond which the cost-opportunity goes down significantly unless the sample size is increased significantly.
6. Think Alouds

6.1. Overview

Think Alouds are one of the most simple, flexible, easy to conduct and cheap usability testing methods [6-7], and at the same time one of the most robust methods in drawing large amounts of contextual user experience and usability testing insights. Nielsen and Norman recommend Think Alouds as the primary go to usability testing methods in a usability engineer or user experience designer’s toolkit [50]. In industry, user experience researcher and usability engineers use Think Alouds to create powerful user research studies in a variety of user-centric design areas ranging from interface and interaction design, to user centric service design.

The nature of Think Alouds allows the surfacing on underlying assumptions in a user’s experience through surgical nudges from the usability evaluator. With a relatively low inertial effort to train themselves in conducting Think Alouds -- and supporting survey frameworks [51] -- evaluators can create holistic pictures about the experience and product design layers that drive the user’s experience with the product (or service.)

Usability design and guidelines are driven by human behaviour, and have a long shelf life [52] because of the nature of human-computer and human-system interactions. Human behaviour is largely driven by similar motivations, which evolve gradually and iteratively with respect to iterative and incremental improvements and changes in the nature of technologies, interfaces, and products that are being used. Nielsen argued that usability methods have a shelf life even longer than that of usability guidelines.

The reason, as described by Nielsen and Norman is the rate at which human behaviour changes with respect to changing technologies: that behavioural changes are slower than the technological changes. It is for this reason that Think Alouds have come to become time tested discount usability testing methods with an industry wide adoption know to deliver high value results at minimal cost offsets.

As a general principle, Think Alouds don't require special equipments, allow the evaluator to explore details of interaction and behaviours, are quite flexible in that they can be used for alpha and beta products as well as for low fidelity prototypes, allow the demonstration of users’ experience to developers and engineers -- which is usually an eye opening experience that helps illustrate users’ mental models and experiences very overtly, and are not extensively rigorous to train to conduct.

Think Alouds do not, however, don't precipitate much statistical data unless large and expensive studies are undertaken.

Nevertheless, Think Alouds have proven to be a demonstrated usability testing method that help unravel significant feedback with relatively low initial investment of resources.
6.2. Method

This experiment is aimed at generating insights about the usability and user-experience of a web-based interface through a retrospective think aloud. The number of verbalisation problems that were identified through traditional Retrospective Think Alouds (TA).

The test object used in this experiment was the HCC Workshop developed at the Australian National University, Human Centered Computing Group, by Christopher Chow and colleagues. Eight Think Alouds studies were undertaken. Four user-pathways were tested as part of the tasks expected of the users to complete for this experiment: Create an account (register)/login; Use Eye Tracking workshops and complete activities; Use electrocardiography activity; Use neural network activity.

The subjects were emailed the participant information sheet a day ahead of the study. The study began with a brief overview of the study, coupled with 5 minutes allocated for the test subjects to go over the participant information sheet. This was followed by the presentation of screener questionnaire, and an optional demographic information questionnaire.

The participants were briefed on Think Aloud and how the experimenter would ‘task’ the subjects during the study.

The experimenter followed a structured framework to document the user experience during the Think Aloud using the following two templates:

1. Surveying Document - Think Aloud: This document was based off of a metacognitive teaching framework [53] [54] [55] originally intended to assist educators understand the nature of cognitive processes that students undergo when ‘thinking out loud’ in a structured and methodological manner. Based upon principles of learning theory and metacognitive-drivers of Think Aloud, a framework with a tally was developed to identify user behavioral interactions and experience as part of the usability study. The experimenter tallied and categorised verbalisations during the course of the study.

2. Surveying Document - Task Completion: This document allowed the experimenter to document task completion, and document observations during the course of the study.

The experiment was designed for a 40 minutes run-time, 10 minute of onboarding of users and calibration of experimental settings, and 10 minutes for post-study debrief and exit interview.

The exit interview involved a 5 minute debrief by the experimenter with the subject, followed by the exit interview questionnaire that was created to draw user feedback about the experience of participating in the study, and the nature of their experience in finding usability issues as a result of thinking out loud.
6.3. Analysis

6.3.1. Participants Selection and Representativeness

This study was undertaken with 16 participants. Additionally a pilot study, and 2 dry run studies were conducted to develop experimental maturity and work out the goodness or lack thereof of the experimental procedure and its elements.

The results of the pilot study, and the 2 dry run have not been included in this analysis as they were intended to develop process and procedural maturity, and were not ‘clean’ enough to be considered ‘studies’. This decision was informed in part by the recommendation made by the Nielsen Norman Group's guide on ‘How to conduct usability studies for accessible web design’ [56].

The participants were recruited as a representative sample of intended user groups for the test object (HCC Workshop), ‘ANU students’, to address and reflect the specific user characteristics. The decision to recruit a representative sample of user groups for Think Aloud was informed by experimental design recommendations by the Nielsen Norman Group’s ‘234 Tips and Tricks for Recruiting Users as Participants in Usability Studies’ [57].

6.3.2. Metacognitive processes analyses

The following are the results that elucidate the cognitive and metacognitive strategies that users employed during the course of the Think Aloud:

<table>
<thead>
<tr>
<th>Strategy Think Aloud</th>
<th>% of Total Documented Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicting</td>
<td>7.8</td>
</tr>
<tr>
<td>Questioning</td>
<td>5.9</td>
</tr>
<tr>
<td>Visualising</td>
<td>6.6</td>
</tr>
<tr>
<td>Personal Response</td>
<td>15.6</td>
</tr>
<tr>
<td>Clarifying</td>
<td>18.3</td>
</tr>
<tr>
<td>Synthesis</td>
<td>16.6</td>
</tr>
<tr>
<td>Summarising</td>
<td>7.9</td>
</tr>
<tr>
<td>Reflecting</td>
<td>8.5</td>
</tr>
<tr>
<td>Making Connection</td>
<td>13.2</td>
</tr>
</tbody>
</table>
The average responses tallied per session is 73.7, this is to say that for each 1-hour session an average of about 74 responses were tallied by the experimenter using the strategic think aloud survey toolkit. The toolkit was designed to allow easy detection of the strategy used by eliciting a range of possible cues that indicative of specific strategies.

6.3.3. Task Completion

The following are the results that document the task completion by the participants of the study:

<table>
<thead>
<tr>
<th>Task Completion</th>
<th>% of Total Documented Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register to create a new account for yourself</td>
<td>100</td>
</tr>
<tr>
<td>Open up the HCC Workshop, and try to understand what it's about</td>
<td>100</td>
</tr>
<tr>
<td>Open up the Eye Tracking Activity, and find the meaning of saccades</td>
<td>100</td>
</tr>
<tr>
<td>Find out how you can use Eye Tracking as a controller</td>
<td>68.8</td>
</tr>
<tr>
<td>Find out the use of Eye Tracking in reading</td>
<td>56.3</td>
</tr>
<tr>
<td>Find out if law is a subject area associated with HCI</td>
<td>25</td>
</tr>
<tr>
<td>Find out what good mapping means</td>
<td>62.5</td>
</tr>
<tr>
<td>Find out how conceptual models help you tell the time</td>
<td>18.8</td>
</tr>
<tr>
<td>Logout of the HCC Workshop</td>
<td>100</td>
</tr>
</tbody>
</table>

The overall task completion 70.1% for the test user pathways.

6.3.4. Exit Interview Feedback

The following are the results from the exit interview questionnaire given to the participants of the study at the end of the experiment:
### Exit interview Questions

<table>
<thead>
<tr>
<th>Exit interview Questions</th>
<th>Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I found it easy to find issues with the interface</td>
<td>3.37</td>
</tr>
<tr>
<td>I found it easy to create an account</td>
<td>4.56</td>
</tr>
<tr>
<td>I liked the Neural Networks activity in the HCC Workshop</td>
<td>3.38</td>
</tr>
<tr>
<td>I liked the Human Computer Interaction activity in the HCC Workshop</td>
<td>3.25</td>
</tr>
<tr>
<td>Think Aloud was useful</td>
<td>3.68</td>
</tr>
<tr>
<td>I found it easy to logout of the HCC Workshop</td>
<td>5</td>
</tr>
</tbody>
</table>

Total average score (rounded off): 3.87

### 6.4. Insights

#### 6.4.1. Metacognitive processes analyses

The three leading strategies that seem to be used by the participants of this experiment are clarifying responses (18.3%), synthesis responses (16.5%), and personal responses (15.6%), while questioning responses (4.3%) appear to be the least observed.

These results reflect the thinking about thinking nature of a Think Aloud experiment. This is to say the nature of a think aloud requires a participant to verbalise their thought process while interacting with a given user interface, which requires the participant to think about their thoughts as they are thinking them in real time during the course of the study.

For instance, when a participant was interacting with the Human Centred Interaction activity in the HCC Workshop, and trying to *find out if law is a subject area associated with HCI* they verbalised associations that law as a discipline may have with cyber security, psychology and design. In doing so, the participant was drawing from “knowledge in their head” and trying to match their mental model with the conceptual models as they are presented in the HCC Workshop. It can therefore reasonably be inferred how clarification, synthesis and personal responses were leading metacognitive processes implemented by the participant.

Another example of illustration of this combination of metacognitive processes was when one of the users was attempting to *understand what the HCC Workshop is about*. Their verbalisations reflected that they were drawing from previous experience with websites to try and understand...
what the HCC workshop is about. This involved rapidly scrolling up and down, the verbalisation of their thought during this was, “this is how I browse Facebook to see what’s up”. Again, the personalisation of previous experiences and the nature of these past experiences collated to draw expectation from the given interface to match the knowledge in their head, to the knowledge in the world.

Similarly, the logout task has a 100% completion rate, and the verbalisations from multiple participants indicate the emergence of affordance, consistency and mapping as design principles that lead the participants to *synthesize* knowledge and past experiences, to *clarify* the rationales for making (or not making) a choice in a personalised manner. The logout button on the HCC Workshop is on the top right corner and uses an icon commonly used as a logout icon on multiple websites. The combination of the use and identification of the design principles by the users, through the aforementioned metacognitive processes is apparent.

A key insight that was realised during the course of the studies was the taxing nature of Think Aloud as a usability testing method for the participants. The exit interview debriefs, and general observational analysis of the participant during the studies revealed that thinking out aloud was found to be tiring and a bit frustrating by certain participants, and the general feedback around the nature of think aloud by the participants indicated that although it helped them discover issues with the interface’s usability more so than they would have had they just used it per usual (not thinking out loud), it was tiring to be conscious of their thoughts -- that the unnatural nature of thinking out loud made them feel tired.

This is not entirely surprising as it can be reasonably be observed that thinking out loud requires conscious verbalisation on cognitive process of which an individual would normally not be aware given the tendency of the mind to make leaps and bounds in reasoning to abstract and encapsulate the complexity of normal decision making processes [58] through the experiential development of heuristics, metaphors and ideologies [59].

The nature of qualitative analysis enables the furbishment of layers of complexity in a rapid, and relatively inexpensive manner. In addition to creating for the generation of rapid insights, context around the insight can be developed in-situ through a structured and methodological approach by the experimenter. There is scope for the creation of deep learning platforms [60] that may potentially be able to replicate such processes [61-62], but the the cost-opportunity of creating a such system would need to be addressed to justify the development of a such system.

### 6.4.2. Task Completion

The overall task completion for all of the participants was 70.13% which is higher than the average task completion rate for lab-based usability tests [63], and based on existing research [64] it can be
extrapolated that similar results would be achieved for up to 97% participants for a study with up to 40 participants [65].

It can be inferred from the relatively high percentage of task completion that the user interface is generally usable. This inference is along the same lines as the findings from the expert heuristics evaluations.

Specifically, however, it appears that the tasks of finding if law is a subject areas associated with HCI (25% task completion) and finding how conceptual models can help tell time (18.75% task completion) have a lower than desirable task completion percentage. In both cases, the tasks were designed to test if the user will be able to make second-degree inferences and be able to complete these tasks. So, the low rate of task completion can be reasonably understood given that second-degree inferences are not reasonably expected to be made.

To further explain this, law is now one of the icons under the subject area section in the HCI activity. The task to identify if law was one of the areas was intended to understand how users may try and use knowledge in their heads to meet this task. The observations from the think aloud verbalisations illustrated that the participants could easily make first-order deductions [66] with respect to the other subject areas in the section in question, but to then deduce if law was in fact associated the users followed one of two pathways: went back to the dashboard/home to find if they could find law elsewhere (mostly in the cyber security activity); asked the experimenter if saying that law is not a subject areas associated with law is an acceptable answer.

Further evaluation of this user behaviour may need to be undertaken to fully understand what such behavioural patterns and interactions indicate with substantive reasoning.

The task to determine the use of Eye Tracking in reading had a relatively low completion rate (56.25%). Observation of the user interaction leads to the inference that this can be explained due to the mouse hover over function in the Eye Tracking activities. The text description of these activities is revealed when the cursor hovers over the icons. The participants’ verbalisations leads to the inference that the users desired a present text description for these activities.

This is in contrast to the observation from the verbalisations around the law task. The icons under the subject area section are designed with mouse on-click and unravel the text on a mouse click. Part of the reason why the task had a low completion rate was because users did not click on the icons (it appeared as though that was not the expected behaviour), instead navigated away from the section because they could not locate law in the section. The few users that did successfully complete the task gave the feedback that the icons unraveling the text when they hovered over it was what they were expecting the icons to do. This difference in the users’ expectation of on-click in one area, and on-hover on another is something that requires deeper evaluation.

An additional insight into the users’ interaction with the HCC Workshop was their general curiosity towards the neural networks activity. The neural networks activity was purposefully left out of the
usability test tasks, to observe the participants’ natural affinity towards it -- this was based on preliminary observations from STEM outreach workshops that the experimenter was a part of at an earlier stage, where the participants tended towards the neural networks activity. The general feedback from the verbalisations as part of this study reflected that the participants found ‘neural networks’ interesting/curious. Further evaluation may be required to fully understand the dimensions of this observation.

6.4.3. Exit Interview Feedback

Creating a new account and logging out were scored well by the participants of the experiment, which is indicative of the relative ease with which both these tasks could be performed with the HCC Workshop interface. Verbalisations from the user with respect to the interface design of the login screen indicated a preference for the colour of the screen and the background image being simple yet informative -- especially the subtle animation of the background image appeared to communicate a common message for all participants that can be roughly described as the HCC Workshop having to do something with eyes and human behaviour.

Furthermore, the feedback about the interface design illustrated well executive of signified, affordances and consistency as design principles on the HCC Workshop interface. This is to say that the layout, icons, and the placement/interaction of buttons allowed users to quickly draw from knowledge in their heads to match the system image.

The initial hypothesis of participants having a tendency towards the neural network was the reason why the neural networks activity related question was present in the exit questionnaire but not included as a task in the usability test itself. With the exception of 2 participants, all other participants interacted with the neural networks activity as part of their task #2 in the study, or during the course of the study. The activity scored at an average of 3.38 (on the Likert). The user feedback from the verbalisations indicated the reason for a tendency towards the activity being it seeming interesting/curious. The reason for this could not be unravelled -- perhaps future studies could implement a level-3 verbalisation inspection to further understand this [67].

Moreover, the question *think aloud was useful* rated 3.68 -- the participants were briefed to answer the questions with their experience of the think aloud in helping them discover issues with the user interface in mind. The score, and the feedback from the participants illustrated that verbalising their thoughts assisted the participants in discovering usability issues in an enhanced manner than internal thought processing (or natural use) would have rendered. This help support the initial hypothesis that think alouds allow an enhanced discovery of usability issues with ease, low resource input and inexpensive experimental settings.

The feedback about the nature of the think aloud, as experienced by the participants, illustrated that they felt more *conscious* of their thoughts while thinking them out loud, in addition to having to slow down their actions to take in all of the variables at play during the course of their interaction. This helps to infer that think aloud allow a strategic use of metacognitive and cognitive processing that allows the unravelling of multiple layers of user behaviour [64], [68] across different dimensions of complexity with
a mix of both self-regulating user behaviour and strategic prompting by the experimenter to facilitate assumption testing [69] and user experience analyses.

6.5. Comparison With Heuristics Analysis

Heuristics analysis is predicated on some existing expertise or knowledge in the subject matter of human computer interaction or usability engineering. As such, it can only be performed by expert users. Albeit, it is a discount usability testing method that is valuable in drawing early feedback during the initial stages of development, and help evaluate the holistic health of the usability and interaction of a given interface.

Think alouds on the other hand prefer a lack of subject matter expertise with human computer interaction or usability engineering, as it is intended to draw feedback from end or naive users. They help pinpoint inconsistencies in the interface design with an inexpensive resource input. However, they are unnatural to the users and unless administered in a consistent manner, can help draw feedback, but may leave the opportunity for the development of context around the observations and feedback.

A commonality in both these usability testing methods is the opportunity for a structured and methodological approach for non-repudiation and consistency of experimental design, and therefore of the results drawn as a consequence of the usability study. Moreover, given the nature of the design of these studies, useful feedback can be drawn in a quick and relatively inexpensive manner early on in the development lifecycle, therefore positively ensuring that the overall product design can be developed in the right manner.

The heuristics analysis allowed the experts to apply generally accepted design principles by drawing from industry standard practices to holistically evaluate the HCC workshop. With the framework that was provided to them, they were able to do so in a replicable manner that could be documented in a platform independent manner and evaluated against ISO standards for usability and ergonomics. However, the feedback from the experts rendered the insight that the framework needs further simplification in its expanse to meet the requirements necessary for it to be deemed a discount usability testing method. This will require further abstraction and encapsulation of the factors upon which the framework was based, but is beyond the scope of this project due to the limitation of time and resources.

Moreover, the think alouds allowed rapid generation of insights similar to those surfaced from the heuristics analysis, with slightly lower fidelity given the relative non-structure typical of a think aloud. This issue around non-structure was addressed with the use of appropriately designed survey toolkits for the experimenter, which allowed easy documentation of both task completion and analysis, and simultaneous tally of the metacognitive strategies that the users seemed to employ through the use of preprocessed heuristics.

It could also be argued that the comparison between think alouds and heuristics analysis may require more context to determine the function and objective for which either method is implemented or used. Nothing means anything without context. To, therefore, make an assertion that either method is wholly better than the other would be over simplification of the functional context. It must therefore be noted that
it is the objective and the desired outcomes from the usability test or study that determine which method is contextually relevant and more likely to yield desirable insights at the right stage of the product development.
7. Retrospective Think Alouds with Eye Tracking

7.1 Overview

A retrospective think aloud is similar to a traditional think aloud in that it drives verbalisations as a means to draw user feedback by the experimenter. The key difference is the division of the experimental protocol into two components.

The first is a traditional think aloud with verbalisations -- in addition to the participants’ eye movements being tracked, which does not change the nature of interaction with the interface but works as a simultaneous (and background process).

The second component is after the completion of the tasks designed for the study, the participants are asked to pause and reflect on their experience interacting with the interface, and verbalise it (think aloud). The fact that these verbalisations are post-hoc -- as opposed to in tandem -- is what makes this a retrospective think aloud.

A retrospective think aloud with eye tracking includes following this part of retrospective verbalisations by showing the participant a heatmap or an opacity map of their interactions with the interface, which would ideally be developed based on timing data from the session.

Showing the user this eye tracking data is supposed to theoretically boost the participant’s recollection of the nature of their interaction. This is to say, recall rather than recollect. Looking at the heat map would allow the participant to recall which specific part of the interface created a certain experience for them, and since they will be verbalising their thoughts, this would allow the experimenter to be able to prompt the participant to understand the nature of the user experience in a deeper capacity.

It must be noted though, that given the amount of resources (development time to write API to create heat maps -- unless a custom off the shelf solution is being used, and the effort associated with such an experimental design), it could be argued that a retrospective think aloud with eye gaze tracking ceases to be a discount usability testing method, and beings to lean on the relatively objective quantitative side of usability testing, also relatively more expensive.

7.2 Analysis

Three eye tracking studies were performed. However, schedule slippage on account of time required for ethics approval, and the development of technical skills required to create scripts and APIs required to convert eye tracking data into heat maps and opacity maps led to the data not being analysed.

The ethics approval delay appeared to be due to the use of the HCC Workshop as an outreach tool which
could involve participants under 18, which led to a number of cycles of questions from the committee, before approval was eventually granted.

This data has nevertheless been collected, and can be used for future studies for further evaluation. Given that this experiment was designed for a specific use, the data from the verbalisations was also left out to maintain experimental integrity. Analyses on these datasets can be revisited at a future opportunity.

8. Conclusions and Future Work

The heuristics analysis framework (HAF) was developed in part based upon human information processing theory, applied industrial ergonomics design, and industry standard practices in factorial interface design. The questionnaire items in the framework abstracted and encapsulated the theoretical elements that drive the design of the principles in question through a subject matter expertise driven approach of the researcher. Peer review of the framework would therefore be critical in drawing invaluable feedback from the human computer interaction and usability testing community at large to improve its quality. The intention of the HAF is to create a tool that can help benchmark the usability of a given interface. Feedback from the participants of the initial study were instrumental in evaluation of HAF’s usability. For resource purposes, further evaluation could not be performed as part of this project. However, future quantitative studies would help test HAF and make the adequate and necessary improvements to it.

Questioning is recognised to have a strong relationship to knowledge building and construction in metacognitive processes [70]. Therefore, the low incidence of questioning strategies among the participants of this study as a strategy for thinking aloud beckons further investigation. Perhaps the consideration of response-evaluate-understand models of metacognition [67] may help understand the nature of this behavioural pattern among users when thinking aloud.

Metacognition has been used extensively in educational and behavioral psychology to understand learning processes and knowledge construction. Given that think alouds are used extensively in usability engineering, and user experience research and design, understanding the deeper drivers of the underlying cognitive and metacognitive processes may help further the methods employed to understand user interactions and behaviours. During the studies as part of this project, it could be observed that the users were thinking about thinking, in addition to developing knowledge about knowledge part of their interaction with the HCC workshop. The question is therefore, what sorts of insights could a usability engineer unravel with this in mind? Further research in metacognition and cognitive processes that drive user behaviour and interaction with digital interfaces could help understand new and innovative layers to the complexities of these human-computer interaction.

Due to resource and skill constraint, the project objective of using eye tracking data to understand and evaluate the efficacy of retrospective think alouds with eye tracking data could not be fully achieved. Although experimental data has been collected to this end, analysis of this data is now beyond the scope of the project, and will have to be revisited at a future opportunity. Understanding how recall may assist (or not assist) in finding usability issues will help determine the extent to which discount usability tests
may need to be pushed to achieve desirable granularity in specific feedback from the user interactions.

An interesting observation from the expert heuristics analysis was around the nature of expertise, and the limitations to which this expertise can be replicated in, between and within studies. This is to say that although *experts* are trained with similar backgrounds, skills and knowledge, the nature in which they apply this expertise heavily influences how they test the usability of a given interface. This was apparent in the studies ran with experts as part of this project, each expert had a slightly different background, and the nature of their feedback reflected the biases that these factors had on how their observed the interface.

Furthermore, it may be possible -- through the use of deep learning -- to replicate this expertise through the creation of expert rules and heuristics with which a given machine learning platform can be trained. The nature, extent and feasibility of a such platform will require further analyses, but from a simplistic perspective the marginal cost and entropic risks associated with the development of a such platform, versus the benefit of long-term application in a standardised and indexed manner may make this a lucrative opportunity for further investment and research.

Lastly, a clear difference between the nature and application of mouse on-click and mouse hover over in user experience design was apparent during the think alouds. It might be worthwhile to investigate the contextual nuances that precipitated this difference to unravel the underlying design principles, and assumptions that lead to this observation.

9. Acknowledgements

I would like to thank my supervisors Tom Gedeon, and Chris Chow for giving me the opportunity to undertake this project, and for the incredible support and guidance that they have both given me over the course of this project.

I would also like to thank my colleagues, friends, and participants for the experiments that were run as part of this project for helping me achieve my aspirations with this project.
10. References


[67] M. Bannert and C. Mengelkamp, “Assessment of metacognitive skills by means of instruction to


11. Appendix 1

11.1 Project Description

The project is intended to generate insights about user experience testing for web-based applications. This was to be achieved through an initial analysis of the numerous methods and approaches used in industry and academia for user experience and usability testing, and discovery of the most germane methods that generate maximum output, with minimum input.

The intended approach was to use grounded-theory as a guiding principles and use a mix-methods approach combining quantitative and qualitative data to generate and report insights.

The specific expectations from this project are as followed:

1. Research and analysis of existing usability and user experience testing methods
2. Understand experimental design, and the ethics approval process in human-research
3. Design and execution of expert heuristics analysis
4. Design and execution of think aloud studies
5. Generate insights from the studies conducted for the test object

Create a technical report detailing methods, approaches, results and recommendations to improve the user experience for real world web applications.

\[\text{2} \text{ The Heuristics Analysis Framework that has been developed was not originally in scope. The importance of a structured framework for discount usability tests was realised during the research phase of the project, and as a result effort was allocated to creation of this gap in the existing usability testing methods.}\]

\[\text{3} \text{ The metacognitive approach used has not been done before. This approach is novel, and this project is the first instance of its implementation to the best knowledge of the author of this report.}\]
12. Appendix 2

12.1 Study Contract

INDEPENDENT STUDY CONTRACT

Note: Enrolment is subject to approval by the projects co-ordinator

SECTION A (Students and Supervisors)

Unit No.: U5660667
Surname: Mahajan
First Name: Manik
Project Supervisor (may be external): Tom Gedeon, Chris Chow
Course Supervisor (an ANU academic):

Course Code, Title and Unit: COMP8715, Computing Project

Semester: S2 Summer 2016 S1 2017

Project Title:
User Experience Testing for Web Applications

Learning Objectives:
- Identify modern user experience (UX) testing methodologies and approaches that are suitable for web applications
- Design and perform an expert heuristic analysis of a real-world web application against a formal set of design principles
- Design and perform a small scale user study to collect empirical data related to the UX of a real-world web application
- Analyse empirical data in order to inform improvements to UX
- Understand the ethical considerations involved in performing research with human subjects
- Write a technical UX report detailing the methodology, empirical results, and suggested improvements of a real-world web application

Project Description:
This project focuses on performing UX testing on the HCC Workshop, a real-world web application in use by ANU BECS for outreach programs. It aims to expand the knowledge of HCI and UX testing introduced in COMP9006/6390 and apply them in an advanced setting to a real-world web application.

During this project, the student will:
- Research UX testing methodologies and approaches that are utilized within the industry for investigating web applications, including a literature survey
- Design and perform an expert heuristic analysis of the HCC Workshop based on the formal set of design principles previously introduced in COMP9006/6390
- Design and perform a small scale user study to collect qualitative empirical data relating to the UX of the HCC Workshop
- If time permits, expand the user study to incorporate the collection of quantitative data through common UX testing mediums, including eye-gaze tracking or mouse click tracking
- Collate and analyse findings to highlight positive/negative UX aspects of the HCC Workshop
- Develop an own heuristic analysis framework for use on the HCC workshop
- Evaluation extended by monitoring of site metrics and user behaviour
- Explore the use of eye tracking, gaze fixations, heart rate, electrodermal activity, and mouse click rates to understand user behaviour and infer attitudes when using the workshop
- Write a report detailing the methodologies followed, the results and insights learned, and recommended improvements to the HCC Workshop to benefit UX

Research School of Computer Science

Form updated Jan-12
ASSESSMENT (as per course's project rules web page, with the differences noted below):

<table>
<thead>
<tr>
<th>Assessed project components</th>
<th>% of mark</th>
<th>Due date</th>
<th>Evaluated by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artefact: UX evaluation framework and results dataset</td>
<td>45%</td>
<td>May 2017</td>
<td>Prof Tom Gedeon</td>
</tr>
<tr>
<td>Report: Final overall research report</td>
<td>45%</td>
<td>May 2017</td>
<td>Dr Sabrina Caldwell</td>
</tr>
<tr>
<td>Oral Presentation: Presentation</td>
<td>10%</td>
<td>May 2017</td>
<td></td>
</tr>
</tbody>
</table>

MEETING DATES (IF KNOWN):
Weekly with Tom Gedeon and Chris Chow

STUDENT DECLARATION: I agree to fulfil the above defined contract:

_________________________________________   …18 Oct 2016…+…22 Feb 2017
Signature   Date

SECTION B (Supervisor):
I am willing to supervise and support this project. I have checked the student's academic record and believe this student can complete the project.

_________________________________________   …18 Oct 2016…+…22 Feb 2017
Signature   Date

REQUIRED DEPARTMENT RESOURCES:
Access to HCC Workshop tool (available)

SECTION C (Course coordinator approval)

_________________________________________   
Signature   Date

SECTION D (Projects coordinator approval)

_________________________________________   
Signature   Date
13. Appendix 3

13.1. Participant Information Sheet

**Participant Information Sheet**

**Researcher:**
The primary researcher for this experiment is Manik Mahajan, a postgraduate Masters of Computing student within RSCS at ANU. This research project is being undertaken for the partial fulfillment for the course COMP8755, Individual Computing Project, at the ANU College of Engineering and Computer Science. This research is being conducted under the supervision of Professor Tom Gedeon, and Mr. Christopher Chow of the RSCS, ANU.

Project Title: Improving user interfaces through heuristics analysis, think alouds, and eye gaze tracking

**General Outline of the Project:**
- **Description and Methodology:** This project intends to evaluate the effectiveness of different usability testing methods. The participants will be given a screener questionnaire to check compliance with acceptance criteria of this study. An optional demographic survey will be made available to help contextualise this research with sufficient detail. The participants will be using the HCC Workshop as the test object, and will be asked to complete certain tasks which will be detailed by the experimenter. The purpose of this study to observe how the participants interact with the test object, and the efficiency of the method of usability testing being used to do so. It must be noted that it is not the participant that is being tested, but the test object and the method of testing. Information will only be collected and analysed upon providing consent by generating a unique username. Information relating to your eye gaze will only be recorded if you choose to participate in eye gaze tracking on dedicated machines. An optional exit survey will be made available so that you can share your thoughts and opinions about the workshop and the activities.
- **Participants:** Eye gaze, questionnaire data and activity metrics will be recorded from participants. A minimum 10 participants are required for this experiment. Participants will be above the age of 18 and will be selected based on prior experience with the HCC Workshop, HCI, user experience or design.
- **Use of Data and Feedback:** The data collected will be used to draw conclusions about certain interaction techniques and the nature of the tasks. Any data collected, either raw or processed, may be used in a thesis and other research and publications. The data will be made unidentifiable so that no participant will be able to be identified from any data collected.

**Participant Involvement:**
- **Voluntary Participation & Withdrawal:** This usability experiment is completely voluntary. You may end the test session or ask for a break at any time. You may request that any or all data...
collected from you be destroyed. You have the right to completely withdraw from the experiment at any point with no explanation to the researcher. In this case your data and personal information will be destroyed in accordance with the ANU Code of Research Conduct. You can ask that your name be deleted from our contact list for future testing at any time. You can withdraw until data is prepared for publication.

- What does participation in the research request of you? The main purpose of the user study is to collect data to enable useful information to be gained on the interface, the interaction techniques, and tasks. The data will be collected as part of normal participation. We will provide an opportunity to complete a post-task questionnaire that may contain some questions of an identifying nature. You do not need to complete these or any of the other questions if you have any objections to them. Some tasks carried during the session will involve recording of eye gaze.

- Location and Duration: The study will take place in Computer Science and Information Technology (CSIT) Building 108 on the ANU Campus. The time needed to complete this user study will be about 60 minutes in one standalone session. This time will include an introduction to the tasks, setup, and completion of the tasks mentioned above.

- Risks: All care will be taken to make you as comfortable as possible, given the nature of the interaction tasks. Some physical discomfort such as eye and muscle strain may occur with some people including, in rare cases, motion sickness. You are free to request that your participation in the user study cease at any stage without explanation.

Confidentiality:

- Confidentiality: The data from the experiment will be made unidentifiable so that no participant will be able to be identified from any data collected. All results published will be in regards to the overall findings from the cohort of participants and not on an individual basis. Until that time, if you give your permission, your contact details will be retained for follow-up testing. The data may be used in follow-up research by researchers not listed on this form. All researchers that will gain access to the data collected in this research will be listed under the same human ethics protocol as the current researcher. Confidentiality will be maintained as far as the law allows.

Privacy Notice:
In collecting your personal information within this research, the ANU must comply with the Privacy Act 1988. The ANU Privacy Policy is available at https://policies.anu.edu.au/ppl/document/ANUP_010007 and it contains information about how a person can:

- Access or seek correction to their personal information;
- Complain about a breach of an Australian Privacy Principle by ANU, and how ANU will handle the complaint.

Data Storage:
- Where: The data from the research will be stored securely and will be made unidentifiable to
retain privacy of each participant. Access to the data is available only to the primary researchers.

- How long: In accordance with the ANU Code of Research Conduct all data collected for the research will be stored for a minimum 5 years from the date of last publication.

- Destruction of Data: After this 5 year period the data will be archived for follow-up research. The data will be kept in secure storage at the Research School of Computer Science, ANU.

Queries and Concerns:
- Contact Details for More Information: If you have any further requests for information or queries regarding the study participants should be directed to the primary investigator, Manik Mahajan:
  
  Email: u5860667@anu.edu.au

Ethics Committee Clearance:
The ethical aspects of this research have been approved by the ANU Human Research Ethics Committee (Protocol 2012/006). If you have any concerns or complaints about how this research has been conducted, please contact:

  Ethics Manager
  The ANU Human Research Ethics Committee
  The Australian National University
  Telephone: +61 2 6125 3427
  Email: Human.Ethics_Officer@anu.edu.au
13.2. Screener Questionnaire

Screener Questionnaire

1. Do you wear contacts or eyeglasses in order to read the computer screen?
   [ ] Yes CONTINUE
   [ ] No Skip to 3

2. Are your glasses for:
   [ ] Reading only CONTINUE
   [ ] Seeing distant objects only CONTINUE
   [ ] Both (Do you wear bifocals, trifocals, layered lenses, or regression lenses) TERMINATE

3. Can you read a computer screen and the Web without difficulty with or without your contacts and/or eyeglasses on (whichever applicable)?
   [ ] Yes CONTINUE
   [ ] No TERMINATE

4. Do you use a screen reader, screen magnifier or other assistive technology to use the computer and the Web?
   [ ] Yes TERMINATE
   [ ] No CONTINUE
13.3. Written Consent for Participants

WRITTEN CONSENT for Participants

Retrospective Think Alouds: Improving user interfaces through eye tracking and verbalisation

I have read and understood the Information Sheet you have given me about the research project.

I have had the following questions or concerns about the project addressed to my satisfaction:

...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................
...........................................................................................................................................................................

I acknowledge that I will not be attributed in any way within research outputs.

I agree to participate in the project. YES ☐ NO ☐

Signature:..........................................................
13.4. Demographic Information Questionnaire

Demographic Information Questionnaire

The purpose of this test is to evaluate how user-friendly HCC Workshop is. There are no "right" or "wrong" answers. Your ability or inability to accomplish requested tasks provides us with valuable information about how well Usability Tests are designed. (Please note, your information will not be sold or given to outside entities. It is for internal use only. Please circle the correct answer.)

1. Profession: ______________________________

2. Institution: ____________________

3. Would you consider yourself comfortable with using the internet (please encircle)?  Y  N

4. Do you normally use a Mac or a PC (please encircle)?

Mac  PC  Other (specify): ______________________________

5. Age Group: (please encircle)

11-20  21-30  31-40  41-50  51-60  over 60

6. Gender:  □  Female    □  Male

7. How often do you use the internet (please encircle)?

Daily  Weekly  Monthly  Occasionally  Never

8. On the following scale, please rate your expertise in Usability Testing:

Very Poor  Below Average  Average  Above Average  Excellent
13.5. Experimental Plan

1. Research Intent, Questions and Hypothesis
   1.1. This experiment is aimed at generating insights about the usability and user-experience of a web-based interface through a retrospective think aloud.
   1.2. The number of verbalisation problems that will be identified through traditional Retrospective Think Alouds (TA) and Retrospective Think Alouds with Eye Gaze Tracking (TE) by the participants will be evaluated by this experiment.
   1.3. **Question:** Do participants report more usability/user-experience issues when participating in a retrospective think aloud in which they are shown eye gaze maps from their interaction with the user interface, than when they are not?
   1.4. Further evaluations will help generate insights into:
      1.4.1. The number of usability/user-experience issues identified by the participants ($N_p$)
      1.4.2. The manner in which the usability/user-experience issues are identified:
         1.4.2.1. Through observations of the participants by the experimenter: $M_o$
         1.4.2.2. Through verbalisations by the participants of their experience: $M_v$
         1.4.2.3. Through a mixed approach (observations + verbalisation): $M_{ov}$
      1.4.3. The nature of usability/user-experience issues discovered:
         1.4.3.1. Interaction
         1.4.3.2. Content
         1.4.3.3. Navigation
         1.4.3.4. Design
   1.5. There is no formal hypothesis in this experiment at this stage, but is is expected that the participants are more likely to have a positive experience with regards to their participation in a Think Aloud (verbalising their thoughts) with the use of eye gaze based retrospective think aloud.

2. Overview
   2.1. The independent variable in this experiment is the Think Aloud, which was tested in two variations: Retrospective Think Aloud (Traditional); Retrospective Think Aloud with Eye Gaze Tracking.
   2.2. The dependent variable in this experiment would be: $N_p$, $M_o$, $M_v$, $M_{ov}$ and the nature of usability/user-experience issues identified.

3. Test Object
   3.1. The test object used in this experiment was the HCC Workshop developed at the Australian National University, Human Centered Computing Group, by Christopher Chow and colleagues.
   3.2. The HCC Workshop ([http://hcc-workshop.anu.edu.au](http://hcc-workshop.anu.edu.au)) is a collection of hands-on activities that demonstrate the research that the HCC Research group carries out at the Australian National University.

4. Participants
   4.1. This experiment will aim to recruit 10-20 participants. Half of the total number of participants will be parts of traditional TA, and half will be part of a TE.
   4.2. Participants will be recruited from the Australian National University.
   4.3. There will be no financial compensation or course-credit incentive offered to participants
to participate in this experiment.

5. User Pathways Tested
5.1. Four user-pathways will be tested as part of the tasks expected of the users to complete for this experiment:
   5.1.1. Create an account (register)/login
   5.1.2. Use Eye Tracking workshops and complete activities
   5.1.3. Use electrocardiography activity
   5.1.4. Use neural network activity
5.2. This experiment is expected to run for one-hour per participant during which the participants will be accompanied by an experimenter who will observe the participants’ interactions and take notes.

6. Questionnaire
6.1. The experience of the participants will be evaluated using a five point Likert scale (range: strongly agree - strongly disagree) at the end of the experiment.
6.2. The Questionnaire can be found in the Appendix.

7. Devices Used
7.1. Eye Tribe
7.2. Empatica E4

8. Experimental Protocol
8.1. The participants will be provided information about the experiment one day before.
8.2. The participants will be asked to sign the Participant Consent Form, and interviewed to collect demographic data.
8.3. The audio/video equipment (if being used) for recording the experiment, Eye Tribe, and Empatica E4 will be explained.
8.4. The HCC Workshop will be introduced.
8.5. The participants will be given brief instructions about the nature of the experiment.
8.6. Eye Tribe will be calibrated to match the individual participant.
8.7. Empatica E4 will be calibrated to match the individual participant.
8.8. The participants will be given a task list enumerating and outlining the Task Scenario (to test the target user pathways).
8.9. The experimenter will explain Think Aloud to the participants, and emphasize that it is not the participants being tested, but the goodness of the HCC Workshop.
8.10. The experimenter will facilitate interaction with the HCC Workshop, and refrain from leading the participants.
8.11. After the completion of each task, the experimenter will ask the participant to verbalise their experience, and take notes. (If this is a TA, the experimenter will skip to 8.11)
   8.11.1. For TEs:
   8.11.1.1. The experimenter will prompt the participant to undertake the next task on the list, until all the tasks have been completed.
   8.11.1.2. The experimenter will show the participant eye gaze maps from the participant’s interaction with the HCC Workshop (and the target tasks), and ask the participants to verbalise their thoughts about the nature of interaction.
8.12. The Empatica E4 and Eye Tribe will be removed.
8.13. The participants will be given the questionnaire to fill.
8.14. The participants will complete an exit interview.
13.6. Task/Activity Information Sheet

Task/Activity Information Sheet

Activity Description:
This experiment is aimed at generating insights about the usability and user-experience of a web-based interface through a retrospective think aloud. You will be asked to use the HCC Workshop to undertake 3 activities. As you are performing these activities, the experimenter will ask you to 'Think Aloud' and verbalise your thoughts.

Project Title: Retrospective Think Alouds: Improving user interfaces through eye tracking and verbalisation

Outline of the Activities:
- Activity A - Create a HCC Workshop account
- Activity B - Use the Eye Tracking Activity
- Activity C - Use the HCI Activity

Tasks Description:

<table>
<thead>
<tr>
<th>Activity</th>
<th>Task Number</th>
<th>Task Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>Register to create a new account for yourself</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Open up the HCC Workshop, and try to understand what it's about</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>Open up the Eye Tracking Activity, and find the meaning of saccades</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Find out how you can use Eye Tracking as a controller</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Find out the use of Eye Tracking in reading</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>Find out if law is a subject area associated with HCI</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Find out what good mapping means</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Find out how conceptual models help you tell the time</td>
</tr>
</tbody>
</table>
### 13.7. Surveying Document - Think Aloud

**Surveying Checklist**

<table>
<thead>
<tr>
<th>Strategy Think Aloud</th>
<th>Tally for Each Time Strategy Used</th>
<th>Cue Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicting</td>
<td></td>
<td>I predict that…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In the next part I think…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I think this is…</td>
</tr>
<tr>
<td>Questioning</td>
<td></td>
<td>Why did…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What did…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How did…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Where was…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Should there…</td>
</tr>
<tr>
<td>Visualising</td>
<td></td>
<td>I see…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I picture…</td>
</tr>
<tr>
<td>Personal Response</td>
<td></td>
<td>I feel…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I like…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I dislike…</td>
</tr>
<tr>
<td>Clarifying</td>
<td></td>
<td>I got confused when…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am not sure of…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I did not expect…</td>
</tr>
<tr>
<td>Synthesis</td>
<td></td>
<td>I am not sure of…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I did not expect…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I think this is about…</td>
</tr>
<tr>
<td>Summarising</td>
<td></td>
<td>I think this is mainly about…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The most important idea is…</td>
</tr>
<tr>
<td>Reflecting</td>
<td></td>
<td>I think I will… next time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maybe I will need to… next time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I realised that…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I wonder if…</td>
</tr>
<tr>
<td>Making Connections</td>
<td></td>
<td>This is like…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This reminds me of…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This is similar to…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If it were me…</td>
</tr>
</tbody>
</table>
## 13.8. Surveying Document - Task Completion

### Surveying Checklist

<table>
<thead>
<tr>
<th>Task</th>
<th>Completed without Guidance (Y/N)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register to create a new account for yourself</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open up the HCC Workshop, and try to understand what it's about</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open up the Eye Tracking Activity, and find the meaning of saccades</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find out how you can use Eye Tracking as a controller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find out the use of Eye Tracking in reading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find out if law is a subject area associated with HCI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find out what good mapping means</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find out how conceptual models help you tell the time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logout of the HCC Workshop</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Exit Interview

Please answer the following questions based on your experience with this study:

<table>
<thead>
<tr>
<th>Question</th>
<th>Score (Difficult → Easy)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>I found it easy to find issues with the interface</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>I found it easy to create an account</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>I liked the Neural Networks activity in the HCC Workshop</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>I liked the Human Computer Interaction activity in the HCC Workshop</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>Think Aloud was useful</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>I found it easy to logout of the HCC Workshop</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
</tbody>
</table>
13.10. The Heuristics Analysis Framework

The following is a high-level description for the heuristics analysis framework:

Framework Guide:

Column 1 indicates the code of the ‘factor’ group for the heuristic groups

Column 2 indicates individual heuristics being used to evaluate interface and UX

Column 3 indicates the ISO/IEC standard usability criteria for usability and ergonomic

Columns 4-10 indicates ISO/IEC standards for performance measurement of usability

Column 11 indicates the weighted average relative score for each individual heuristic

Column 12 indicates a feedback column for the evaluator to make annotation/comments/notes.

Usability Score:

Cell J250 indicates the weighted average relative performance score for performance measurement

Cell K249 indicates the weighted average relative performance score for usability heuristics

Cell K251 indicates the weighted average relative score for overall usability of the interface and UX.
<table>
<thead>
<tr>
<th>Code</th>
<th>Usability Heuristics</th>
<th>Usability Criteria</th>
<th>Evaluation Criteria</th>
<th>Satisfaction</th>
<th>Relative Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A9</td>
<td>Feedback: The system should always keep users informed about what is going on, through appropriate feedback and responsive help when necessary.</td>
<td>Feedback: The system should always keep users informed about what is going on, through appropriate feedback and responsive help when necessary.</td>
<td>Feedback: The system should always keep users informed about what is going on, through appropriate feedback and responsive help when necessary.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>Provide error recovery</td>
<td>Recovery for task</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F7</td>
<td>Feedback: 'Integrate feedback and error recovery so that users know when they have made an error and can take corrective action.'</td>
<td>Feedback: ‘Integrate feedback and error recovery so that users know when they have made an error and can take corrective action.’</td>
<td>Feedback: ‘Integrate feedback and error recovery so that users know when they have made an error and can take corrective action.’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E13</td>
<td>What features of the wizard change as users carry out tasks? What do features change as users carry out tasks?</td>
<td>Satisfaction</td>
<td>Relative Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>Speak the user's language: the dialogue should be expressed clearly as a natural, fluent, conversational form.</td>
<td>Speech</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>Contain familiar terms and natural language</td>
<td>Comprehensibility</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G7</td>
<td>User familiar should be considered</td>
<td>Compatibility</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>Metaphors from the real world take advantage of people's knowledge of the world</td>
<td>Self-documentation</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>Users and nodes should be supported</td>
<td>Self-documentation</td>
<td>User control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>Direct manipulation: subjects can perform visible while users perform physical actions on them, and the impact of these operations is immediately visible.</td>
<td>Usability for task</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Factor 1: Usability and efficiency of use**

<table>
<thead>
<tr>
<th>Q14</th>
<th>Accelerators should be provided</th>
<th>Usability for task</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q17</th>
<th>Feedback: Accelerators around the user's eyes – may often cause the user to react to the interaction for the expert user such that the system can cater to their experience and expert user needs.</th>
<th>Usability for task</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q9</th>
<th>User familiarity: Allow users to perform operations and changes to the user interface</th>
<th>Usability for task</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q6</th>
<th>User control allow the users to initiate and control actions</th>
<th>Usability for task</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q12</th>
<th>System should be efficient to use</th>
<th>Usability for task</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>User control</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flexibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error management</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compatibility</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-desirability</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Workload</td>
<td>0</td>
</tr>
</tbody>
</table>